Variational Autoencoders for Data Integration in Cognitive IoT

Variational Autoencoders for Heterogeneous Data Integration: Applications in Remote Sensing, Fusion, and Anomaly Detection

By
ALESSANDRO GIULIANO,
BEng. (Mechanical Engineering)

A Thesis Submitted to the School of Graduate Studies
in the Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in
Mechanical Engineering

McMaster University
Hamilton, Ontario

Doctor of Philosophy (2025)
Mechanical Engineering
McMaster University
Hamilton, Ontario, Canada

TITLE: Variational Autoencoders for Heterogeneous Data Integration: Applications in Remote Sensing, Fusion, and Anomaly Detection

AUTHOR:
Alessandro Giuliano,
BEng (Mechanical Engineering)

SUPERVISOR:
Dr. S. Andrew Gadsden
Associate Professor, Department of Mechanical Engineering,
McMaster University, ON, Canada

SUPERVISORY COMMITTEE CHAIR:
Dr. Gregory Wohl,
Professor, Department of Mechanical Engineering,
McMaster University, ON, Canada

SUPERVISORY COMMITTEE MEMBERS:
Dr. Stephen C. Veldhuis
Professor, Department of Mechanical Engineering,
McMaster University, ON, Canada

Dr. John Yawney
Adjunct Professor, Department of Mechanical Engineering,
McMaster University

SUPERVISORY COMMITTEE EXTERNAL MEMBER:
Dr. Apurva Narayan

Assistant Professor, Department of Computer Science and Electrical & Computer Engineering
Western University, ON, Canada

NUMBER OF PAGES: xix, 260

Abstract

This sandwich thesis comprises a comprehensive survey of Cognitive IoT and remote sensing systems, followed by three technical contributions that advance the state-of-the-art in data compression, multi-modal fusion, and anomaly detection. The increasing integration of the Internet of Things (IoT) and remote sensing systems has created an unprecedented need for efficient data processing, transmission, and integration. These systems often rely on heterogeneous data (spanning modalities such as numerical measurements, textual information, and imagery) each with unique characteristics and structures. While effective at reducing data size, traditional data compression and processing techniques often fail to retain the contextual and relational information required for downstream analytical tasks. This limitation is particularly acute in resource-constrained environments, where computational power, bandwidth, and energy are restricted. This thesis explores Variational Autoencoders (VAEs) as a unifying framework to address these challenges. VAEs provide a mechanism for encoding complex, multi-modal data into low-dimensional latent representations that are simultaneously compact, efficient to transmit, and inherently structured for interpretability. The overarching goal of this research is to establish a methodology for representing information such that heterogeneous data can be processed, compressed, and fused seamlessly. The research is organized around three key objectives: (1) developing and fine-tuning VAE architectures that generate compressed latent spaces optimized for direct classification and reconstruction, minimizing the reliance on reconstructive processing while preserving interpretability, (2) investigating the capacity of VAEs for multi-modal data fusion by combining disparate data types, such as Synthetic Aperture Radar (SAR) and optical imagery, into a unified latent representation, and (3) evaluating the potential of VAE-derived latent spaces for anomaly detection, particularly in applications where identifying critical events or failures is essential. These results collectively underscore the potential of VAEs not only as tools for compression but also as versatile foundations for diverse analytical and predictive tasks across varied datasets. In the broader context of remote sensing and IoT, these methods align well with the overarching theme of the thesis to increase system efficiency through multi-level intelligence and distributed computing. By leveraging compressive sensing and latent representations, these approaches facilitate reduced data transmission and enhanced computational efficiency, supporting the development of scalable architectures for data-rich applications in IoT and remote sensing environments. The results also demonstrate that compressive VAEs generate

rich latent spaces, enabling their dual use for direct downstream tasks and reconstruction as well as for data fusion and anomaly detection. This implies that deploying VAEs for compression on edge devices could fundamentally transform data transmission workflows. Rather than transmitting raw data, edge devices could send compressed, machine-learning-interpretable representations, reducing bandwidth requirements while preserving essential information for analysis and data fusion. This approach not only enhances efficiency but also lays the groundwork for intelligent, resource-aware systems capable of performing complex, real-time tasks through distributed and interpretive data handling. This thesis highlights the transformative potential of VAEs for addressing the critical challenges associated with processing and fusing heterogeneous data. By leveraging their inherent flexibility and capacity for structured representation, VAEs provide a scalable, interpretable, and resource-efficient approach for data-intensive applications in IoT. cognitive IoT (CIoT) and remote sensing. The findings lay a foundation for future research into compressive neural networks and their broader applications in intelligent systems.

Acknowledgements

This thesis marks the culmination of nearly four years of dedication and perseverance, shaped by some of the most profound highs and lows of my life. Much like life itself, the path was not always clearly defined, and this journey stands as a reflection of that uncertainty. It has been a teacher in many ways, guiding me through different phases of life, beginning at the close of my bachelor's degree, carrying me through moments of immense joy, and putting me through some of the most challenging times. Yet, we continue to move forward, undeterred, striving to become better versions of ourselves, shaped and strengthened by the experiences we endure and, with resilience, ***I will keep enduring and growing stronger.***

A mia madre*, e al suo amore che mi ha sempre seguito ovunque sia andato. Al suo supporto e alla sua dedizione e premura cercando sempre di essere sicura che avessi tutto e che stessi bene. A lei che mi ha insegnato a prendermi cura di me stesso e a volermi bene, a non essere troppo duro con me stesso.*

A mio padre, *e al suo supporto, che instancabile riesce sempre a darmi, non importa come o quando; ci sei sempre stato per me. Al suo incoraggiamento che infallibile riesce a tirarmi su nei momenti più duri e a farmi andare avanti. A lui che nella vita mi ha insegnato a essere forte, a reagire e a non mollare mai, massicci e incazzati.*

A mia zia*, che mi ha accolto come un figlio nella sua casa e mi ha reso parte della sua famiglia. Nel corso degli anni, mi hai sempre aiutato e insegnato tanto nella vita. Sei un angelo, sempre pronta ad aiutare gli altri, chiunque essi siano.*

A nonna Mila e nonno Tonino*, anche se non siete più qui, questa tesi è dedicata anche a voi. Spero che siate fieri di me. Quando guardo il cielo di notte, le due stelle più luminose siete sempre voi.*

A nonno Adriano e nonna Maria*, mi avete preceduto in un viaggio che pochi possono comprendere, amandomi e incoraggiandomi sin da quando da bambino correvo per il balcone. Grazie per la vostra saggezza e inestinguibile sostegno.*

A mia sorella, *che riesce sempre a farsi sentire vicina, anche da lontano, e a strappare un sorriso nei momenti più difficili. Grazie per esserci sempre: anche se la vita ci ha fisicamente diviso, so che non ci allontaneremo mai davvero. Prometto di scriverti più spesso.*

To Dr. Gadsden, *who has given me this opportunity and encouraged me to keep going throughout the years. Thank you for being so supportive, and for the opportunities you gave me. Thank you for all the laughs and responsibility you trusted me with, they helped me grow immensely.*

To my friends, *thank you for all of the laughs and all of the experiences, it was some of the best years of my life. To Can and all of the nights out just the two of us causing trouble. To Wally and all of the crazy things we went through, you are like a big brother to me and I love you and hate you for it. To Simone, you have always been there for me with a kind word and a positive remark, bringing out the best in me. Your ability to uplift and inspire others is remarkable, you see the good in everyone and help them reach their potential. You are truly an inspiration to all who have the privilege of knowing you.*

To all of the members of the coldest lab in the University, *you are some of the brightest people I have ever had the privilege to meet, thank you for all of the laughs we shared, all of the soccer games, and all of the trivia nights lost, keep drinking Ben's milk.*

Ai miei amici, *che negli anni mi sono sempre stati affianco anche molto dopo che sono partito per questa avventura. A Roberto, e a tutte le notti fonde spese a giocare insieme da quando eravamo piccoli fino ad oggi e per sempre. Nella vita ci sono poche amicizie che sai trascenderanno qualsiasi cosa accada e io so che la nostra è una di quelle. Grazie per esserci sempre e grazie per riuscire sempre a farmi dimenticare tutti i miei problemi spendendo ore e ore insieme a me, anche se da remoto. A Giovanna che nonostante tutto riesce a contattarmi dai posti più assurdi in giro per il mondo. Se una forza della natura non te lo dimenticare mai. A Davide, una delle persone più pure che abbia mai incontrato non cambiare mai. Ad Arianna che riesce sempre a prendersi cura di tutti con gran premura. A Edoardo, che sa sempre accoglierti con un sorriso caloroso e un drink al momento giusto. A Giuliano, con cui condivido molto, e con cui riesco sempre a trovare una sintonia naturale, come se i nostri pensieri fossero uno. A tutti i membri di tutti in carrozza con cui ho passato i migliori momenti della mia vita siete sempre con me ovunque vada.*

To my beautiful girlfriend*, thank you for always being by my side and taking care of me even during some of the hardest times. Your work ethic and kindness inspire me every day to be a better person. You are so strong, even if you don't always realize it. You are my rock, and I love you more than words can ever express.*

Contents

List of Figures

List of Tables

Declaration of Academic Achievement

I, Alessandro Giuliano, hereby declare that this thesis, titled **Variational Autoencoders for Heterogeneous Data Integration: Applications in Remote Sensing, Fusion, and Anomaly Detection**, and the work presented herein, are my own. I confirm that I have independently conceived, conducted the experiments, and authored all papers included in this thesis.

Chapter 1

Introduction

The rapid expansion of interconnected devices within the Internet of Things (IoT) has revolutionized industries ranging from healthcare and transportation to agriculture and environmental monitoring by enabling real-time data collection, enhanced decision-making, and improved operational efficiency. In healthcare, IoT devices facilitate remote patient monitoring and personalized treatment plans, while in transportation, they optimize traffic management and enable autonomous vehicles. Agriculture benefits from IoT through precision farming techniques, such as monitoring soil conditions and automating irrigation systems, and environmental monitoring leverages IoT to track air and water quality, helping to address pollution and climate change. IoT systems collect and process vast volumes of data from diverse sources, often encompassing multiple modalities such as numerical measurements, images, and textual information. However, this growing complexity poses significant challenges in data integration, transmission, and analysis. Traditional IoT architectures, which predominantly rely on centralized systems and conventional compression techniques, often fail to manage the heterogeneity and immense scale of data generated by modern IoT networks.

To address these limitations, Cognitive IoT (CIoT) has emerged as a paradigm that incorporates elements of artificial intelligence and cognition into IoT systems. CIoT systems aim to adapt dynamically to changing environments and perform tasks with minimal human intervention by mimicking cognitive processes such as perception, memory, and intelligence. Despite these advancements, existing data compression, fusion, and anomaly detection methods face significant challenges in real-world applications. These methods often struggle to meet demands for scalability, efficiency, and interpretability. For example, traditional compression methods frequently sacrifice contextual and relational information, limiting their utility for downstream analytical tasks. Similarly, existing techniques for fusing data from multiple modalities often fail to capture the

nuanced relationships between different data types, making it difficult to extract meaningful insights from diverse sensor inputs and data streams, resulting in suboptimal system performance.

This thesis investigates the application of Variational Autoencoders (VAEs) as a unifying framework to address these challenges. VAEs, a class of generative neural networks, provide a mechanism for encoding complex, multi-modal data into low-dimensional latent spaces. These latent representations are compact, efficient for transmission, and structured in ways that preserve the critical information necessary for downstream tasks. By leveraging VAEs, it becomes possible to integrate data compression, fusion, and anomaly detection into a single, cohesive framework.

1.1 Motivation and Problem Statement

The integration of IoT systems into critical applications such as remote sensing, smart cities, and autonomous vehicles requires novel approaches to manage the heterogeneity and volume of data. For example, in remote sensing, the combination of Synthetic Aperture Radar (SAR) data with optical imagery provides a richer understanding of the environment. However, current methods for fusing such disparate data types often require substantial pre-processing and computational resources, making them unsuitable for deployment in resource-constrained environments. Furthermore, traditional anomaly detection techniques are often ineffective in high-dimensional, noisy datasets, which are characteristic of IoT applications.

At the heart of these challenges lies the need for a framework that can:

- Compress heterogeneous data efficiently while preserving essential information.
- Fuse disparate data modalities into a unified representation that facilitates downstream tasks such as classification and anomaly detection.
- Operate effectively in resource-constrained environments, such as edge devices with limited computational and energy resources.

This thesis addresses these needs by leveraging VAEs to create structured latent spaces that serve as both compressed representations and feature-rich inputs for analytical models. Unlike traditional methods, VAEs enable the direct utilization of these latent spaces for tasks such as classification, significantly reducing the need for reconstructive processing.

1.2 Variational Autoencoders: A Bayesian Framework

Variational Autoencoders (VAEs) are a probabilistic framework rooted in Bayesian reasoning, designed to model and generate complex data distributions. At their core, VAEs encode high-dimensional data into a probabilistic representation, where each data point is described as a distribution over latent variables rather than a single deterministic point. This probabilistic nature allows VAEs to capture the inherent uncertainty and variability in the data, making them highly effective for downstream tasks such as classification, fusion, and anomaly detection. This section introduces the mathematical foundation of VAEs, highlighting their connection to Bayesian principles and the advantages of representing data through distributions.

1.2.1 Bayesian Reasoning and VAEs

Bayes' rule forms the foundation of probabilistic modeling in VAEs. For observed data x and latent variables z, the posterior distribution $p(z|x)$ is defined as:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)},$$

where:

- $p(z)$ is the prior distribution over the latent variables.
- $p(x|z)$ is the likelihood of the observed data given the latent variables.
- $p(x) = \int p(x|z)p(z)\,dz$ is the marginal likelihood, serving as a normalization constant.

The posterior $p(z|x)$ encapsulates our updated beliefs about z after observing x. However, the computation of $p(x)$ often involves an intractable integral, especially in high-dimensional latent spaces. To address this, VAEs approximate the posterior $p(z|x)$ using a variational distribution $q_\phi(z|x)$, parameterized by a neural network.

Evidence Lower Bound (ELBO)

The Variational Inference approach maximizes the Evidence Lower Bound (ELBO), a tractable surrogate to the intractable log-marginal likelihood $\log p(x)$. The ELBO is derived as:

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p(x|z)] - \mathrm{KL}(q_\phi(z|x)\|p(z)),$$

where:

- $\mathbb{E}_{q_\phi(z|x)}[\log p(x|z)]$ is the reconstruction term, ensuring that the latent representation can faithfully reconstruct the observed data.
- $\mathrm{KL}(q_\phi(z|x)\|p(z))$ is the regularization term, which enforces the learned latent space to remain close to the prior $p(z)$, typically a standard Gaussian distribution $\mathcal{N}(0, I)$.

1.2.2 Encoder-Decoder Architecture

VAEs employ an encoder-decoder architecture to approximate the posterior $q_\phi(z|x)$ and reconstruct the data via $p_\theta(x|z)$. The encoder maps x to a latent representation z, parameterized by a mean $\mu_\phi(x)$ and variance $\sigma^2_\phi(x)$:

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \mathrm{diag}(\sigma^2_\phi(x))).$$

The decoder reconstructs x from z using the likelihood:

$$p_\theta(x|z) = \mathcal{N}(x; \hat{\mu}_\theta(z), \mathrm{diag}(\hat{\sigma}^2_\theta(z))).$$

1.2.3 Reparameterization Trick

To enable gradient-based optimization, VAEs use the reparameterization trick to sample z in a differentiable manner:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where $\odot$ denotes element-wise multiplication. This formulation ensures that gradients can flow through the stochastic sampling process during backpropagation.

1.2.4 VAE Loss Function

The loss function for VAEs is derived from the negative ELBO:

$$\mathcal{L}_{\mathrm{VAE}}(x; \phi, \theta) = \mathrm{KL}(q_\phi(z|x)\|p(z)) - \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)].$$

Minimizing this loss balances two objectives: regularizing the latent space to align with the prior and ensuring accurate reconstruction of the observed data.

1.2.5 Bayesian Context in CIoT Applications

The Bayesian framework of VAEs aligns closely with the principles of cognitive reasoning observed in humans. At its core, Bayesian reasoning involves updating beliefs based on new evidence, a process that mirrors human cognition when encountering and processing information. In the context of Cognitive IoT (CIoT), leveraging this framework provides several key advantages for handling the challenges of uncertainty and data heterogeneity:

- **Latent Space Structuring**: The learned latent representations in VAEs are inherently probabilistic, modeled as a posterior distribution $q_\phi(z|x)$ approximating $p(z|x)$. This structured approach enables seamless fusion and classification of heterogeneous data by capturing nuanced relationships between modalities. This mirrors the way humans synthesize information from multiple sensory inputs, such as sight and sound, into a coherent perception of their environment.

- **Uncertainty Quantification**: The variational posterior $q_\phi(z|x)$ provides a measure of uncertainty in the latent space. For example, when a VAE encounters incomplete or noisy data, the posterior reflects this uncertainty, guiding the system to make probabilistic decisions. Similarly, humans operate under uncertainty by assigning confidence levels to their beliefs, adjusting them as new information becomes available. This is critical in anomaly detection, where the ability to quantify and respond to uncertainty can improve system reliability in dynamic, unpredictable environments.

- **Scalability through Prior Knowledge**: In Bayesian reasoning, the prior $p(z)$ serves as a foundation of existing knowledge, which is updated as evidence accumulates. For VAEs, this prior ensures the latent space remains regularized, avoiding overfitting and enabling generalization across diverse datasets. In CIoT systems, this translates to efficient data compression and transmission, reducing bandwidth and computational overhead. Humans similarly rely on prior knowledge to make rapid and scalable decisions, refining their understanding with experience.

- **Probabilistic Decision-Making**: VAEs, grounded in Bayesian principles, enable probabilistic decision-making based on the likelihood of observed data under the generative model, where decisions often weigh multiple probabilistic outcomes rather than deterministic absolutes. For instance, in CIoT systems, probabilistic reasoning can guide resource allocation or anomaly detection under conditions of limited or ambiguous data.

In the broader context of CIoT, the ability to handle uncertainty, heterogeneity, and dynamic data environments is paramount. VAEs, as Bayesian models, address these challenges effectively, supporting applications such as multi-modal data fusion, anomaly detection, and real-time decision-making. This thesis demonstrates how grounding VAE-based methods in Bayesian reasoning not only enhances their technical capabilities but also aligns them with the cognitive principles that underpin human intelligence, paving the way for smarter and more adaptable IoT systems.

1.3 Research Objectives

The primary objectives of this research are as follows:

1. **Study VAE Architectures for Compression and Classification:** Design and optimize VAE models that produce latent representations capable of direct classification and high-quality reconstruction, balancing efficiency and interpretability.
2. **Explore Multi-Modal Data Fusion:** Investigate the use of VAEs to combine heterogeneous data types, such as SAR and optical imagery, into unified latent spaces that enhance analytical performance.
3. **Utilize VAEs for Anomaly Detection:** Evaluate the potential of VAE-derived latent spaces for identifying anomalies in dynamic, resource-constrained environments.

Through these objectives, the thesis aims to establish VAEs as a versatile tool for addressing critical challenges in CIoT applications.

1.4 Contributions and Significance

This thesis makes several significant contributions to the fields of data compression, multi-modal data fusion, and anomaly detection. One of the primary contributions is the demonstration of the direct utilization of latent representations constructed by neural compression models for downstream machine learning tasks, such as classification. This approach leverages the efficiency and compactness of learned latent spaces by bypassing the need for explicit reconstruction or inverse transformation, enabling immediate application to analytical tasks. This represents a departure from conventional methods, which typically rely on full data reconstruction, and validates the potential of using lower-dimensional representations directly for diverse machine learning applications.

Another key contribution lies in developing methodologies for the fusion of heterogeneous data modalities using VAE latent spaces. By integrating disparate data types, such as Synthetic Aperture Radar (SAR) and optical imagery, into unified latent representations, these methodologies significantly enhance the accuracy and efficiency of downstream tasks. The resulting fusion approaches enable more robust and interpretable analytics, particularly in scenarios involving complex and multi-faceted data, such as remote sensing applications.

Finally, this thesis validates the application of VAEs for anomaly detection in high-dimensional datasets. By leveraging the structured and probabilistic nature of the latent space, VAEs are shown to identify critical events and anomalies in challenging environments effectively. This capability addresses real-world problems that require precise and reliable anomaly detection, even under conditions of uncertainty and incomplete data. Together, these contributions highlight the transformative potential of VAEs for addressing core challenges in data-rich, heterogeneous, and resource-constrained environments.

The findings presented in this work have broad implications for the design of next-generation CIoT systems. By integrating compression, fusion, and anomaly detection into a single framework, VAEs enable the development of intelligent, resource-efficient systems capable of real-time analytics and decision-making. This research also lays the groundwork for future studies on compressive neural networks and their applications in domains such as autonomous systems, smart cities, and distributed sensor networks.

1.5 List of Publications

The following is a comprehensive list of publications completed during my graduate studies, reflecting the outcomes of my research efforts and collaborations.

Journals:

[4] A. Giuliano, S. Andrew Gadsden and J. Yawney, "Optimizing Satellite Image Analysis: Leveraging Variational Autoencoders Latent Representations for Direct Integration," in IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1-23, 2025, Art no. 5603123, doi: 10.1109/TGRS.2024.3520879.

[5] A. Giuliano, S. Andrew Gadsden and J. Yawney, “Cognitive internet of things: A review of frameworks, applications, and recent advances,” IEEE Communication Surveys and Tutorials, 2024, First round of revisions.

[6] A. Giuliano, S. Andrew Gadsden and J. Yawney, "Enhancing data fusion and classification of sentinel-1 and sentinel-2 imagery using neural compression," Information Fusion, 2024, Under Review.

[7] A. Giuliano, S. Andrew Gadsden and J. Yawney, "Anomaly detection of under-over current in magnetorheological damper suspension using variational autoencoders," IEEE/ASME Transactions on Mechatronics, 2024, Under Review.

[8] A. Giuliano, S. Andrew Gadsden and J. Yawney, "Transformer-based transfer learning for battery state of health estimation," Energy Reports, 2024, Under Review.

Conferences:

[9] A. Giuliano, S. A. Gadsden, W. Hilal, and J. Yawney, "Convolutional variational autoencoders for secure lossy image compression in remote sensing," in Sensors and Systems for Space Applications XVII, K. D. Pham and G. Chen, Eds., National Harbor, United States: SPIE, Jun. 2024, p. 18, isbn: 978-1-5106-7442-4 978-1-5106-7443-1. doi: 10.1117/12.3013451.

[10] N. Alsadi, A. Giuliano, S. A. Gadsden, and J. Yawney, "An adaptive approach to blockchain in smart system applications," in Big Data V: Learning, Analytics, and Applications, vol. 12522, SPIE, 2023, pp. 27–32.

[11] A. Giuliano, G. Bone, S. A. Gadsden, and M. AlShabi, "A Comparative Analysis of Control Methods Applied to Horizontal 2 DOF Robotic Arms," in 2023 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates: IEEE, Feb. 2023, pp. 01–09, isbn: 978-1-66545-474-2. doi: 10.1109/ASET56582.2023.10180701.

[12] A. Giuliano, W. Hilal, N. Alsadi, J. Yawney, and S. A. Gadsden, "Normalized determinant pooling layer in CNNs for multi-label classification," in Computational Imaging VII, J. C. Petruccelli and C. Preza, Eds., Orlando, United States: SPIE, Jul. 2023, p. 17, isbn: 978-1-5106-6160-8 978-1-5106-6161-5. doi: 10.1117/12.2663916.

[13] N. Alsadi, W. Hilal, O. Surucu, A. Giuliano, A. Gadsden, and J. Yawney, "Visual attention for malware classification," in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV, T. Pham, L. Solomon, and M. E. Hohil, Eds., Orlando, United States: SPIE, Jun. 2022, p. 74, isbn: 978-1-5106-5102-9 978-1-5106-5103-6. doi: 10.1117/12.2619009. []

[14] N. Alsadi, W. Hilal, O. Surucu, A. Giuliano, S. A. Gadsden, and J. Yawney, "An optimized volumetric approach to unsupervised image registration," in Big Data IV: Learning, Analytics, and Applications, F. Ahmad, P. P. Markopoulos, and B. Ouyang, Eds., Orlando, United States: SPIE, May 2022, p. 15, isbn: 978-1-5106-5070-1 978-1-5106-5071-8. doi: 10.1117/12.2618647.

[15] N. Alsadi, W. Hilal, O. Surucu, et al., "An anomaly detecting blockchain strategy for secure IoT networks," in Disruptive Technologies in Information Sciences VI, M. Blowers, R. D. Hall, and V. R. Dasari, Eds., Orlando, United States: SPIE, May 2022, p. 18, isbn: 978-1-5106-5110-4 978-1-5106-5111-1. doi: 10.1117/12.2618301.

[16] N. Alsadi, W. Hilal, O. Surucu, et al., "Neural network training loss optimization utilizing the sliding innovation filter," in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV, T. Pham, L. Solomon, and M. E. Hohil, Eds., Orlando, United States: SPIE, Jun. 2022, p. 76, isbn: 978-1-5106-5102-9 978-1-5106-5103-6. doi: 10.1117/12.2619029.

[17] A. Giuliano, W. Hilal, N. Alsadi, S. A. Gadsden, and J. Yawney, "A Review of Cognitive Dynamic Systems and Cognitive IoT," in 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada: IEEE, Jun. 2022, pp. 1–7, isbn: 978-1-66548-684-2. doi:10.1109/IEMTRONICS55184.2022.9795834.

[18] A. Giuliano, W. Hilal, N. Alsadi, et al., "Efficient utilization of big data using distributed storage, parallel processing, and blockchain technology," in Big Data IV: Learning, Analytics, and Applications, F. Ahmad, P. P. Markopoulos, and B. Ouyang, Eds., Orlando, United States: SPIE, May 2022, p. 3, isbn: 978-1-5106-5070-1 978-1-5106-5071-8. doi: 10.1117/12.2618891.

[19] W. Hilal, A. Giuliano, S. A. Gadsden, and J. Yawney, "A Review of Cognitive Dynamic Systems and Its Overarching Functions," in 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, ON, Canada: IEEE, Jun. 2022, pp. 1–10, isbn: 978-1-66548-684-2. doi: 10.1109/IEMTRONICS55184.2022.9795764.

[20] W. Hilal, C. Wilkinson, N. Alsadi, et al., "A topic modeling-based approach to executable file malware detection," in Disruptive Technologies in Information Sciences VI, M. Blowers, R. D. Hall, and V. R. Dasari, Eds., Orlando, United States: SPIE, May 2022, p. 4, isbn: 978-1-5106-5110-4 978-1-5106-5111-1. doi: 10.1117/12.2619033.

[21] W. Hilal, C. Wilkinson, A. Giuliano, et al., "Minority class augmentation using GANs to improve the detection of anomalies in critical operations," in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV, T. Pham, L. Solomon, and M. E. Hohil, Eds., Orlando, United States: SPIE, Jun. 2022, p. 69, isbn: 978-1-5106-5102-9 978-1-5106-5103-6. doi: 10.1117/12.2618858.

[22] O. Surucu, C. Wilkinson, U. Yeprem, et al., "PROGNOS: An automatic remaining useful life (RUL) prediction model for military systems using machine learning," in Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV, T. Pham, L. Solomon, and M. E. Hohil, Eds., Orlando, United States: SPIE, Jun. 2022, p. 71, isbn: 978-1-5106-5102-9 978-1-5106-5103-6. doi: 10.1117/12.2618913.

[23] O. Surucu, U. Yeprem, C. Wilkinson, et al., "A survey on ethereum smart contract vulnerability detection using machine learning," in Disruptive Technologies in Information Sciences VI, M. Blowers, R. D. Hall, and V. R. Dasari, Eds., Orlando, United States: SPIE, May 2022, p. 12, isbn: 978-1-5106-5110-4 978-1-5106-5111-1. doi: 10.1117/12.2618899.

1.6 Thesis Organization

The remainder of this thesis is structured as follows:

- Chapter 2 provides a comprehensive review of the CIoT paradigm, highlighting its key components, challenges, and enabling technologies.
- Chapter 3 introduces the proposed VAE architectures and their application to data compression and classification.
- Chapter 4 explores the use of VAEs for multi-modal data fusion, with a focus on remote sensing applications.
- Chapter 5 examines the role of VAEs in anomaly detection, presenting experimental results and analyses.
- Chapter 6 concludes the thesis, underscoring the key findings, their implications, and potential directions for future research.

This thesis represents a significant step forward in the development of scalable, intelligent systems for CIoT applications, demonstrating the transformative potential of VAEs in addressing the challenges of heterogeneous data integration.

Chapter 2

What is Cognitive IoT

The content of this chapter is a second revision of the manuscript text for publication under the following citation:

Giuliano, A. (2024). Cognitive Internet of Things: A Review of Theory, Applications, and Recent Advances. IEEE Communications Surveys & Tutorials.

Cognitive Internet of Things: A Review of Theory, Applications, and Recent Advances

Alessandro Giuliano
Faculty of Engineering
McMaster University, Hamilton, ON, Canada
Email: giuliana@mcmaster.ca

S. Andrew Gadsden
Faculty of Engineering
McMaster University
Email: gadsdesa@mcmaster.ca

John Yawney
Adastra Corp.
Email: john.yawney@adastragrp.com

Abstract

With the development of increasingly interconnected cyber-physical systems (CPSs), the Internet of Things (IoT) paradigm must be expanded further to account for the collection, transmission, and processing of unprecedented amounts of data in uncertain and changing environments. Cognitive Internet of Things (CIoT) introduces a paradigm shift in IoT systems by integrating the engineering perspective of cognition, as formulated in cognitive dynamic systems (CDS), into traditional IoT frameworks. This survey systematically examines how CIoT leverages the five pillars of cognition: perception, attention, memory, language, and intelligence, to enable context-aware, autonomous, and adaptive functionality. We trace the evolution from standard IoT architectures to this cognitively enriched model, detailing how data acquisition and storage, combined with enabling technologies such as data fusion, reinforcement learning, cognitive communications (via cognitive radios), and the integration of foundation models and large language models (LLMs), facilitate advanced data analytics and introduce a new intelligent layer for deeper contextual understanding and adaptation. By emphasizing the

synergy between CDS principles and emerging technologies, the paper demonstrates how CIoT can address longstanding challenges in scalability, interoperability, and resource management. Through a critical evaluation of current limitations and lessons learned, we offer a forward-looking perspective on how these cognitively inspired frameworks can further enhance intelligent IoT ecosystems. Ultimately, this work serves as a foundational resource for aligning IoT systems with the engineering-driven notion of cognition, guiding future research and innovation in autonomous, scalable IoT environments.

Keywords: *Cloud Computing, Cognitive Computing, IoT, Edge Computing, Federated Learning.*

Nomenclature

AI Artificial Intelligence

AIOTI-HLA Alliance for Internet of Things Innovation - High-Level Architecture

ANN Artificial Neural Network

AMI Advanced Metering Infrastructure

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

BIOCAS Biomedical Circuits and Systems Conference

BLE Bluetooth Low Energy

CAV Connected Autonomous Vehicle

CBTC Communication-Based Train Control

CDS Cognitive Dynamic Systems

CIoT Cognitive Internet of Things

CLIP Contrastive Language–Image Pre-Training

CNN Convolutional Neural Networks

CoAP Constrained Application Protocol

CPS Cyber-Physical Systems

CPU Central Processing Unit

CPSS Cyber-Physical-Social System

CR Cognitive Radio

CRN Cognitive Radio Networks

CT Computerized Tomography

CTC Cross Technology Communication

CVO Cognitive Virtual Object

D2D Device-to-Device Communication

DAG Directed Acyclic Graph

DHT Distributed Hash Table

DNS Domain Name System

DoF Degrees of Freedom

DRL Deep Reinforcement Learning

DRAM Deep RL-Based Resource Allocation

ECC Edge Cognitive Computing

ECG Electrocardiogram

EHR Electronic Health Record

EMG Electromyography

ERA Economic Resource Allocation

ETSI European Telecommunications Standards Institute

EU European Union

EVD Eigenvalue Decomposition

FA Factor Analysis

FCC Federal Communication Commission

FDI False Data Injection Attack

FL Federated Learning

fMRI Functional Magnetic Resonance Imaging

FPT Frozen Pre-Trained Transformer

FSYNC Fog Sync Differential Algorithm

GPT Generative Pre-trained Transformer

GPRFCA Gaussian Process Regression for Fog-Cloud Allocation

GSM Global System for Mobile Communications

HDFS Hadoop Distributed File System

HMM Hidden Markov Model

HS Hyper-spectral

HTTP HyperText Transfer Protocol

HTSA Hybrid Tabu-Based Simulated Annealing

HVAC Heating, Ventilation, and Air Conditioning

IaaS Infrastructure as a Service

IBM International Business Machines

ICA Independent Component Analysis

ICU Intensive Care Unit

IEEE Institute of Electrical and Electronics Engineers

INS Inertial Navigation System

IPFS InterPlanetary File System

IoHT Internet of Health Things

IoT Internet of Things

IoV Internet of Vehicles

ISO International Organization for Standardization

KBL Kernel-Based Learning

KNN K Nearest Neighbor

KNX Konnex Protocol

LED Light Emitting Diode

LIDAR Light Detection and Ranging

LLM Large Language Model

LLaMA Large Language Model Meta AI

LoRA Low-Rank Adaptation

LULC Land Use and Land Cover

LUSI Lunar Spectral Irradiance

LTE Long-Term Evolution

M2M Machine-to-Machine Communication

MAC Media Access Protocol

MDP Markov Decision Process

MEC Mobile Edge Computing

MLE Maximum Likelihood Estimation

MQTT Message Queuing Telemetry Transport

MRI Magnetic Resonance Images

MVAE Multimodal Variational Autoencoders

NCS Networked Control System

NLG Natural Language Generation

NPU Neural Processing Unit

OFDM Orthogonal Frequency-Division Multiplexing

P2P Peer-to-Peer

PAC Perception-Action Cycle

PaaS Platform as a Service

PBRA Prediction-Based Resource Allocation Algorithm

PCA Principal Component Analysis

PCEN Portable Cognitive Emergency Network

PCR Principal Component Regression

PET Positron Emission Tomography

PLC Programmable Logic Controller

PPDR Public Protection and Disaster Relief

PU Primary User

QoD Quality of Data

QoE Quality of Experience

QoI Quality of Information

QoS Quality of Service

RA Reference Architecture

RAM Random Access Memory

RACE Risks Sensitive, Autonomous, and Connected Electrical Vehicles

RF Radio Frequency

RFID Radio Frequency Identification

RL Reinforcement Learning

RLHF Reinforcement Learning from Human Feedback

RM Reference Model

RNN Recurrent Neural Network

RS-FSYNC Reed-Solomon Fog Sync

RSM Request Situation Matching

SAS Symbiotic Autonomous Systems

SAR Synthetic-Aperture Radar

SG Smart Grid

SMOTE Synthetic Minority Over-sampling Technique

SPO2 Blood Oxygen Saturation

SU Secondary User

SSL Secure Sockets Layer

SVD Singular Value Decomposition

TCP Transmission Control Protocol

TLS Transport Layer Security

TinyML Tiny Machine Learning

TinyTL Tiny Transfer Learning

TUF Tensor Unified Fusion

TUFF Task-Utility Function Framework

UAV Unmanned Aerial Vehicle

UHF Ultra High Frequency

UID Unlicensed IoT Devices

UWB Ultra-Wideband

V2V Vehicle-to-Vehicle Communication

V2X Vehicle-to-Everything Communication

VAE Variational Autoencoders

VATT Video-Audio-Text Transformer

VO Virtual Object

VHF Very High Frequency

2.1 Introduction

The proliferation of electronic devices and the integration of processing and communication capabilities into physical objects have become key drivers of the fourth industrial revolution. Many aspects of everyday life are becoming increasingly interconnected, particularly in areas such as wearable technology, healthcare, home appliances, and transportation. The Internet of Things (IoT) paradigm has emerged in response to this progress, envisioned as a seamless integration of physical objects with the cyber world. This technology's potential impact is vast and being actively explored across various disciplines. For example, in industrial applications, IoT can transform simple sensor and actuator-based control systems into sophisticated networks that cooperate to share data, enabling more accurate and effective decision-making through multi-level intelligence.

In this new era of increased interconnectivity, novel cyber-physical systems (CPS) are constantly being introduced, enhancing objects and infrastructure with the ability to perceive and control the environment around them, process the collected data, and communicate it through the internet [24]. The crux of IoT development is the unification of the physical and digital realms required for such advanced systems. New IoT architectures are hypothesized to be able to make semantic inferences from obtained and contextual data, increasing the system's performance and efficiency autonomously and allowing self-adjustment in response to variable conditions and unexpected events. In the pursuit of this level of autonomy, researchers have turned to the most adaptable system known, the brain, for inspiration. Biomimicry is not a novel concept, having been demonstrated as a powerful tool in engineering design and computation for many years [25, 26, 27, 28]. Emulating the human mind is considered the pinnacle of biomimicry, as it is proficient in its capability to move between multiple functional states to meet the demands of the environment [29]. The critical question is then how can we draw from elements of human cognition to implement them in engineering systems and allow for contextual dynamic adaptability?

The research of Dr. Simon Haykin formally introduced the novel concept of cognitive dynamic systems (CDS) in an article published in 2006 [30]. He later expanded his ideas

to radio and radar systems in two very influential publications in electrical engineering, [1] and [31]. In these, the author defines the workings of CDS as follows:

"Cognitive dynamic systems build up rules of behavior over time through learning from continuous experiential interactions with the environment, and thereby deal with environmental uncertainties."

This early concept was later refined by Haykin in [32], in collaboration with neuroscientist Dr. Joaquin Fuster, outlining the correlations between adaptation and cognition, and presenting the principles by which an engineering system can be defined cognitive; namely the perception action cycle (PAC), memory, attention, language, and intelligence.

Based on Haykin's work, Wu *et al.* proposed Cognitive IoT (CIoT) in [33] to enhance current IoT systems by mimicking human cognitive capabilities, efficiently utilizing large datasets and addressing the challenges inherent to scalability and adaptability. The goal of CIoT is to enable IoT systems to understand and develop contextual awareness in response to variable environmental conditions, crucial considerations for ever-changing dynamic systems.

The CIoT paradigm in a broader sense can be defined as a worldwide network of objects which are interconnected and uniquely addressable based on standard communication protocols [34]. This definition embraces a theoretical model of complex multidimensional systems composed of interlinked and interdependent objects [35], gated by encryption protocols, passwords, and private networks. As part of this framework, smaller subsystems work together to achieve local or global goals in the most optimal manner. Tools like social network analysis are also important, used to describe the network of relations present among the various objects composing the broader IoT and to discover their effects on data analysis, context extrapolation, and semantic derivation. This is particularly useful in large-scale dynamic IoT systems, such as smart cities, smart grids [36], and distributed manufacturing [27]. By effectively integrating cognition processes in a variety of IoT systems, we can expect to obtain the ability to sufficiently learn and understand the dynamics of physical and social environments through data and interaction. This ultimately enables the creation of a new generation of systems that require minimal human intervention [33]. However, to construct such systems, numerous technical challenges for a scalable and efficient IoT must be addressed and resolved a priori.

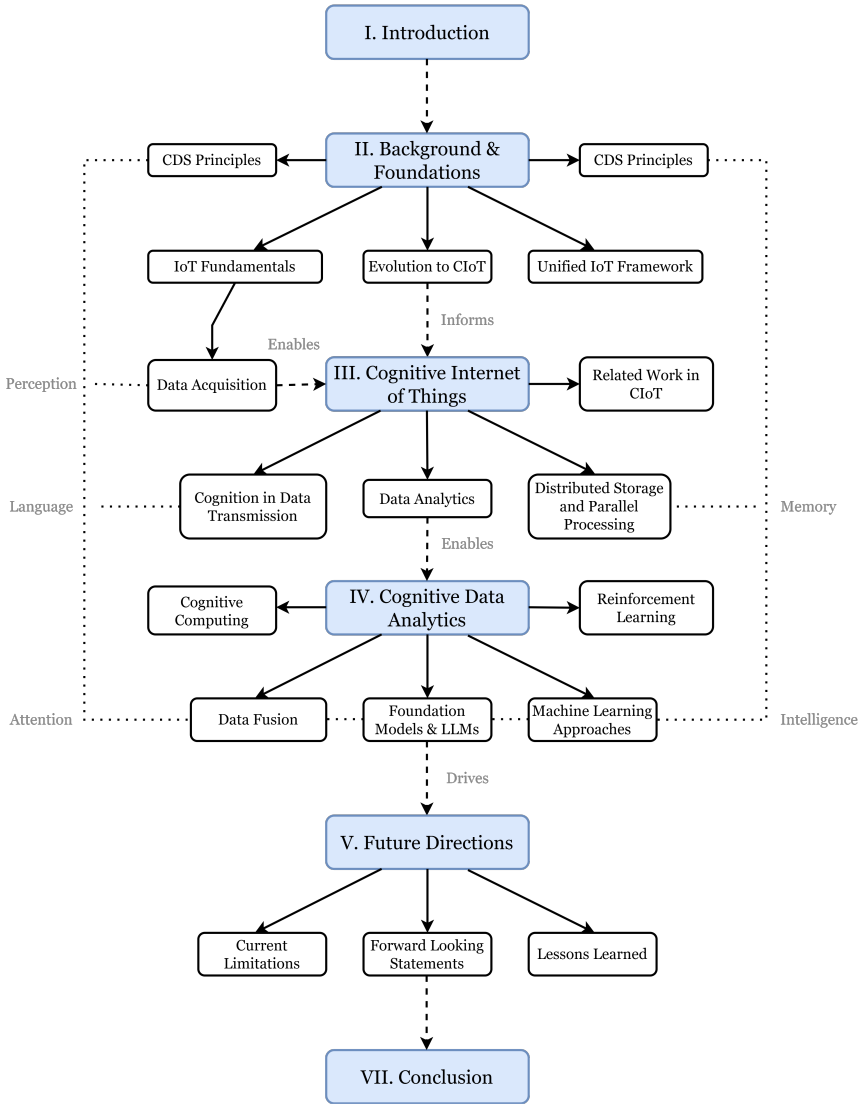


FIGURE 2.1: This flowchart organizes the survey structure, illustrating the interconnections between core topics and their subsections in the CIoT field. Beginning with foundational concepts like CDS principles and IoT fundamentals, it outlines the evolution towards CIoT, emphasizing the roles of perception (data acquisition), language (cognition in data transmission), memory (data storage and processing), attention and intelligence (cognitive data analytics, including data fusion and machine learning approaches). The framework culminates in future directions, addressing current limitations, lessons learned, and forward-looking statements, providing a comprehensive roadmap for understanding and advancing the CIoT paradigm. This flowchart uses various types of connections to represent different relationships: **solid arrows** indicate direct, hierarchical relationships or causal dependencies; **dotted arrows** represent indirect or inferred relationships, such as conceptual links or secondary influences; and **dotted lines without arrowheads** signify associative or parallel relationships between topics that coexist or share contextual relevance but lack a hierarchical or directional dependency.

The primary concern in scaling such systems lies in the limitations of current wireless technologies and mobile networks. Although advancements in smart host technologies and communication interfaces help alleviate bandwidth constraints in localized environments, significant challenges persist in broader, large-scale IoT deployments. Limited range, data bandwidth, and spectrum availability remain critical issues for large-scale CIoT systems [37]. These challenges are exacerbated in high-density scenarios, such as smart cities and industrial IoT ecosystems, where diverse devices interact across heterogeneous networks. Additionally, the number of devices that can be connected simultaneously is limited by the underlying technology. For instance, while LoRaWAN and ZigBee are effective for low-power, low-bandwidth use cases, they support fewer devices per gateway compared to cellular networks. Conversely, 5G offers massive device connectivity, but its deployment costs and spectrum availability may still restrict scalability in certain regions. Even when a large number of devices can be connected simultaneously, issues such as network congestion, latency, and quality of service (QoS) degradation can arise in scenarios with dense device clusters or highly dynamic communication requirements. Given the rapid growth of IoT and device usage, these issues could have cascading effects in the future.

Another major barrier is the wide variety of protocols used across IoT devices, which hampers seamless machine-to-machine (M2M) communication. IoT architectures are often tailored to specific applications due to the diverse domains in which IoT is deployed (e.g., smart vehicles, cities, grids, healthcare, transportation, manufacturing, and homes). While this specialization has been effective, it limits interoperability and hinders progress toward cross-domain architectures [38]. Efforts toward protocol standardization have focused on multiple layers, including the application layer (e.g., CoAP, MQTT), service discovery layer (e.g., mDNS, DNS-SD, uBonjo), and infrastructure layer (e.g., IEEE 802.15.4, 6LoWPAN, LoRaWAN, ZigBee). Although these initiatives have improved compatibility within specific domains, a unified solution remains elusive [39]. Cross-technology solutions [40], which harmonize communication across heterogeneous standards, offer a promising alternative by reducing reliance on full protocol unification. However, a definitive solution has yet to emerge.

Recently, cognitive radio (CR) and cognitive radio networks (CRN) have attracted significant attention from industry and academia as potential solutions to these challenges [37]. CR's core functions—spectrum sensing, decision-making, management, and mobility—aim to maximize licensed spectrum utilization through dynamic allocation to fill existing spectrum gaps. The applications and mechanisms of CR will be further

explored in Sections 2.2.1 and 2.3.

Another limitation in large-scale IoT systems is the difficulty in aggregating data that is diverse in nature. Sensory data in a multi-sensor IoT system can be highly heterogeneous, encompassing various data types such as numerical readings, categorical labels, images, or text. This heterogeneity necessitates the adaptation of data analytics techniques to effectively process and interpret the data. Additionally, the collected data can be nonlinear, high-dimensional, or incomplete, further complicating its application for intelligent decision-making and service provisioning [30, 33].

Cognitive computing stands as the intelligence pillar of CDSs, providing the adaptive decision-making and data analytics capabilities necessary for CIoT to cope with various data types, contexts, and environments. By leveraging machine learning algorithms and artificial intelligence, CIoT systems can dynamically select suitable processing methods based on the characteristics of incoming data—whether structured, unstructured, numerical, or categorical—and adjust to changing environmental conditions such as network constraints, resource availability, or hazards.

Moreover, CIoT systems incorporate contextual information (e.g., time, location, user behavior, and surrounding conditions) into their data analyses, enabling them to interpret information within the correct situational framework and generate more precise insights. This approach is further strengthened by techniques such as association analysis, clustering, and regression, which allow the system to identify patterns, group similar data, and predict future trends for intelligent, context-aware decision-making. Such decision-making is vital for big-data-driven applications that benefit from flexible, efficient, and self-optimized operations.

From a broader perspective, cognitive computing is more than just a collection of machine learning or data analytics methods. It is a cohesive framework composed of multiple processing subsystems that extract features and knowledge from the environment, feeding insights to a cognitive engine capable of enacting corrective actions and interacting with humans and machines in a dynamic and adaptive manner. CDSs, in turn, learn from the past, understand the present, and anticipate the future. By integrating this cognitive engine within a feedback control loop, CDSs update their knowledge over time and adapt to diverse environmental scenarios in real-time.

Within CIoT, both cognitive computing and CDSs are foundational. Cognitive computing imparts the computational intelligence for data processing and decision-making, while CDSs emphasize real-time adaptation driven by continuous feedback from the

environment. This synergy grants CIoT devices an elevated degree of autonomy and intelligence, allowing them to interpret large volumes of data, make informed decisions, and respond effectively to shifting conditions. As a result, these systems exhibit superior problem-solving capabilities, enhanced adaptability, and more efficient decision-making.

Advancements in architectures that integrate data analytics, cognitive computing, and multi-level intelligence for IoT systems will be explored in Section 2.4.

Furthermore, central data processing poses a substantial limitation for large-scale IoT systems. Single-node failure, limited scalability, and colossal exchange overhead are characteristic problems associated with the central data processing structure [33]. Considering these inhibiting factors, parallel and distributed data processing is a preferable solution for IoT systems, though it also presents challenges of its own and will be explored further in Section 2.3.

Surveys in the field have predominantly focused on the application of machine learning techniques in the context of CIoT for various components of IoT architectures. Examples include analyses of various model-based approaches and CR applied to wireless communication systems in IoT [41, 37, 42]. Others have focused on the use of machine learning for data processing and context information sharing in IoT systems [43, 44]. It has been identified that there has not been a comprehensive survey discussing cognition-integrated IoT systems in all its components and aspects. The objective of this paper is to further the discussion on CIoT making use of recent technological advances and expanding on the CIoT paradigm as introduced by Wu *et al.* [33]. The CIoT topic has been briefly explored in existing literature, but to the author's best knowledge, it has not yet been considered in relation to CDS [17, 19]. Furthermore, previous research fails to discuss relevant emerging technologies crucial to cognitive computing, such as the rise of LLMs and multimodal machine learning models. Utilizing these methods in conjunction with typically adopted technologies can potentially overcome current limitations in IoT and CIoT data processing. Moreover, existing literature on this topic does not address the concept of multi-level intelligence, introduced in this paper as a significant breakthrough for realizing CIoT's potential. The paper's contributions can be summarized as follows:

- The concept of cognition in engineering systems as defined by Haykin and Fuster is thoroughly examined, providing a novel framework for CIoT and highlighting its history and development.
- The CIoT paradigm is newly extended to modern enabling technologies, exploiting machine learning, data fusion techniques, and foundation models to enhance

context awareness and adaptability.

- Distributed learning through multi-level intelligence and federated learning is explored within the CIoT framework, providing innovative solutions to the scalability problems of IoT systems.
- The history of CIoT and its possible future directions are explored through an extensive review of the existing literature.
- Parallelisms are drawn to relate modern machine learning techniques and cognitive paradigms with the structure and functioning of brain operations in humans, providing further insights under a biomimicry lens.

This foundational understanding provides the basis for exploring the evolution of cognitive systems and their integration into IoT, which is detailed in the following sections. The remainder of this paper is structured as follows. Section 2.2 provides a comprehensive discussion on CDS and IoT, outlining their fundamental principles, key applications, and the role they play in modern technological ecosystems. Section 2.3 discusses the CIoT paradigm and how it is defined in the literature, including the communication and architectural design components of related applications specific to CIoT. Section 2.4 covers the fundamental data analytics aspects of CIoT, beginning with conventional IoT data analytics techniques discussed in the previous section and progressing to cognitive computing, multi-level intelligence, and large language models (LLMs). Section 2.5 encompasses the current limitations of CIoT systems, lessons learned, and future directions. Finally, Section 2.6 concludes the paper, summarizing the contents and key findings.

2.2 Background and Foundations

2.2.1 Cognitive Dynamic Systems

The refined CDS paradigm outlined by Haykin [45] builds on Fuster's five pillars of human cognition for engineering applications. A system can be considered cognitive if capable of performing five fundamental processes: the PAC, memory, attention, language, and intelligence. Applied to CIoT, the PAC refers to the cyclical process of using sensors to derive information describing the system's state and performance, where actions are initiated according to this information, affecting the environment and the system itself. This process is comparable to feedback control systems but uses advanced data analytics and relies on the other pillars to achieve intelligent decision-making. When multiple

PAC systems are employed in CIoT, system intelligence is significantly enhanced. Each PAC system can focus on different aspects of the environment or various subsystems, collecting specialized data and making localized decisions. By integrating the insights and actions of multiple PAC loops, the overall system benefits from distributed intelligence and collaborative decision-making. This multi-PAC architecture enables the system to address complex, large-scale challenges by leveraging diverse data sources and perspectives, leading to more robust and scalable intelligence.

Building on the PAC, memory is used to store relevant data about the environment, the system, and the previous actions taken to improve the response for future scenarios. As outlined below, memory can be divided into three separate components: perceptual, executive, and working memory.

- **Perceptual memory** stores the information extracted by the system's sensors, providing both long-term and short-term records of the collected data.
- **Executive memory** keeps track of the decisions undertaken by the cognitive controller to be used as a reference for future cognitive actions.
- **Working memory** couples the previous two components and records the outcome of the actions taken, which the system will then use to learn from and model future behavior.

In a machine-learning sense, executive memory refers to the model's capacity to store and utilize learned representations. It is primarily encoded in the form of the learned weights and biases of the model. These parameters are optimized during training to capture patterns in the data and are subsequently used to guide decision-making during inference. Executive memory is what enables a model to generalize from past experiences, allowing it to improve its responses to future inputs.

Perceptual memory, on the other hand, refers to the system's ability to retain and process sensory input or raw data from the environment. In machine learning, perceptual memory can be thought of as the input data that the model is exposed to over time, such as time-series data, images, or sensor readings. This memory represents the model's continuous interaction with its environment and how it 'perceives' new information. While executive memory deals with learned weights, perceptual memory is concerned with the representation and integration of incoming data, typically in the form of feature extraction or pre-processing steps.

Together, executive memory and perceptual memory form the working memory of a machine-learning model. Working memory represents the dynamic interaction between the model's learned knowledge (executive memory - past experiences) and the incoming data (perceptual memory - new experiences), enabling the model to handle new situations. Mathematically, this interaction occurs as the model processes new inputs by combining these inputs with the learned parameters, leading to predictions, decisions, or actions. It is not a static memory but rather a fluid process that combines feature extraction and parameter-based decision-making during inference. This interaction allows the model to generalize from previous data while adjusting to new inputs, resulting in a more adaptive and intelligent system. In practice, working memory is realized through the architecture of neural networks, where the incoming data flows through layers of the network, interacting with stored weights to produce outputs. The working memory of the model evolves during training and is constantly updated through learning algorithms, allowing the model to improve over time.

Attention is an extension on the PAC and memory components of CDS, enabling a cognitive system to optimize these processes by efficient data interpretation and resource assignment. In [32], it is defined as providing "the mechanism needed to prioritize the allocation of computational resources to mitigate the information overload problem." As such, actively filtering the processed information by relevance can facilitate learning and improvement of the cognitive controller. Attention in CDS is not represented by a physical state but manifests itself within the framework through an algorithmic mechanism. Modern attention mechanisms include self-attention and multi-head attention, both widely used in deep learning models like transformers. These mechanisms allow the model to dynamically focus on the most relevant parts of the input data by assigning different weights to different pieces of information, improving tasks such as natural language processing, machine translation, and image recognition.

Language in engineering systems is defined by communication protocols adopted by machines to communicate information to other system components. Cognitive systems should be able to adapt to any communication protocol to be able to exchange information, although efforts towards the standardization of such protocols in IoT aim to resolve this issue as well (i.e., interoperability). These systems should also be able to process natural language to receive instructions directly from humans in a seamless fashion. Foundation models are a promising enabling technology towards this goal and will be discussed in Section 2.4, which is focused on cognitive data analytics.

Intelligence bases itself on the previous four cognition traits (perceiving, memorizing, communicating, and adapting), incorporating them into an algorithmic mechanism capable of optimal decision-making. In the face of unpredictable circumstances, intelligent systems can understand the situation, enact corrective actions, and learn from the scenario's outcome in a Bayesian statistical fashion.

2.2.1.1 Cognitive Radio

Applying the pillars of CDS, CR aims to maintain highly reliable communications while efficiently utilizing the radio spectrum by analyzing the radio scene, identifying channels, and transmitting information using dynamic spectrum management. The receiver carries out the radio scene analysis, which senses the environment to discover spectrum holes, categorized as black and grey spaces, and estimates the interference temperature [1]. Correspondingly, this passive task involves processing nonstationary temporal signals to account for the spatial characteristics of RF stimuli, resorting to adaptive beamforming for inference control [1]. The continuous monitoring of the spectrum and the calculation of alternative routes to identified spectrum holes is vital to this system, providing redundancy when a primary user needs the spectrum for its own use. A graphical representation of the dynamic assessment of spectrum holes can be seen in Figure 2.2.

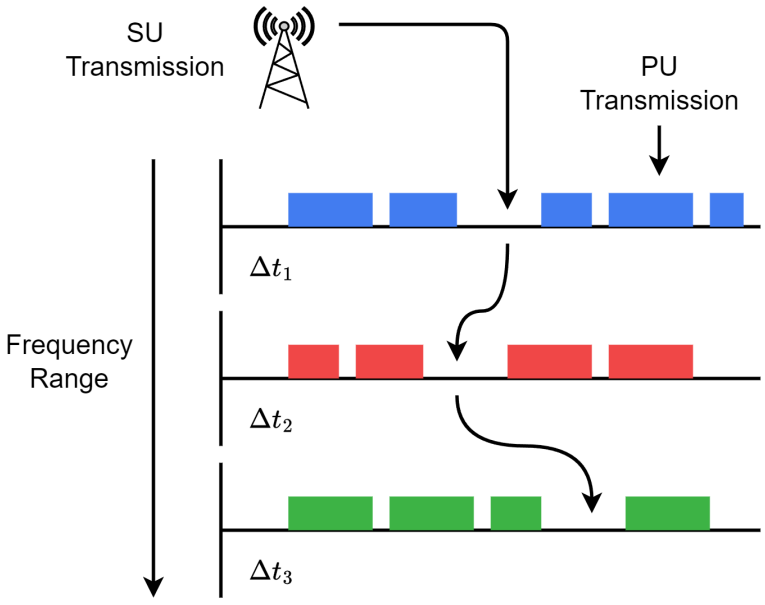


FIGURE 2.2: Dynamic spectrum transmission (adapted from [1])

The problem of channel state estimation is addressed using semi-blind training of the receiver, first introduced in [46], implementing a receiver with two operation modes: supervised training mode and tracking mode. The first mode uses a short training sequence to acquire and calculate the channel estimate. In contrast, the second is meant to be the operational mode and iteratively assesses the channel state. The calculations are carried out using a state-space model of the channel parameters, with process and measurement equations, assuming linearity. Autoregressive (AR) coefficients, dynamic noise, and measurement noise are addressed by selecting an appropriate tracking strategy and filter selection.

Spectrum licenses represent a cost for companies; therefore, the deployment of large-scale heterogeneous networks composed of multiple objects can sum up to be significant capital expenditures. The application of a CRN can reduce this cost by efficient utilization of the available spectrum. The size of the data to be shared is another limitation of large-scale IoT adoption, depending on the application under consideration and the type of sensors deployed. The introduction of edge computing could provide a solution to this, reducing the amount of data to be transmitted by filtering the data and partially processing it on the edge layer [47, 48, 49, 50]. The integration of CR and edge computing as an essential need for CIoT will be discussed more in Section 2.3.2.

Beyond IoT applications, the theoretical advancement of CR has seen useful applications in a variety of technical fields over the past decade. Villardi *et al.* [51] apply CR for emergency broadcasting operating in the television white space. With their portable cognitive emergency service network (PCEN) and novel fractional service area metric, they determine that a relatively small amount of channels is sufficient to ensure emergency services, minimizing area degradation. Ferrus *et al.* propose the application of CR for public service networks in similar critical public protection and disaster relief (PPDR) situations, developing an overview of spectrum-sharing models to mitigate issues as a result of limited network capacity [52]. In their comprehensive survey, Joshi *et al.* [53] outline several potential areas where CR can be implemented for wireless sensor networks, with security applications (e.g., military or public), health care, home appliances, transportation, and surveillance. CR's utility in healthcare is especially prevalent, due to interference-vulnerable patient data and critical communications transmitted across the Wireless Medical Telemetry Services band. CR has been proposed as a solution to this problem by several, offering reduced network load with increased QoS and safety of transmission [54]. Further efforts in CR have been developed recently in this sector considering IoT aspects of health monitoring [55] and security of CR [56].

CR has also seen extensive use in autonomous instrument communication. For unmanned aerial vehicles (UAVs), CR provides an alternative mode of transmission for increasing general communication performance, security, energy efficiency, and addressing spectrum scarcity. Santana *et al.* [57] discusses this application, providing a review of challenges, IoT, and prospective applications. Similarly, the application of CR can alleviate some of these same concerns in satellite communication, using its software-defined radio concept to adapt transmission parameters and demonstrate flexibility against obsolescence [58]. The authors of [59] propose CR and rate splitting multiple access for the improvement of low Earth orbit communication (SatCom), increasing spectral efficiency for massively connected systems, where Li *et al.* [60] apply CR for integrated satellite-terrestrial communication towards 6G mobile networks. Further, several authors propose to implement CR into smart grid communication frameworks, improving QoS, traffic scheduling, and implementing backup protection [61, 62].

2.2.1.2 Cognitive Control

Fatemi *et al.* [63] extend upon the CDS framework, describing a new way of thinking about cognitive control, focusing mainly on two essential components: learning and planning. Each of these are built upon the following fundamental notions.

- The two-state model is composed of the target state, or target of interest, and the entropic state of the perceptor. The entropic state can be viewed as a measure of the lack of sufficient data in the cyclic flow of information from the global PAC. Mathematically it is represented by a state-space model of the environment defined by a process and measurement equations. Alternatively, the entropic state can be modeled using an entropy-based model, where entropy quantifies the amount of uncertainty or disorder within the system. In this context, a high entropy state indicates insufficient or noisy data, which impacts the model's ability to accurately track or predict the target state. By incorporating these two states, the model balances the interaction between uncertainty (entropy) and the target of interest, enabling better decision-making and control.
- The first principle of cognition, the global PAC, which in this context is a cyclic directed flow of information from the environment.

The goal of cognitive control, through learning and planning, is to optimize cognitive policy, specifically the probability distribution of cognitive actions, including the influence of previous actions on current states. Shannon's entropy concept is used to describe

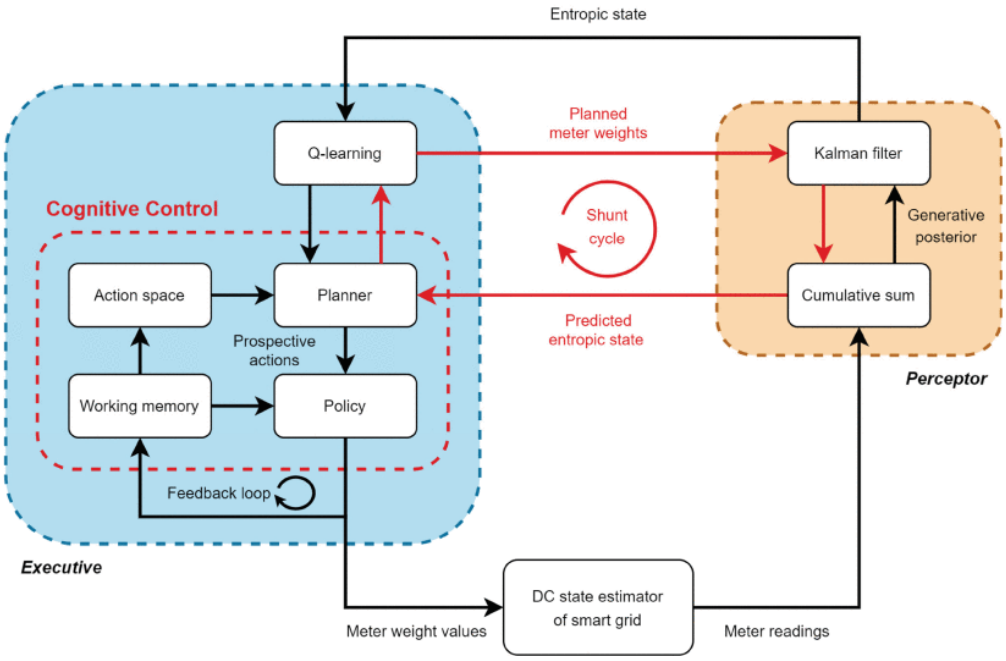


FIGURE 2.3: Cognitive Control Diagram (adopted from [2])

the current state of the perceptor, quantifying the noise present as a distribution of collected data. The system modifies the entropic state through incremental deviations, formalized by an immediate rewarding process at the end of each cycle, attempting to predict the future entropic state of the system and use it in the planning phase of the cycle [63]. The algorithm that converges to the optimal policy is the core of the cognitive controller functionality, as illustrated in 2.3. It is derived and demonstrated as a particular case of dynamic programming, inheriting the basic properties of such convergence and optimality. To accelerate this property, the authors also describe a mixed strategy of pure explore (selecting actions randomly) and pure exploit (selecting actions based on maximum value criterion) called the ϵ-greedy strategy adopted from Powell *et al.* [64]. The implementation of this trade-off may be viewed as a facilitator for attention. As previously touched on, allocating computational resources is vital to continuously improve the knowledge of the environment without falling into local sub-optimal solutions [63].

A computational experiment is presented in their seminal work by applying this new concept of the cognitive controller to a radar tracking problem. The cognitive controller adapts the system's variables to improve the estimation of the object's position, velocity, and ballistic coefficients. Dynamically changing the radar waveform resulted

in an improvement of four orders of magnitude compared to the fixed waveform radar. The field of cognitive control has since been further developed for a variety of applications, focused in robotics and networked systems. Wang *et al.* adopted this technology for communication-based train control (CBTC) systems, utilizing the entropic state to quantify the communication between trains and ground [65]. Using their reinforcement learning (RL)-based decision maker, the cognitive control approach significantly improved the CBTC performance and efficiency through the compensation of channel fading and packet loss. Fatemi later applied cognitive control to optimize complex sensor networks, maximizing information [66]. For mobile robots utilizing networked control systems (NCS), Wang *et al.* [67] apply a Q-learning-based cognitive control framework with a backstepping controller to detect and compensate for random packet dropouts. The entropic state is used to characterize this phenomenon, induced through the application of a wireless network as an intermediate feedback medium. This idea is further extended to 3 DoF robotic manipulators [68] and compensating for additional time delay [69] with robust and event-triggered control schemes.

2.2.1.3 Cognitive Risk Control

In the presence of unexpected uncertainty, cognitive control lacks a mechanism of predictive adaptation to adverse events or obstacles, commonly referred to as risk. In [70], the authors take this concept a step further, recognizing the need for a subsystem capable of interacting with the various aspects of CDS such as the perceptor, working memory, and executive memory to predictively adapt the system to a new uncertain environment. This newly defined subsystem uses a Bayesian filtering mechanism and Bayesian generative model to guide the CDS through timely risk-avoiding actions [70]. Under this generative model, the posterior is computed on the current iteration based on previous actions, where there might be several iterations for each PAC cycle.

The objective of the Bayesian filtering process is to capture relevant information from the generative model and reject irrelevant information, jointly improving the relevant information fed to the entropic information processor. This process is referred to as top-down attention. A shunt cycle is also defined to bring bottom-up attention from the planner to the RL algorithm, resulting in local feedback between the two. The internal rewards process results from the entropic state calculation and feeds into the executive for RL and the task switch control. To characterize different situation, the internal rewards are either always positive under the assumption that the physical system is free of uncertainties or consistently negative under the presence of uncertainties. The rewards

are computed and transformed by the RL algorithm in a value-to-go function that constitutes the input to the cognitive controller. The value-to-go function is influenced by the actions space (containing all hypothesized actions), internal rewards, discount factor (weight assigned to discount previous actions exponentially), and the policy (the action taken at the immediately preceding PAC).

The cognitive controller is composed of the planner (which schedules the possible prospective actions) and the policy (the function that leads to decision-making). Under uncertainty, the risk-sensitive cognitive actions are selected by a classifier responsible for decision-making. Given N past experiences, the classifier assigns the posterior to past perturbed cognitive actions recorded in the executive memory.

Furthermore, task switch control is introduced to prevent the perturbed cognitive actions from affecting the executive memory. This correlates directly with the double nature of the internal reward systems, such that pre-adaptation is achieved by successfully classifying events that occurred under risk-sensitive uncertain environments and non. A pair of switches is used to direct the flow of information to different sections in the CDS framework, requiring further analysis if perturbed cognitive actions are necessary [70], illustrated in 2.4. Cognitive control is particularly relevant in the IoT domain, as the application of optimal policy selection are of primary importance in the management of such systems.

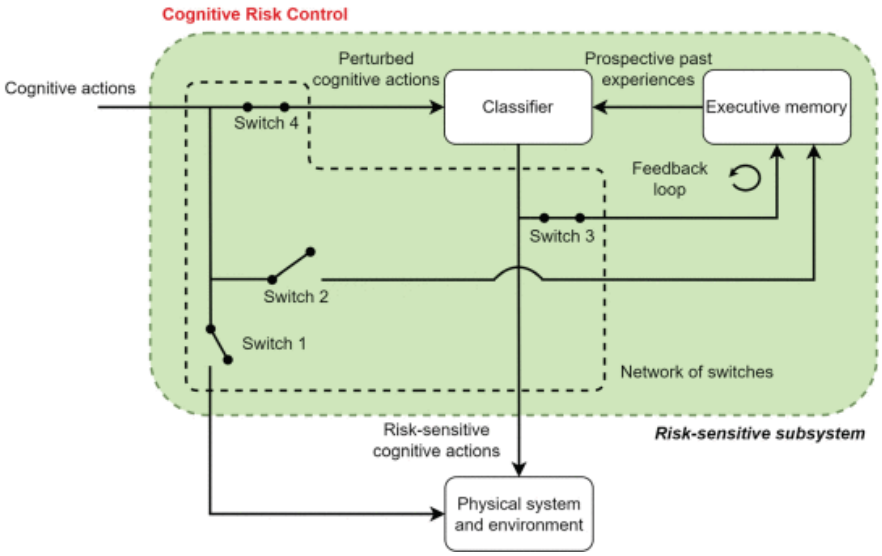


FIGURE 2.4: Cognitive risk control switching mechanism (adopted from [2])

Building on the cognitive principles discussed, the next section explores how these concepts relate to the IoT paradigm through cognition.

2.2.2 Internet of Things

2.2.2.1 Overview of Standard IoT Components

IoT refers to a network of interconnected devices that utilize sensors and actuators to gather and exchange data over the Internet. These systems comprise several key components that operate in harmony to support automation, data collection, and remote control across diverse applications.

At the heart of every IoT system are the devices themselves, which include a diverse array of sensors, actuators, and embedded systems. These components act as a bridge between the physical and digital worlds, enabling the acquisition and manipulation of environmental data to execute meaningful actions [71].

Sensors Sensors are devices that detect events or changes in the environment or the system and send the information to other electronics, typically a computer processor. They gather data from various sources, such as:

- **Temperature Sensors**: Measure heat energy to detect temperature changes, crucial for climate control systems and weather monitoring.
- **Humidity Sensors**: Monitor moisture levels in the air, important for environmental monitoring and agricultural applications.
- **Motion Sensors**: Detect movement or vibrations, used in security systems, automated lighting, and smart appliances.
- **Light Sensors**: Sense ambient light levels, enabling automatic brightness adjustment in devices and energy-saving lighting systems.
- **Pressure Sensors**: Measure pressure in gases or liquids, essential in industrial processes and fluid dynamics.
- **Proximity Sensors**: Detect the presence of objects without physical contact, utilized in touchless interfaces and obstacle detection.
- **Gas and Chemical Sensors**: Identify the presence of gases or chemicals, vital for air quality monitoring and hazardous material detection.

- **Accelerometers and Gyroscopes**: Measure acceleration and orientation, used in mobile devices, drones, and wearable technology.

These sensors convert physical parameters into electrical signals that can be processed and analyzed. They are fundamental in collecting real-time data, which forms the basis for informed decision-making in IoT applications.

Actuators Actuators are devices that take electrical input and convert it into physical action, allowing IoT systems to interact with the environment. They perform actions based on received commands, such as:

- **Electric Motors**: Provide rotational movement, used in robotics, conveyor belts, and adjustable components.
- **Solenoids**: Generate linear motion, applicable in locking mechanisms and valve controls.
- **Servos**: Offer precise control of angular or linear position, velocity, and acceleration, essential in robotics and automated systems.
- **Heaters and Coolers**: Regulate temperature, important in HVAC systems and thermal management.
- **LEDs and Display Units**: Provide visual feedback or illumination, used in user interfaces and signaling.
- **Speakers and Buzzers**: Deliver audio output for alerts and communication.

Actuators enable IoT devices to affect changes in the physical world, executing tasks such as adjusting environmental conditions, controlling machinery, or providing feedback to users.

Embedded Systems Embedded systems integrate sensors and actuators with processing units and communication interfaces to execute tasks autonomously. They are specialized computing systems that perform dedicated functions within larger systems[71]. Key features include:

- **Microcontrollers and Microprocessors**: Serve as the brains of the device, processing data from sensors and sending commands to actuators. Examples include Arduino, Raspberry Pi, and ESP32.

- **Memory and Storage**: Store firmware, operating systems, and data collected from sensors.
- **Communication Modules**: Facilitate connectivity through Wi-Fi, Bluetooth, Zigbee, LoRaWAN, or cellular networks, enabling data exchange with other devices and cloud services.
- **Power Management Units**: Manage energy consumption efficiently, crucial for battery-powered and remote devices.
- **Input/Output Interfaces**: Allow interaction with other hardware components, such as analog/digital converters and serial communication ports.

Embedded systems are designed to be resource-efficient and reliable, often operating under real-time constraints. They execute programmed instructions to perform specific tasks, from simple control functions to complex data processing.

Integration of Components The integration of sensors, actuators, and embedded systems allows IoT devices to function intelligently. For example:

- **Smart thermostats** use temperature sensors to monitor room temperature, processing the data with embedded algorithms, and controlling heating or cooling systems through actuators to maintain desired settings.
- **Automated irrigation systems** employ soil moisture sensors to assess hydration levels and activate water valves via actuators to irrigate crops as needed.
- **Wearable health monitors** collect biometric data like heart rate and activity levels using sensors, process the information to track health metrics, and can alert users or healthcare providers if anomalies are detected.

These examples demonstrate how the synergy of sensors, actuators, and embedded systems enables IoT devices to perform complex tasks autonomously, improving efficiency, safety, and user experience across various domains.

Communication Protocols Communication between IoT devices and networks is facilitated through various protocols, each designed to meet specific requirements regarding data transmission, energy efficiency, and connectivity [71]. Standard IoT protocols involve:

- **MQTT (Message Queuing Telemetry Transport)**: A lightweight messaging protocol ideal for constrained devices and low-bandwidth networks.
- **CoAP (Constrained Application Protocol)**: Designed for use with resource-constrained devices, allowing them to communicate over the internet.
- **HTTP (HyperText Transfer Protocol)**: A widely-used protocol for transmitting hypermedia documents, commonly used in web services.
- **BLE (Bluetooth Low Energy)**: Enables wireless communication over short distances with low energy consumption.

Network Architecture IoT devices connect through layered network architectures that include edge computing, fog computing, and cloud services. Edge computing brings data processing closer to the data source, reducing latency and bandwidth usage. Fog computing extends cloud computing to the network edge, providing intermediate processing. Cloud services offer scalable storage and computational resources for data analytics and application deployment [71, 39].

Security and Privacy Security is a critical aspect of IoT systems due to potential vulnerabilities arising from interconnected devices. As these devices communicate and share data, they can become targets for cyber threats, making it essential to implement robust security measures to protect data integrity and user privacy [72].

Standard security measures include:

- **Encryption Protocols**: Utilizing encryption methods like TLS/SSL to secure data transmission and prevent unauthorized access to sensitive information.
- **Authentication Mechanisms**: Implementing strong authentication processes to verify the identities of users and devices, reducing the risk of unauthorized access.
- **Secure Communication Channels**: Establishing protected channels for data exchange to safeguard against interception and tampering.

By incorporating these security measures, IoT systems mitigate risks associated with data breaches and unauthorized control, ensuring that both system functionality and user privacy are maintained [72].

2.2.2.2 Unified Internet of Things Framework

The development of novel IoT frameworks and architectures has been the subject of research in both industry and academia, due to the variety of IoT applications. However, due to the heterogeneity of the various domains within IoT, these architectures vary in components, functionalities, and often in terminology and protocols used. To add, no holistic architecture of IoT or CIoT exists at present, valuing global goals over local ones within the system. The different scopes to which the architectures are developed have resulted in limited interoperability between the systems, effectively hampering the development of a cross-domain holistic architecture. This discrepancy prompted the development of IoT-A and IoT-I, ETSI M2M, FI-WARE, AIOTI-HLA, and ITU-T, which are large-scale projects focused on designing a comprehensive IoT framework as described in [38].

Even governing bodies such as the European Union (EU) under multiple technical commissions have attempted to create a solution for this problem. Various programs, such as the EU FP7, allowed thousands of organizations to receive EU funding. In addition to the EU efforts, a global initiative under the European Telecommunications Standards Institute (ETSI) auspices was initiated to define a standard structure for the service layer and M2M communication. The architecture proposed under the ETSI initiative focuses on two specific domains, the gateway domain and then the network domain, trying to establish standard protocols and application programming interfaces (APIs) based on the proposed architecture.

Furthermore, several technical committees, working groups, and standardization organizations are working on the standardization of communication protocols, interference control, and data storage. Such groups include EPC Global and ITU SG13, SG16 technical groups, STF 396, and CEN TC 225 for informational security and data privacy, ISO and EMA committees for spectrum usage and radio frequency identification (RFID), as well as IEEE, 3GPP, and IP for smart objects [73]. Moreover, the Federal Communication Commission (FCC) is considering the use of CR's dynamic spectrum access for very high frequency (VHF) and ultra-high frequency (UHF) bands, contributing to its popularization and likelihood for widespread adoption. The generic enablers for offering reusable shared functions serving multiple service areas include cloud hosting, data and context management, applications and services, IoT service enablement, interfaces to networks and devices, and security [38].

Primarily, IoT-A aims to offer IoT architecture developers a common technical ground

to optimize interoperability and avoid IoT architectures being built as stand-alone silos. ARM is used as a structure for common technical ground in designing new architectures. It consists of three interconnected parts involved in developing an IoT framework. The first part is the IoT Reference Model (RM), which provides a set of models used to define architectural views [38]. This framework also outlines a taxonomy of IoT concepts and provides an information model, communication model, functional model, and security, trust, and privacy models. The second part is the Reference Architecture (RA), which uses the views and perspectives of various stakeholders to analyze and address design problems. The last part is guidance, which defines the process that will lead to creating a concrete architecture based on RM and RA [38]. These initiatives will foster new architectures based on standard design techniques, especially on the FP7 IoT projects, like IoT-A ARM, effectively enabling straightforward repurposing of results, functionalities, and components [73].

A significant challenge in the IoT domain remains the development of standardized communication protocols capable of handling massive device-to-device transmissions. However, recent advancements in smart host technologies have mitigated the need for extensive, distributed inter-device communication. These smart hosts act as centralized gateways, connecting devices across multiple network types and efficiently managing data flows. By aggregating and forwarding data to relevant entities or services while translating between different protocol requirements, smart hosts reduce network congestion and simplify interoperability by limiting the number of direct communication pathways each device must support.

Furthermore, cross-technology communication solutions, such as those discussed in [74], enable seamless interaction among heterogeneous devices operating under diverse standards and protocols. By introducing an abstraction layer between disparate communication mechanisms, these solutions eliminate the need for rigid, unified protocols, which can be difficult to deploy across fragmented IoT ecosystems. Instead, interoperability platforms, protocol translation layers, and context-aware gateway solutions collectively create a flexible and adaptable infrastructure that simplifies integration among a wide array of devices.

As IoT systems architectures continue to evolve, the need for enhanced intelligence and autonomy has led to the emergence of CIoT. With these theoretical foundations, having defined cognition in the context of engineering artificial systems, explored what IoT is, and examined its current limitations, we now delve into CIoT. We analyze how

CIoT has been conceptualized through the lens of CDS, exploring its key components, applications, and future directions.

2.3 Cognitive Internet of Things

The International Telecommunication Union (ITU) has documented a four-stage program for future IoT development in smart IoT or CIoT [73]. This program suggests that future IoT architectures integrate key capabilities: service sensing, data sensing, environment sensing, and intelligent cognitive abilities. CIoT is still a relatively young field, but it is gaining popularity following research efforts in CDS and cognitive computing. It can be considered a framework for developing novel IoT architectures, addressing specific issues within the IoT domain by equipping systems with a *cognitive layer* that enables them to learn, think, and understand both the physical and social worlds [33]. The successful and widespread application of CIoT allows for the integration of many disciplines and fields, such as computer science, mathematics, neuroscience, and engineering. The adaptive features of CIoT allow it to be deployed across various domains and industries, bridging the physical and cyber worlds to facilitate smart resource allocation, automatic network operation, and intelligent service provisioning.

Modern autonomous systems and self-operating technologies are enhanced by disruptive innovations such as artificial intelligence, machine learning, and robotics, which augment human capabilities by performing tasks without direct human intervention. These advancements have enabled researchers to work towards a future where *human-machine interaction* and *interdependence*, are increasingly prevalent.

Many definitions have been proposed to describe future generation systems capable of full autonomy, such as CDS and Symbiotic Autonomous Systems (SAS). CDS and SAS share many similarities [75]; they both aim to enhance the autonomy and capabilities of synthetic machines, focusing on embedding IoT systems with the ability to evolve, adapt, and learn. A crucial aspect of this evolution is the interplay and bridging between the physical and cyber spaces [76]. These new technologies aim to create heterogeneous and synergized IoT structures capable of autonomous behavior through *collective intelligence* [77].

This will give rise to new hybrid societies where the symbiosis between humans and smart machines is an integral part of every aspect of life. This vision entails that machines will be capable of full autonomy in decision-making, exploration, goal setting,

and replication [77], finding applications in transportation, healthcare, consultancy, education, manufacturing, and more [78, 79, 77].

By building upon the foundation of standard IoT components, CIoT introduces a cognitive layer that empowers IoT systems with artificial intelligence and machine learning capabilities. This integration enables systems to not only perform predefined functions but also learn, adapt, and make autonomous decisions, thereby enhancing their effectiveness and efficiency in complex environments.

2.3.1 Related Work in CIoT

Wu *et al.* argue that only being connected is not enough. IoT systems should be capable of learning, thinking, and understanding the physical and social world themselves, empowering them with "high-level intelligence" [33]. The paper develops a new paradigm, the CIoT, based on the foundations laid down by Haykin and Fuster. The article proposes a new operational framework built upon the interactions among five fundamental cognitive tasks: the PAC, massive data analytics, semantic derivation, knowledge discovery, intelligent decision-making, and on-demand service provisioning [33] and shown in Figure 2.5. The authors present a new network paradigm in which physical/virtual objects are interconnected and behave as agents with minimum human intervention by bridging the physical, digital, and social world, while also enabling smart resource allocation, automatic network operation, and intelligent service provisioning.

From a bottom-up point of view the system can be divided into four layers:

- The sensing control layer, directly related to the global PAC, interfaces directly with the environment by processing incoming stimuli and feedback observations.
- The semantic knowledge layer, related to semantic and ontological derivations, further processes the data to enable context-awareness.
- The decision-making layer uses the knowledge abstracted from the previous layer to reason, plan and select the most suitable action for the interacting agents to implement.
- The service evaluation layer assesses the provisioned services and feedback evaluation through novel performance metrics related to the social world.

The paper thoroughly presents the challenges of processing nonlinear, high-dimensional, and heterogeneous data before discussing steps for intelligent decision-making. Decision-making is generally characterized by two components planning and selecting. Within

the decision-making process focus is placed on the choosing an action from the action set based on collected data and inferred information motivated by the learning ability in CRNs. Cognitive selection is defined as the ability to adjust based on historical and current data intelligently. There are three kinds of cognitive selecting highlighted in the paper: Markovian decision process, multi-armed bandit, and multiagent learning.

Given the expected large number of decision-makers in a distributed CIoT architecture, the authors focus on game theory, and investigating the learning approach with uncertain, dynamic, and incomplete information [33]. Precisely, non-cooperative game models fit the problem, characterizing the interactions among single decision-makers, in which each player maximizes their own utility function. The main concerns in developing such a system is the convergence to desirable stable solutions. In large-scale CIoT systems, further challenges arise in the context of local interactions between agents and spatial game models. Although global information exchange is unrealistic in a classical large-scale CIoT system, local interactions among agents are possible through regional cooperation, leading to near-optimal results.

The system's performance evaluation is a complicated task in CIoT and is simplified by categorizing the metrics in two dimensions, cost, and profit. In the profit dimension, three main metrics are discussed. The quality of data (QoD) metric evaluates the data acquisition process and the quality of sensed data. In addition, the QoD should be able to quantify the data completeness, truthfulness, and currentness. The second metric is the quality of information (QoI), which represents the amount of valuable data the decision-maker obtains for a specific task based on precision, accuracy recall, and quantity. These details characterize the quality of the information quantity provided to the decision-maker. Third, quality of experience (QoE) is the last metric used in the profit dimension and evaluates user experience based on access, stable operation, efficiency, and user application [33]. On the other hand, the cost dimension metrics proposed in the paper are device utilization efficiency, computational efficiency, energy efficiency, and storage efficiency.

2.3.1.1 Energy Management and Resource Optimization

Energy efficiency is a cornerstone of scalable CIoT systems, particularly in resource-constrained environments. Several studies have explored how cognitive capabilities can address this challenge by leveraging predictive modeling and optimization techniques. Following the works of Wu *et al.* the CIoT paradigm started to become popular in the literature, inspiring further studies in the subject. Braten *et al.* built a testbed

to explore autonomous resource management and autonomous learning processes using machine learning, framed in the CIoT paradigm [80]. Arguing that the optimal control of an IoT system cannot be obtained through centralized device management, Braten *et al.* also proposed the autonomous orchestration to be carried out by a cognitive device manager. Specifically, the problem considered was energy management, a prominent issue in resource constrained IoT systems. The cognitive manager was tasked with handling the collection of meta-data, triggering the machine learning process and planning the energy consumption based on external factors. Adaptation was implemented in the system through predictive model selection to mitigate the bootstrapping problem of predicting prior to have collected enough data for reliable training [81]. The work was expanded in subsequent publications, fusing the data collected with contextual data, which aggregated solar intake data with weather forecast data [82]. Furthermore, the authors recently published a structured review of IoT device management and cognitive model in which adaptation mechanisms for IoT management are discussed in detail [83]. The emphasis when explaining cognitive architecture is placed on the adaptation mechanism and the separation of knowledge to be processed by different computational models, depending on case scenario and data type [83].

In the context of bridging the cyber-physical world and human experience, smart homes provide an excellent ground to analyze the potential of CIoT as a people centric IoT, enhancing quality of life by intelligently adapting to the living environment. Furthermore, the growing presence of smart objects in houses makes it possible to bring cognition to modern-day smart homes.

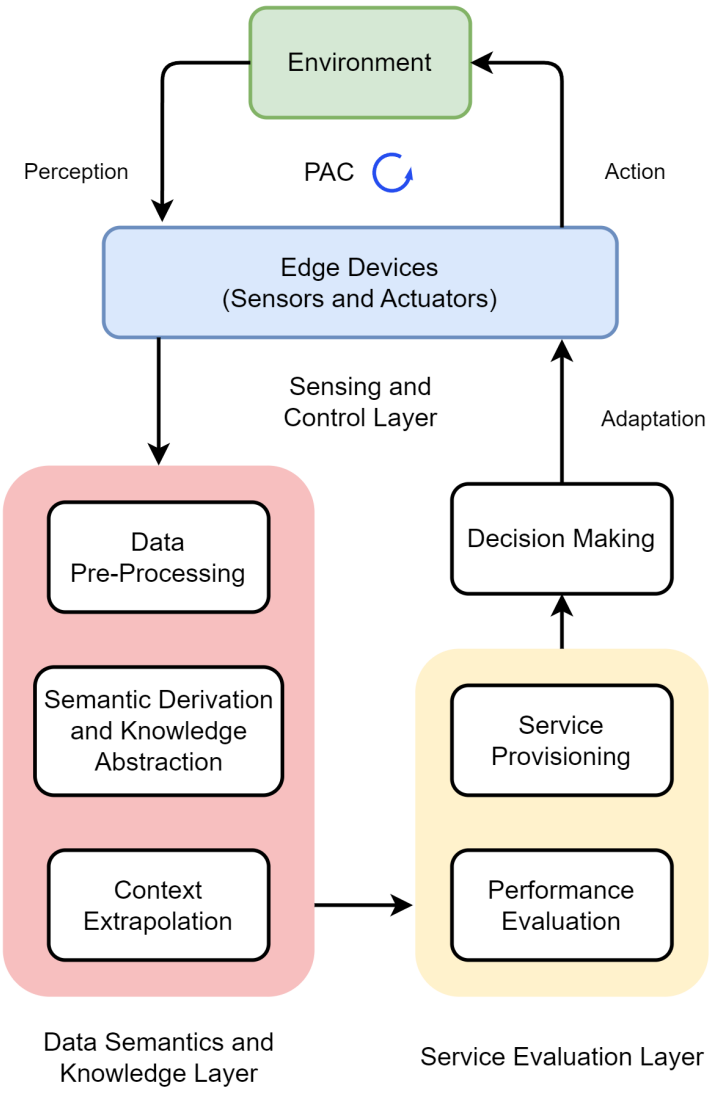


FIGURE 2.5: CIoT Structure as defined by Wu *et al.*, adapted to fit CDS conceptual framework.

2.3.1.2 Smart Home and Personalized Adaptation

Smart homes provide an ideal testbed for CIoT technologies, showcasing how systems can adapt to user behaviors and preferences to enhance comfort and energy efficiency.

Feng *et al.* introduce the use of CDS in smart home scenarios [84]. The authors use the "falling asleep problem" as an example scenario of the functioning of the smart home on a day-to-day basis. In this scenario, a person is gradually falling asleep on the couch of their house. With the aim to maximize the user's comfort and ensure a good sleeping environment, the home can recognize the event taking place and adjust its variables to such conditions. Based on the information perceived, for example, the house could lower the TV volume or gradually turn it off while modifying the room temperature to an optimal temperature suitable for sleeping and changing the shape of the couch to a laid-back position resembling a bed.

The paper bases itself on the previously mentioned five pillars of cognition. It uses a Bayesian filter coupled with a modified RL algorithm in its structure, similar to the cognitive control architecture outlined in Section 2.2.1. Moreover, attention plays a crucial role in the CIoT smart home scenario; for example, referencing the "falling asleep problem", the CDS may wrongfully interpret actions such as movements while sleeping and interpret them as a sign the person is waking up [84]. Therefore, the CDS should be able to use various sensors present in the house to understand the behavioral pattern of the resident, predicting the time interval over which the person will not move and adjust accordingly. The sensors used for this purpose could be, for example, a pressure sensor installed directly into the couch and a temperature sensor present in the room. The authors define three types of cognitive actions in the context of an intelligent home scenario [84].

1. Cognitive actions applied to the environment to affect people's perception process.
2. Cognitive actions applied to the system itself to reconfigure sensors and actuators.
3. Cognitive actions are applied as part of the state control actions, therefore modifying the environment to decrease the information gap.

This architecture leverages the same set of sensors already used in smart homes, enhancing the entire system to establish a comfortable and efficient living environment. By utilizing the perceptual capabilities through the executive part of the CDS equipped with attention and intelligence, it adapts to the occupants' lifestyle preferences and habits. This approach not only improves operational efficiency, such as optimizing energy

consumption and resource usage but also enriches the residents' lifestyle by providing personalized and responsive services that align with their daily routines and comfort needs.

The management of thermal comfort in smart homes has also been explored by Serianni *et al.* [85]. This research does not explicitly follow the definition used in this paper for CDS and CIoT (that is, make use of all of Haykin's pillars of cognition, such as attention and language). Still, it proposes a different architecture based on message queuing telemetry transport (MQTT) and neural networks to elaborate suggested action prediction and anomaly detection based on the user's habits. An ANN is used to actively control the HVAC system to maintain an optimal living temperature in the house environment while also reducing energy consumption costs, training the ANN directly on the user's comfort habits. The system's performance was evaluated based on the analysis of the interaction between the user and the HVAC system used to manage thermal conform in the home environment and showed promising results.

Expanding the scope beyond individual homes, researchers have also investigated how similar strategies can be applied to larger-scale environments. Rinaldi *et al.* propose a framework for cognitive buildings, involving user preferences and recurring patterns to be used for learning and planning day-to-day management [86]. The goal is to create a building capable of maintaining responsive environments and adaptable indoor spaces optimizing energy consumption and enhancing user comfort. The authors showcase the results obtained using the prototypes of the ELISIR installation, using Schneider Electric hardware for home automation and the KONNEX (KNX) protocol. The experiment focused on the use of smart windows and smart manifold for indoor environmental management and natural resource management [86].

2.3.1.3 Smart Cities and Infrastructure

The success of CDS and CIoT for smart home applications in previous works demonstrates the applicability of this technology for subtle prediction and decision-making based on human or machine patterns captured with sensor data. Naturally, it can be suggested that this property can be scaled to smart city frameworks, encompassing numerous simultaneous "falling asleep problems" across large distances. For instance, larger-scale modern grids must accommodate load demand variability of renewable resources and employ advanced metering infrastructure (AMI) at smaller scales to collect data for pattern recognition, enhancing distribution efficiency. These frameworks must also consider large-scale interoperability, cybersecurity, and other risks linked to digital

communications, such as QoS. Such issues are discussed at length by Gunduz and Das [87] and Song *et al.* [88]. As an example, the authors of [89] explore the predictive component of CDS for security, detecting bad data and false data injection attacks in IEEE bus systems.

In a broader context, the principles of CIoT can be extended to numerous domains within smart cities. In [90] Feng *et al.* outline a case study using the CDS framework for the Internet of Vehicles (IoV) in smart cities, asserting that the modernization of the transportation system will offer great potential to prevent traffic, vehicle collisions and reduce commute costs. Connected autonomous vehicles (CAVs) are necessary for such purposes since they're capable of adjusting their actions based on perceived environmental information. The article further expands this definition to RACE vehicles (risk-sensitive, autonomous, and connected electrical vehicles) to follow the recent trends in the adoption of electric cars. The deployment of such large-scale CAV networks would benefit private and public transportation, while also opening the door for possible cyber-attacks. Before diving into the CDS framework for smart vehicles, the authors consider the cyber threats that such networks could be exposed to and propose measures that must be put in place to ensure the availability, integrity, and confidentiality of the system. This analysis considers active attacks such as jamming, binding, false data injection (FDI) attacks, and passive attacks such as eavesdropping and stalking [90]. Given the complex, dynamic, and adversarial environment CAV operate in, the addition of CDS as an active supervisor of all subsystems present in a car is desirable to enhance the risk control mitigation through joint interoperability and adaptation. The operational sensors such as LIDAR, video cameras, radio receivers, and radar receivers could be actively adapted to the situation based on the context extrapolation capabilities of CRC, improving their functionality. To achieve this goal, the authors propose an upgraded CRC framework based on the Bayesian generative model and entropic information processing, using the task switch control mechanism to control operational mode depending on the situation. The executive part of CDS would be composed of the RL portion and planner, action library, policy, classifier, working memory, and executive memory to estimate the best cognitive action or policy based on the adapted and filtered feedback information.

Vlacheas *et al.* bring this vision further, proposing a cognitive management framework that interconnects various smart city systems [91]. The paper defines VOs as the virtual representation of any real-world object and cognitive virtual objects (CVOs) as

a set of interoperable semantic VOs. Cognition is used for self-management and self-configuration of VOs using a concept called proximity to define the interoperability of such objects. Proximity is defined as the level of relatedness between any IoT application and the relevant objects that could be used to deliver the desired outcome. Semantic extrapolation is mentioned in the paper as a requirement for high-level description, registration, discovery, and access invocation of the cognitive objects and processes. The request and situation matching (RSM) algorithm is introduced to match potential existing CVOs to be used in any given situation or to task the creation of new CVOs in the case no existing CVOs are suitable. The scenario utilized to describe the operational framework conceptually is the autonomous trigger of medical intervention based on personal health monitoring devices. In this scenario, an elderly woman suddenly has a heart attack recognized by the heart monitoring device she wears. The device triggers an immediate response to an autonomous driving ambulance directed to the scene based on the city traffic monitoring system for faster response. The interoperation of these IoT subsystems shows the potential for this CIoT framework to act autonomously with minimal human intervention, improving the efficiency of associated services and dynamic provisioning [91].

Another architecture proposed by Park *et al.* is CIoT-Net, demonstrating that the various domains in a smart city share similar sets of cognitive data and make use multiple cognitive computing-based applications [92]. Such architecture requires the least number of separate configurations intended for different applications and bases itself on five layers, the smart city platform, IoT layer, data layer, cognitive computing layer, and service layer. For example, the smart city platform detailed in the article consists of multiple sublayers for each IoT domain present in a smart city, such as intelligent buildings, smart homes, transportation, agriculture, industry, and SGs [92]. A visual representation of the architecture can be seen in Figure 2.6 and includes the various subsystems present in each architecture layer. Section 2.4 will discuss the arguments tackled in this paper regarding the management of heterogeneous and high-dimensional data in more detail.

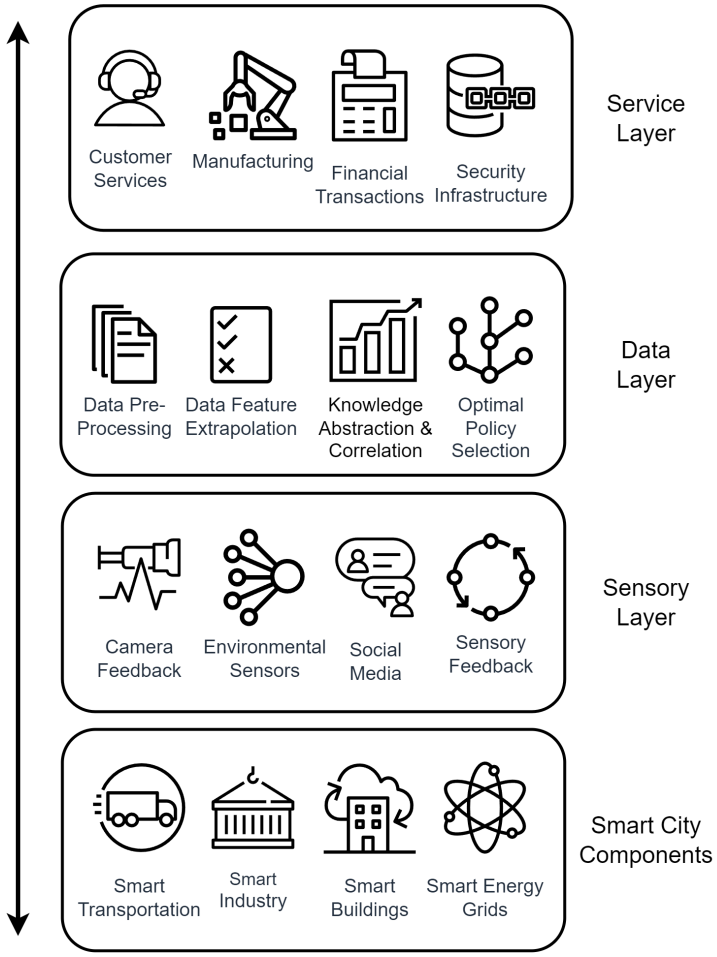


FIGURE 2.6: CIoT layers in smart city scenario, bottom-up view starting from hardware, sensors, data processing, and finally service implementation based on the previous layers.

Within smart cities, the energy infrastructure is a crucial aspect of the realization of sustainable growth. The deployment of SGs has become prevalent with the integration of smart meters for energy monitoring. Moreover, it enables the gathering of urban informatics to improve energy services' availability, efficiency, reliability, economics, and sustainability.

Pranaya *et al.* propose a similar architecture to the one proposed by Park *et al.* consisting of a perception layer, attention-memory layer, and decision layer as a possible solution for energy conservation, thereby reducing operational cost, and exploiting data analytics [93]. In the paper, the authors break down SGs into three layers. The user end layer encompasses smart homes, smart vehicles, and renewable energy, capturing data by RFID, cameras, and environmental sensors. The communication and network layer encompasses the data communication, management, storage, servers, Cloud, transmission means such as GSM, LTE, cable broadband, and private networks. The power generation layer encompasses power generation, transmission, and distribution based on the previous two layers. Uplink and connectivity of can be further enhanced by combining energy harvesting techniques and CR. Considering the challenges in spectrum scarcity and saturated conditions present in crowded cities, energy harvesting could be a reasonable alternative to batteries, with applications in SGs [94, 95, 96]. Guo *et al.* tackle the transmission optimization of energy and spectrum efficiency using a deep reinforcement learning (DRL) approach. The optimization problem is modelled as an incompletely known Markov decision process (MDP) without complete non causal a priori knowledge. The results show that the proposed deep deterministic policy gradient model converges and performs better than random transmission policy, myopic transmission policy, and deep Q RL algorithm [97].

CIoT has also been applied to energy-saving, recognizing users' mobility habits, and understanding such variable patterns with the goal of reducing the unneeded energy consumption of the device during operation. In addition, exploiting location and environmental information derived from smartphones can provide valuable insights into the functioning of a smart city but at the cost of reducing the device's battery life due to the continuous usage of power-hungry sensors such as GPS receivers. The trade-off between accuracy and energy consumption is a relevant problem in the battery management of smartphones. Torres *et al.* tackle this problem by applying CDS to identify mobility states, aiming to reduce the sampling rate in a process that resembles the definition we outlined earlier for attention [98]. In this work, the authors show that

TABLE 2.1: Relevant papers in CIoT, classified by application area and the use of the 5 pillars of cognition: perception (P), attention (A), memory (M), intelligence (I), and language, which in this context is represented by CR.

Year	Authors	Reference	Applications	P	A	M	I	CR
2006	Haykin	[30]	Theoretical	✓		✓		✓
2011	Haykin	[32]	Theoretical	✓	✓	✓	✓	✓
2014	Wu	[33]	Theoretical	✓	✓	✓	✓	✓
2014	Fatemi	[63]	Control	✓	✓	✓	✓	
2016	Sheth	[34]	Theoretical	✓			✓	
2017	Haykin	[70]	Risk Control	✓	✓	✓	✓	
2017	Al-Turjman	[100]	Networks	✓		✓	✓	✓
2017	Khan	[37]	IoT	✓	✓	✓	✓	✓
2017	Feng	[84]	Smart Home	✓	✓	✓	✓	
2018	Ploennigs	[101]	Smart Buildings	✓	✓	✓	✓	
2017	Braten	[80]	IoT	✓		✓	✓	
2018	Braten	[81]	IoT	✓		✓	✓	
2019	Park	[92]	Smart City	✓		✓	✓	
2019	Feng	[90]	Smart City	✓	✓	✓	✓	✓
2019	Zhang	[102]	IoT	✓		✓	✓	
2020	Foukalas	[103]	Industrial IoT	✓		✓	✓	
2020	Li	[73]	IoT	✓		✓	✓	✓
2021	Braten	[83]	IoT	✓		✓	✓	

applying a CDS approach to the user mobility problem effectively extracted features describing human mobility for mobile mining without strongly compromising the device's energy consumption. The authors further refined these concepts in another subsequent publication, which can be found in [99], proposing a cognitive controller with event-based processing for energy efficiency and spatio-temporal accuracy in mobility-based and location-based sensing.

Furthermore, cognitive computing and CIoT have been proposed as possible frameworks for smart crowd management within a smart city environment. Crowd monitoring has been a difficult problem to tackle due to the variety of behavioral traits that can be extracted from a crowd [104]. For example, identifying abnormal movements, group formations, or sudden dispersals can be challenging in large gatherings. The aim is to develop smart surveillance systems, operating within the framework of a smart city with minimal human intervention, that can effectively monitor events such as concerts, protests, or sporting events, and automatically raise alarms when suspicious or dangerous activities are detected. Varghese *et al.* also tackles the issue noting how cognitive

computing itself is not enough to accomplish this task, the need for distributed processing through edge/fog computing must also be considered due to the high latency and high bandwidth usage of cloud servers for video processing [105].Table 2.1 summarizes the relevant literature by categorizing key contributions in CIoT according to their application areas and the integration of the five cognitive pillars—perception (P), attention (A), memory (M), intelligence (I), and language (CR)—providing a clear overview of the field's evolution.

Additional related work that falls outside the scope of this literature review can be found in the following papers regarding smart homes [106, 84, 85], astronomy [107, 108], industrial applications [103, 73, 109], smart cities and more [110, 111].

Although much progress has been made in the abstraction of CIoT architectures, multiple research gaps have yet to be addressed. The role of distributed processing in the creation of architectures leveraging new machine learning methods and the potential of new LLMs and multimodal models have yet to be addressed in the CIoT context. These models could potentially provide CIoT systems with the ability to understand, interpret, generate, and translate human language, and pair it with contextual data presented in different modalities. This will enable more natural and intuitive interactions between humans and IoT devices. Thus, there is a pressing need for more research to bridge these gaps, which could potentially unlock new capabilities and applications for CIoT systems. Furthermore, the current literature fails to bridge the theoretical and practical aspects of CIoT, while this manuscript emphasizes enabling technologies capable of bringing the idea to fruition while further expanding it.

2.3.2 Communication Components in CIoT

The rapid expansion of the IoT, combined with the emergence of intelligent systems, is driving an ever-growing demand for spectrum and wireless bandwidth. This trend is expected to intensify as global deployments of massive CPSs increase, thereby exacerbating bandwidth constraints in localized areas. For instance, Khan *et al.* [37] argues that semantic-oriented IoT applications are unfeasible without the integration of CR technologies. Recent advances—including cross-technology communication (CTC) frameworks and smart host systems—have partially mitigated these local capacity bottlenecks. For example, modern smart control centers can seamlessly connect over 500 devices using technologies such as PLC, Wi-Fi, and BLE within a single apartment, illustrating that careful system design and emerging standards can ease the pressures imposed by high-density IoT environments.

A major challenge for deploying advanced CIoT communication components is ensuring their compatibility with legacy IoT systems. Many currently deployed devices rely on outdated protocols or proprietary standards, which hinders seamless integration with modern solutions such as cross-technology interfaces, CR frameworks, and smart host systems. Overcoming this barrier requires the development of backward-compatible protocols, modular gateways that translate legacy communications into contemporary formats, and middleware that standardizes data exchange. These measures not only reduce the technical debt associated with older systems but also facilitate a gradual transition toward scalable, interoperable CIoT infrastructures.

Currently, most IoT devices operate on unlicensed spectrum bands (e.g., WiFi and Bluetooth). However, as trillions of devices are expected to connect in the near future [42], spectrum congestion will likely intensify. One promising solution is the efficient sharing of licensed 4G and 5G spectrum resources. By integrating CR with machine learning and advanced signal processing, IoT networks can substantially expand their transmission capacity. As discussed in Section 2.2.1.1, CR addresses issues such as collisions and excessive contention in wireless access networks, enhancing the automation, scalability, reliability, energy efficiency, and QoS of networked communications. Moreover, CR enables dynamic spectrum allocation and management, thereby improving accessibility, usability, adaptability, and interconnectivity within IoT networks.

2.3.2.1 Approaches in Cognitive Radio

CR strategies can be broadly categorized into two approaches:

- **Efficient Networking:** This approach focuses on spectrum management and spectrum-aware optimization to improve QoS. By carefully managing limited frequency resources, efficient networking helps prevent overcrowding and overlapping communications.
- **Flexible Networking:** This approach emphasizes environment discovery, self-organization, reconfigurability, and the coexistence of nodes to dynamically adapt to changing network conditions.

The CIoT paradigm envisions a system of interconnected physical and virtual objects—acting as interdependent agents—that require high levels of interoperability. Social network analysis offers one method for addressing this challenge by examining the interrelationships among these objects. Parameters such as analysis relations, ties, multiplexing, and composition can be explored through ego networks (focusing on a single

node and its neighbors) or whole networks (illustrating the interactions among all nodes) [35].

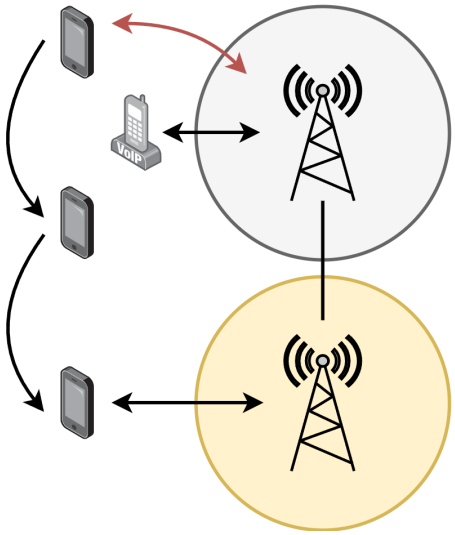


FIGURE 2.7: Routing cooperation among secondary users. Black indicates a direct connection, while red shows that a device is out of range.

Enhancing network performance can also be achieved by leveraging node proximity to mitigate channel impairments such as fading. In cooperative models, nodes relay signals for one another effectively forming a virtual antenna array and harnessing spatial diversity. In [112], two paradigms are discussed: (1) cooperation between primary users (PUs) and secondary users (SUs) and (2) cooperation among SUs. In the first paradigm, PUs with weak transmission links can partner with SUs, who are then incentivized with additional transmission time, access to frequency bands, or licensed spectrum relaying. In the second paradigm, SUs enhance their channel gains through cooperative diversity, relaying each other's signals to provide multiple transmission paths, reducing the risk of deep fades, and improving energy efficiency by lowering transmission power. Figure 2.7 illustrates secondary user cooperation through multi-hop routing, while Figure 2.8 provides an overview of CR integration within CIoT.

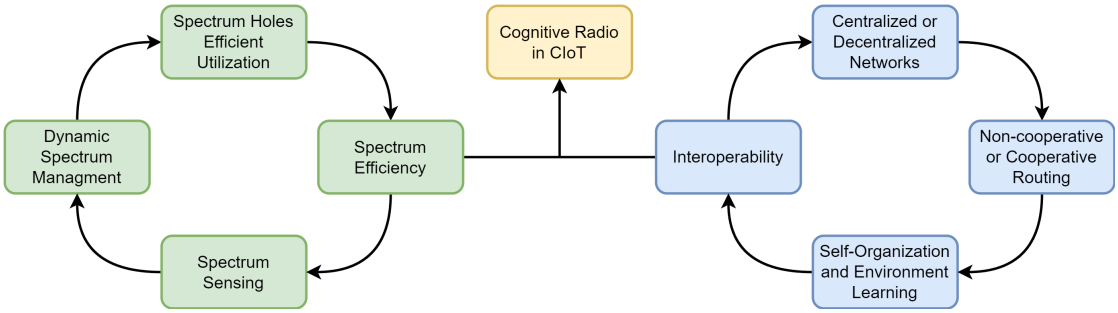


FIGURE 2.8: CR for CIoT. Left: CR elements; Right: decentralized IoT architectures forming the theoretical framework of future CIoT communication components. Arrows indicate the interdependence and continuous cooperation required in CR-CIoT applications.

Both centralized and decentralized networks rely on robust routing protocols. However, traditional single-hop and multi-hop routing algorithms often overlook functionalities introduced by CR, such as dynamic spectrum allocation. Various CR network types have been studied, including CR mesh networks (semi-static) and CR ad hoc networks (adaptable and self-reconfiguring through peer-to-peer communication) [42]. In centralized architectures, a common infrastructure or cluster leader collects spectrum information (e.g., interference zones) and disseminates it to optimize metrics like delay, hop count, energy consumption, channel availability, and route stability. These networks often support single-hop, device-to-device (D2D) connections when within the transmitter's range.

In contrast, distributed architectures typically utilize multi-hop, single-transceiver CR routing protocols [42], where route requests, destination estimates, and spectrum characteristics are propagated among neighboring nodes. Although this method may face scalability challenges, one proposed solution is to map spectrum characteristics to routing delay [112], thereby intentionally delaying routing requests to avoid transient spectrum constraints.

Media access control (MAC) protocols also present challenges in CR networks. They must effectively manage available channels while avoiding collisions with PU transmissions. In dense networks, coordinating a silent period is essential for detecting PU activity, and MAC protocols must be capable of halting transmissions upon collision detection. Time-slotted schemes using rendezvous and backup channels have been proposed [42], along with OFDM-based CRN approaches. For instance, [113] introduces

a contract theory-based model that negotiates cooperation between unlicensed IoT devices (UIDs) and SUs. In this model, negotiations between PUs, UIDs, and SUs are likened to a labor market, where resources, payments, or reputation serve as incentives to encourage spectrum sharing in OFDM-based CRNs.

Gateways, which connect IoT objects and virtual objects (VOs)—such as sensors or digital twins—to the Internet, must be flexible, scalable, secure, and energy efficient. In [114], the authors propose incorporating cognitive analytics and machine learning to enhance IoT performance via cognitive gateways. These gateways classify applications based on their computational and traffic requirements into four types:

- **Type A:** Low traffic, minimal data preprocessing.
- **Type B:** High traffic, minimal data preprocessing.
- **Type C:** Low traffic, substantial data preprocessing.
- **Type D:** High traffic, substantial data preprocessing.

Once an application is classified, the gateway uses a multi-objective optimization scheme to decide whether the service should reside in the cloud or at the fog/edge layer. Such cross-layer techniques not only enhance overall efficiency but also improve performance and QoS. For example, traditional Transmission Control Protocol (TCP) may underperform in CR environments, where unpredictable channel availability and frequent handovers can be misinterpreted as network congestion—leading to unnecessary timeouts and backoff procedures. Short-lived disconnections, due to spectrum sensing or PU-SU handovers, further impact throughput and reliability.

2.3.3 Decentralized Systems and Big Data

Multiple solutions have been proposed in recent years to address the challenge of big data processing. The increasing amount of data collected across domains of IoT applications can be computationally challenging to manage, store and process. Data preprocessing and analytic algorithms can require high amounts of power and time to run on large datasets, increasing the delay of the feedback information flow and increasing the cost of deployment [115]. Hence, posing limitations for applicability and efficiency. Depending on the application, these limitations can be decisive in implementing IoT and CIoT architectures. For instance, real-time data analysis requires data to be transmitted and analyzed at high frequencies, which is not achievable through conventional methods at the edge nodes if the processing is computationally expensive (e.g., anomaly detection

[116, 117]). Since most microcontrollers and CPS have limited hardware and battery capabilities, implementing computationally expensive algorithms such as on a large-scale artificial neural network (ANN) on such devices is often not an option.

Cloud computing technology has been a prominent solution for companies and private users to outsource the hardware and systems required to store, manage, and process data, allowing data to be uploaded to third-party remote servers. Infrastructure as a service (IaaS) and platform as a service (PaaS) are becoming more popular as companies' digitalization expands further, providing clients with a prepackaged system, including hardware and software solutions [118]. Among these providers, there are the biggest tech companies in the world, such as Google Cloud, Microsoft Azure, and Amazon web services, competing for the market share of this new lucrative market.

Cloud storage and processing partially resolve the energy and computational power problems needed to run classification, clustering, and other algorithms based on data, but still present challenges in integrating IoT systems. Mainly, scalability issues arise in continuous, uninterrupted big data transmission to the cloud servers. Depending on the application, aspects such as latency and availability can be crucial to the operation of the IoT system [119]. Moreover, big data processing can be time-consuming, even for the most powerful hardware components. To cope with these issues, three main requirements for CIoT with significant data processing and storage are needed in cloud computing. These requirements are outlined by Cai *et al.* as distributed execution (parallel processing), multitenant storage (distributed file systems), and flexible scalability [120].

Edge computing is an alternative solution for the dynamic management and processing of big data. IoT architectures based on edge computing can vary depending on the domain. Fog computing is the combination of edge computing and cloud computing and is often implemented in the form of cloudlets [18]. Cloudlets are highly virtualized data centers placed closer to the network's edge. Edge and fog computing rely on decentralized server clusters located in the proximity of the nodes, and mobile edge computing (MEC), which utilizes the computational power of mobile devices and CPS for data processing [119]. These architectures are generally more effective, processing the data with low latency, higher efficiency, and agility, but are however constrained by hardware capabilities and cost. For example, several machine learning algorithms need to run on compelling hardware because of their complexity and random-access memory (RAM) requirements, especially deep neural networks and convolutional neural networks.

Furthermore, in a CIoT framework, context-aware algorithms require even more data to be transmitted and processed. To this end, a combination of edge servers and cloud computing can be used to implement data processing. As a result, CIoT architectures can actively manage data flow optimally depending on the application, reducing latency, and selecting relevant data to upload to the cloud. Cloud computing and edge computing are complementary, compensating for each other's limitations. Edge computing ensures availability and low latency since it is physically closer to the end-user, being performed either on the device or a local server. In contrast, cloud computing provides the hardware needed to run more computationally expensive data processing and a redundant distributed storage system.

Zhang *et al.* introduce the discourse of edge intelligence in CIoT systems, asserting how edge computing could enable CIoT to effectively assist humans in real-time through customized service based on the perception cycle of IoT systems [102]. Furthermore, decentralized control mechanisms are crucial for prompt control in smart buildings and cities to achieve flexible communication and scalable computing in such environments [102]. The paper also lists sample services and applications of edge intelligence in CIoT, such as content sharing in vehicular networks, predictive analytics in healthcare, content-oriented caching on edge, and data-driven public bicycle service [102].

Chen *et al.* propose a cognitive computing and edge computing-based healthcare system [121]. The article highlights issues in the current architectures regarding inefficient resource allocation, inflexible network resource deployment, high latency of cloud-based systems, and the limitations of conventional machine learning methods in discovering hidden values and correlations in data [121]. The system employs edge cognitive computing to carry out a comprehensive analysis of the user's physical health data and network resources. Using a cognitive data engine and a cognitive resource engine, the architecture can carry out extensive data analysis involving external and internal data as well as metadata such as network type, service data flow, communication quality and other dynamic environmental parameters to assess the status of the network resource environment [121]. In addition, the article emphasizes the importance of distributed architectures for efficient resource allocation by virtualizing physical infrastructure resources into multiple parallel virtual network slices mutually independent. This handover strategy aims to provide seamless resource connection. The proposed edge cognitive computing system (ECC) comprises three main layers. The first layer is the user side or data-collection layer, composed of CPS such as smart clothing, mobile phones, and portable health monitoring devices, which collect real-time physiological data of the patient [121]. Examples

of data extracted from these devices include electrocardiography (ECG), electromyography (EMG), heartbeat, temperature, and blood oxygen saturation (SpO2). The second layer is the computing-analysis layer or edge computing side, composed of the computing devices of the user computing. The third and last layer is the storage-management layer or cloud platform side, which stores the basic and medical information of the patient and is managed and operated by the hospital [121].

The management of large-scale, heterogeneous IoT architectures is a major challenge and a critical consideration in designing future generation CIoT systems. Device management addresses the orchestration of resources in dynamic spatio-temporal dimensions allowing adaptation in mutable environmental conditions [83]. Perhaps adaptation is again a key component and fundamental characteristic of CIoT and future autonomous systems, which manifests in this case with self-adaptation and self-management of computational resources. Much like Haykin, Braten *et al.* identify three distinct adaptation mechanisms: reasoning, learning, and planning. These manifest in concrete situations through linear and nonlinear programming, probabilistic analysis, fuzzy logic, dynamic programming and recursive optimization, rule-based inference, ontologies, knowledge graphs, case-based reasoning, machine learning, and RL [83].

Moreover, many algorithms have been created specifically to orchestrate and manage existing resources in a network efficiently. Overload, congestion, and low-load resource management challenges are all addressed by workload balance optimization techniques. There are several in the literature, including Gaussian process regression for fog-cloud allocation (GPRFCA) [122], DRL-based resource allocation (DRAM) [123, 124, 125], prediction-based resource allocation algorithm (PBRA) [126], hybrid tabu-based simulated annealing (HTSA) optimization algorithm [127] and economic resource allocation (ERA) [128]. A comparative analysis between these methods can be found in [129]. Depending on the domain to which the IoT architecture will be applied, different optimization methods can be used to create an application-specific cognitive controller. The latency and orchestration need, in the proposed architecture, should be addressed using a combination of resource provisioning and job scheduling techniques. Task scheduling techniques such as fog sync differential algorithm (FSYNC) and Reed-Solomon fog sync (RS-FSYNC) should be able to supplement the design of a comprehensive cognitive controller in addition to any of the stated algorithms [130].

Jalali *et al.* propose an automatic task sharing and switching between cloud and edge/fog computing using machine learning [114]. The specific classification model used by the authors was the support vector machine, where its supervised learning was based

on available CPU, memory, bandwidth, and remaining battery, on top of other factors such as the required execution time and accuracy [114].

Chen *et al.* also proposes a dynamic service migration mechanism ECC based on dynamic systems [131], expanding on previous work [132]. The architecture uses elastic storage and computing services, resource allocation, and user mobility. ECC and the proposed mechanism aim to reduce network load, improve network efficiency, reduce transmission delay or latency, and improve the QoS. This mechanism differentiates from other methods because it considers user or actor behavior. It can decide when to migrate the service from one edge node to another based on behavior and mobility patterns. An optimization problem is then formulated based on migration cost, which depends on the server's capacity, the bandwidth available, migration goal, and desired service resolution based on user demands, mobility, and dynamic network resources. The same author also proposed a DRL approach to solve the resource allocation problem and which is discussed in [133, 134].

2.3.3.1 Tiny Machine Learning

Tiny machine learning (TinyML) is a recent paradigm that tries to integrate machine learning models into hardware with limited computational capability and resource constraints [135]. Next-generation computational sensing systems will be able to reduce the data burden while improving accuracy, bringing the computing closer to the environment in a distributed manner [136]. The results of the processing done on sensing hardware can also be leveraged to create reconfigurable sensor systems, capable of adapting their settings based on a group consensus of the current environmental state in which they operate. At the same time these systems could relay this information to abstract contextual information regarding the environment. Applications of these intelligent multi-sensor systems extend from environmental monitoring to wearable devices.

The integration of TinyML into consumer electronics has revolutionized smart home applications, particularly in smart TVs and voice-controlled systems. Recent advancements in speech recognition have enabled on-device, low-latency voice processing, enhancing user interaction with smart devices. One notable implementation is in speech recognition devices that determine a user's intention based on the direction of the sound source, as described in a patent by Samsung Electronics [137]. This system employs multi-microphone beamforming and deep learning-based direction estimation to improve speech recognition accuracy in noisy environments, allowing smart TVs and AI assistants

to identify which user is speaking and filter out background noise effectively. Additionally, TinyML has been utilized for enhancing media recommendation systems in smart TVs, as demonstrated in a patent outlining an energy-efficient neural network for content personalization and display optimization [138]. These edge AI solutions reduce reliance on cloud-based processing, improving privacy, responsiveness, and power efficiency, which are crucial for embedded AI applications. As TinyML continues to evolve, its deployment in consumer electronics will expand, enabling adaptive, context-aware interactions in real-time while minimizing computational overhead.

Recent advancements in TinyML have focused on optimizing deep learning inference for highly resource-constrained edge devices by leveraging model compression, optimization techniques, and specialized hardware accelerators. Deep learning models, which are typically computationally expensive and memory-intensive, pose significant challenges for deployment on low-power microcontrollers and IoT devices. To address these limitations, various model compression techniques such as pruning, quantization, and knowledge distillation have been employed to reduce model size and inference latency while maintaining high accuracy. Architectures such as MobileNet [139], ShuffleNet [140], and SqueezeNet [141] utilize depthwise separable convolutions and grouped convolutions to minimize the computational footprint of convolutional neural networks (CNNs). Moreover, hardware-aware deep learning optimizations, including algorithm-hardware codesign, have facilitated the development of specialized AI accelerators such as Edge TPUs, FPGAs, and reconfigurable ASICs, enabling efficient execution of deep learning models with minimal power consumption. These advancements are crucial for real-time TinyML applications in domains such as healthcare monitoring, industrial IoT, and autonomous systems, where low-latency, energy-efficient processing is essential. Furthermore, the integration of neural architecture search (NAS) has enabled automated model design tailored for edge devices, ensuring a balance between performance, energy efficiency, and computational feasibility. As TinyML continues to evolve, the synergy between model compression, hardware-aware optimizations, and emerging AI accelerators will drive the next generation of ultra-low-power deep learning applications [142].

Furthermore, tiny transfer learning (TinyTL) has been proposed to train small neural networks on resource-constrained devices with minimal memory usage and loss of accuracy from bigger models. Therefore, optimizing the learning process of application-specific models on edge devices [143].

Further CIoT deployment examples could be a set of environmental sensors capable of adjusting where and how to sense based on a classification algorithm [136], to then

transmit the results of such classification to give context to an unsupervised monitoring system. A concrete application example was implemented by Veiga *et al.*, which applied the CIoT framework to a person counting system in skiing areas through camera sensors. The idea was limiting the amount of information transmitted for processing to save energy in constrained situations, dissecting the image into tiles and only relaying the most relevant ones using a planning algorithm and an attention model [144].

2.3.3.2 Multi-Level Intelligence

While TinyML and edge AI optimizations have significantly improved deep learning inference on resource-constrained devices, there are still scenarios where local processing remains infeasible due to memory, computation, and power limitations. In such cases, hybrid cloud-edge architectures are employed, where edge devices perform partial computations while offloading resource-intensive tasks to cloud servers. According to Shuvo *et al.* [142], three primary hybrid inference strategies have emerged:

- Edge-Server Inference: Raw data from IoT sensors and smart devices is transmitted to edge servers, which store deep learning models and perform inference before sending back results. This approach reduces bandwidth requirements compared to full cloud processing while improving real-time responsiveness.
- Edge-Device Inference: Pretrained lightweight models are deployed directly on resource-constrained devices, ensuring low-latency execution. However, this requires substantial compression, quantization, and model optimization to fit within the hardware constraints.
- Collaborative Inference: Deep learning models are partitioned between the edge device and an edge server or cloud. Early-stage feature extraction occurs locally, and intermediate activations are transmitted for final processing in the cloud. This approach balances computational efficiency and real-time performance.

Notable applications of hybrid cloud-edge architectures include:

- Real-time health monitoring in intensive care units (ICUs): Edge AI processes physiological signals locally for quick anomaly detection, while cloud servers handle deeper analysis for long-term health trends.
- Autonomous driving systems: Vehicles perform immediate perception tasks onboard (e.g., obstacle detection), but offload complex route planning and map updates to cloud infrastructure.

- Industrial IoT and predictive maintenance: Embedded sensors in manufacturing environments run local machine learning models for real-time fault detection, while cloud-based analytics refine predictions and optimize long-term maintenance schedules.

These hybrid approaches ensure deep learning solutions can be deployed effectively across diverse real-world applications by balancing energy efficiency, latency, privacy, and computational feasibility.

In addition to protocol-based advancements, alternative communication methods have emerged to address challenges in scalability and efficiency. For instance, instead of relying on raw data transmission, IoT devices can transmit extracted representations or latent features derived from local processing. This approach not only minimizes bandwidth usage but also enhances privacy and security by reducing the exposure of raw data. Such representation-based communication is particularly valuable in scenarios involving edge computing or federated learning, where only the essential compressed or encoded information is transmitted, ensuring efficient resource utilization.

The integration of data processing capabilities can be implemented in multiple levels of a CIoT architecture, to decrease the magnitude of the information flow and transmission overhead. Complex abstractions can be carried out by implementing lightweight machine learning models such as regression, support vector or small-sized neural networks. The results of the processing can be used either directly from the node to adapt itself or in combination with a bigger overarching model architecture that resides higher in the hierarchy of computational hardware through compressive sensing [136]. By transmitting encoded information to a pre-trained decoder residing on the cloud or the edge the security and transmission size of the data can be improved, as shown in Figure 2.9. The transmission of encoded data will furthermore provide a security layer to some known IoT attacking techniques such as eavesdropping, also known as man-in-the-middle attacks.

Compressive sensing techniques further improve efficiency by encoding data before transmission, allowing a pre-trained decoder at the cloud or edge to reconstruct the information with minimal loss [142, 145, 146]. This method not only reduces transmission size but also enhances security against known IoT attack techniques such as eavesdropping, often referred to as man-in-the-middle attacks. Privacy-preserving AI methodologies, such as federated learning and homomorphic encryption, are also being explored to maintain security in multi-level intelligence frameworks [142].

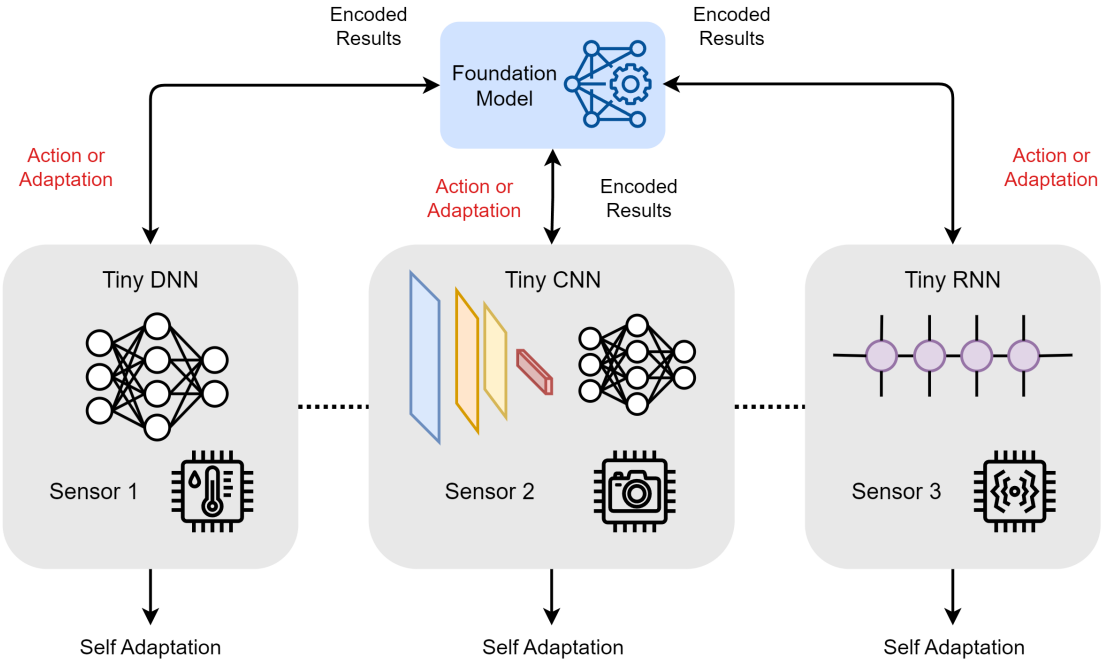


FIGURE 2.9: High-level schematic of multi-level intelligence, depicting the incorporation of TinyML with foundation models for decision fusion. Data is processed both on the edge by sensors equipped with machine learning algorithms and on the cloud by more powerful foundation models that require higher computational capacity.

2.3.3.3 Federated Learning

Building on the concept of edge computing, distributed intelligence, and the implementation of machine learning on edge and resource-constrained devices, federated learning is a prominent research area that aims at further tying these concepts together in a collaborative manner [147]. Multiple actors within a federated learning system process data collected in loco to update a local model and act as workers for a central aggregator of information or overarching machine learning model. Each node of the system receives a model to be used on the data collected to then relay the computed loss back to the central aggregator. Once the aggregator receives all the losses from the distributed devices, it solves an optimization problem to minimize a global loss function [147]. An example of the algorithms that can be used to solve such optimization problems is federated averaging [148]. Before the next round of learning the central computing server broadcasts the updated weights for the local models.

Federated learning can also be modelled without a central aggregator in a fully distributed fashion. These architectures leverage P2P communication to transmit the data needed for collaborative training. The aggregation is performed by every node after it has received updates from neighboring nodes. Blockchain has been proposed as an effective means of communication between nodes, the shared ledger can be used to update a common global perspective of the state of the distributed system. Within this structure, smart contracts can be used to share data among users reliably and securely.

In both centralized and decentralized architectures several benefits of collaborative processing can be extrapolated. Since the raw data is never transmitted among users and centralized servers, eavesdropping and man-in-the-middle attacks would be mitigated, improving data sharing privacy. It is virtually impossible to recreate trained models that could digest the encoded information shared, although this system does not prevent false data injection attacks and denial of service attacks that could be implemented to take down the system by overloading or misleading it. Furthermore, the data transmission would be faster as the raw data is compressed in latent space representations that could encompass a larger window of data. The performance and learning quality of the overall system could also be improved by a more efficient distribution of the processing load, enabling better scalability of CIoT systems.

Moreover, central aggregation algorithms could be biased through the encoding of further semantic contextual information using foundation models. This would allow for distributed systems to benefit from data not directly collected by the nodes of the system

but publicly available through official channels. Contextual data of this form could be, for instance, weather predictions, to be integrated into a distributed system managing a transportation network or autonomous driving network of vehicles. Further improving the adaptability and redundancy of the systems through bias injection [83].

2.3.4 Distributed Storage and Parallel Processing

In order to achieve a high level of autonomy, CIoT systems will have to perform data storage, processing, and retrieval in a timely, efficient, and reliable manner. This is particularly important if taking into consideration smart cities or smart grid (SG)-based CIoT architectures where a failure in the data pipeline could cause widespread outages or malfunctions. Big data processing and storage has been widely covered in the literature [149, 150, 151, 152, 153], where operations are parallelized to improve speed and efficiency at large-scales.

As open-source software, a variety of distributed storage options are available and could be easily integrated into a CIoT architecture. The Hadoop distributed file system (HDFS) is the most well-known and is a distributed storage solution that offers several essential features such as sharding and redundancy to ensure the integrity and reliability of the data storage [154]. For instance, the use of sharding techniques to divide files into smaller chunks of data and store them across multiple network nodes offers redundancy, integrity, and efficiency in an IoT structure. Furthermore, these features ensure that even if one of the nodes fails or is compromised, the data may still be retrieved and restored to its original state. These open-source big data analytics engines are based on distributed file systems that leverage clusters of servers to store their data and implement parallel processing to increase processing time by splitting the computation into multiple tasks distributed among various resources or executors [155]. Specifically, Apache Hadoop is divided into four main modules seen in Table 2.2.

It also has further modules to address real-time data streaming (Apache Kafka) and a scalable multi-master database (Apache Cassandra). Apache Spark is a multi-language engine that works within the Hadoop framework but has its own ecosystem, consisting of the main modules also listed in Table 2.2.

There are also several other platforms for big data analytics, proposing alternative solutions or expanding upon the Hadoop system. A few notable ones are 1010data, Cloudera data hub, Flink, Storm, Samza, SAP-Hana, HP-HAVEn, Hortonworks, Pivotal big data suit, Infobright, and MapR [156, 155].

TABLE 2.2: Software packages and modules for distributed storage and processing.

Environment	Module	Function
Hadoop	Common	Library access
Hadoop	File System	High throughput access to data
Hadoop	YARN	Job and cluster scheduling
Hadoop	MapReduce	Parallel processing based on YARN scheduling
Spark	Core	Data processing engine
Spark	SQL	SQL database interactions
Spark	Streaming	Real-time data streaming
Spark	MLib	Machine learning dedicated library
Spark	GraphX	Graph-based visual representations

Furthermore, new parallel storage solutions arise as open-source projects continuously, innovating and creating more robust and distributed architectures. An outstanding example is the interplanetary file system (IPFS). IPFS is a distributed file system based on the libp2p library for M2M communication that differs significantly from HDFS. The secure Kademlia protocol [157, 158], comparable to a gossip-like protocol, is used to retrieve file chunks using acyclic Merkle directed acyclic graphs (DAGs) instead of a master node. Merkle DAGs are an alternative to distributed hash tables (DHT) for breaking up files into chunks and reconstructing them; the system may collect all the shards that make up any given file from the network of nodes and check data integrity by having the root hash of the file [159]. The real benefit of employing IPFS based storage systems is that it may be deemed truly redundant due to its features and decentralized character.

Therefore, a combination of dynamic resource management coupled with existing open-source distributed storage and processing could be the answer to creating scalable decentralized self-managing architectures in CIoT.

2.3.5 Data Mining in CIoT

Data mining refers to the exploitation of available data to extract knowledge about the environment and process it is associated with. Hence, it extracts high-level information from low-level raw data [120]. To extrapolate such high-level information the data must go through preprocessing, feature extraction, abstraction, and semantic derivation as shown in Figure 2.10.

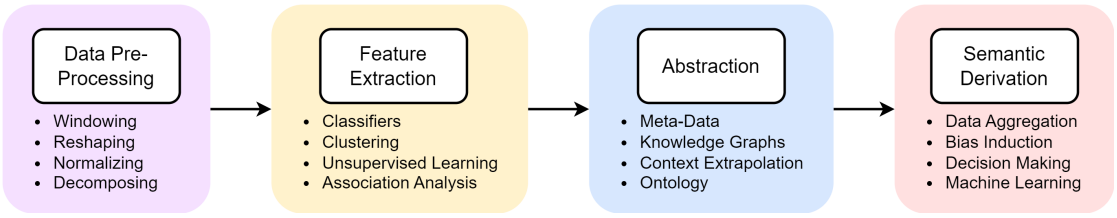


FIGURE 2.10: Conventional analytics stages of data processing.

2.3.5.1 Data Preprocessing

is the first step of data analysis and utilization. Different layers can adapt this step to reduce transmission costs, as mentioned in the previous sections. Generally, some preprocessing will always be carried out by hardware or software features at the node level. Signal preprocessing, for instance, is done using low/high pass filters and bandpass filters using specifically designed circuits.

Mathematical preprocessing, in contrast, does not modify the signal and utilizes the output instead. Such techniques aggregate the data over time windows and transmit such aggregations' characteristics, such as minimum, maximum, mean, median, variance, standard deviation, derivatives, integrations, and correlations [160].

The problems of heterogeneity, nonlinearity, and high dimensionality must be addressed in the processing of the data before abstraction, knowledge discovery, and semantic derivations. Heterogeneous data processing brings both challenges and new possibilities in analyzing sensory data. Mathematically, joint probability density functions can be exploited based on copula theory (couples' multivariate joint distributions to their marginal distribution functions) to model random variables characterized by heterogeneous data. Furthermore, practically adaptive mechanisms could be used to automatically select algorithms developed to address the data type in question. Nonlinear data processing can often outperform the linear counterparts in many applications, as often linear methods are oversimplified to deviate to optimality. Kernel-Based Learning (KBL) is proposed to tackle the mathematical problems associated with nonlinear data processing [33]. High-dimensional data processing is often challenging, due to the large amounts of data being processed. To cope with this, dimensionality reduction techniques can be applied to reduce the size and complexity. Several methods are presented in [160] to address the problem, such as discrete Fourier transformation, wavelet transformation, piecewise aggregation approximation, and symbolic aggregate approximation.

2.3.5.2 Feature Extraction

is performed after the data has been preprocessed. It can be an ambiguous term, but it generally refers to cluster analysis and feature selection processes. An example in image processing would be the use of attention-based mechanisms to weight in the position of different regions or characteristics to improve the performance of image recognition software. Classifying the data or clusters helps make patterns emerge from the data. This process can be carried out using various algorithms in unsupervised learning and following different methods [161]. For example, hierarchical clustering is defined as combining data in subgroups and then constructing subgroups made of subgroups, ultimately forming a hierarchy tree. This is an iterative process of agglomerating, where divisive clustering is helpful for further analysis to define sets of clusters. Another method is partitioning clustering, an iterative solution that reallocates data points between subsets. It can also be used along density-based functions to recognize data clusters; for instance, k-medoids and k-means are examples of this type of clustering [161]. Furthermore, unsupervised machine learning techniques can be used to highlight correlations between groups and association analysis, multilevel association, multidimensional association, and quantitative association. Latent features can also be generated using a Bayesian generative model as discussed by Haykin *et al.* in cognitive risk control [70].

2.3.5.3 Abstraction

is defined as deriving contextual data from sensory data by coupling the raw measurements with meta data and additional knowledge to gain better insights and adjust sensory devices to current factors. In [160], the authors define two types of abstraction. Lower-level abstraction represents static information, such as a single, independent observation made at a specific time step, gathered from sensors along with metadata like sensor range, type, and capabilities. For example, a temperature sensor may report a reading of 25°C, accompanied by details about the sensor's accuracy and operational range. Higher-level abstraction is achieved by analyzing several lower-level abstractions together to better understand complex, multivariate events. For instance, combining temperature, humidity, and wind speed data over time can help detect weather patterns, while aggregating data from motion sensors, door sensors, and camera feeds can infer occupancy patterns in a smart building. Common abstraction techniques are classification, Markov chains, and hidden Markov models (HMMs). Classification is a standard abstraction method to find the correlation between group samples with similar attributes and characteristics. Markov chains are used to represent the likelihood of temporal relations among groups.

Understanding the context in which smart systems exist and to induct such information into the processing is the key enabler to create cognitive systems. Contextless processing, blind to the relativity of the perceived environment, could never lead to informed adaptation. In CIoT context acquisition can be evaluated based on how the processing is biased to account for relative information, the frequency of acquisition, how it is shared within the system, and its relevancy. Many context modeling techniques exist, each aimed at influencing different components of an IoT or CIoT system in various ways. For instance, location-based context modeling can optimize resource allocation in smart homes by adjusting heating or lighting based on occupant behavior, while activity-based modeling can improve wearable health monitoring systems by tailoring alerts based on detected physical activity patterns. Additionally, environmental context modeling can be used in agriculture to automatically adjust irrigation based on soil moisture levels or weather conditions. Logic-based modeling is the most basic technique used to express context through a set of rules and logical expressions to establish direct cause-effect relationships, but fails to define non-linear relationships. Similarly, key-value, ontology, and markup scheme modeling are used to define simple system data structures providing flexible and efficient storage of such relationships. Spatial modeling can be used to integrate physical space and location meta data to be associated with data extracted, providing some level of context to be used in the processing. Defining the relationships between nodes in multi-agent systems is an example of graphical modeling, and embeds the bilateral influence that nodes have on each other expressing their conditional dependencies in a probabilistic fashion. For instance, in a production line composed of multiple sub-stages, the processes influence the subsequent stages by means of variable middle product properties. In a distributed network system, the agents will occupy certain frequencies using communication, crowding the space and possibly influencing other devices that are transmitting on the same frequency. While modeling environmental contexts, uncertainty presents a limitation in establishing contextual information, environmental conditions, especially in relation to complex systems such as the weather can only be modeled and integrated in a probabilistic way, being hard to capture [162].

Furthermore, HMMs are built upon Markov chains adding temporal dimensions for classification purposes. Decision trees and K nearest neighbor (KNN) algorithms are the most notable classification methods. Bayesian networks (including naïve Bayes, selective naïve Bayes, semi naïve Bayes, one-dependence Bayesian classifiers,, and more), and support vector machine algorithms [161].

2.3.5.4 Semantic Derivation, Reasoning, and Decision Making

The semantic derivation is the last step in data analysis and allows models to represent the correlation between related context information, metadata and data. For example, in semantic ontology, events can be linked to reason from simple to more abstract [160]. In addition, domain ontologies-based schemas represent the data from different sources, increasing interoperability.

Once context is modelled, the reasoning and elaboration over abstractions and semantic derivations can be carried out as mentioned. The next step is to integrate the information extracted into inference models capable of reasoning and decision making. To this end, context reasoning can be modelled using rules engines, probabilistic logic, fuzzy logic, supervised learning, unsupervised learning, and RL. This is the last step of the multilevel processing of data that distributes upwards in CIoT systems, and outputs the executive commands to be executed by the systems to self-adapt or influence the environment. The agglomeration of lower-level processing is a challenge in cognitive systems. The curse of dimensionality, the result of the integration of large corpora of multivariate data, presents the first limiting factor, which will be addressed more in Sections 2.4.1 through the decomposition of the processing and parallelization.

The 'no free lunch theorem' for machine learning highlights how no single model can properly generalize without some inductive bias, in the sense that there is no general-purpose learning algorithm [163]. A model can be therefore influenced by the dataset it was trained on in the form of the data distribution, and by the way it was trained, in the form of the learning algorithm and loss function used. Context-dependant processing is a form of inductive bias, integrating sensory signals with the bias brought via contextual information. In this perspective, the main question and major hurdle to cognitive systems in general is how to develop a model capable of learning to piece together a puzzle of relative information and marginal distributions to optimize over a specific loss function of a given process and achieve a goal in an adaptable and flexible way. The human brain is capable of adaptation mechanisms far beyond any computational algorithm ever created. An example is cross-modality reassignment, where previously learned structures assigned to process a specific input are reprogrammed to accept input from a different sensory modality [164]. How to create a system flexible enough to mimic this behavior is still largely unanswered.

2.4 Cognitive Data Analysis

Having detailed how CIoT systems extract high-level insights from raw sensor data, beginning with data preprocessing to address heterogeneity, nonlinearity, and high dimensionality; advancing through feature extraction and abstraction to enrich and contextualize the data; and culminating in semantic derivation that integrates metadata and context, we have established a robust conceptual foundation. Yet, these analytical building blocks represent only the first step. The true potential of CIoT emerges when these insights are harnessed to drive intelligent, adaptive behavior in real-world applications.

Looking ahead, the future of CIoT lies in translating these foundational principles into practical, forward-thinking solutions. In the sections that follow, we present a roadmap for this evolution by exploring cutting-edge approaches in cognitive computing. We delve into the architecture of cognitive systems, examine reinforcement learning and reward-based adaptation, and highlight pioneering systems such as IBM Watson that exemplify the current state-of-the-art. Further, we shift our focus to transformative models, foundation models, and large language models (LLMs), which promise to enable seamless data fusion, efficient transfer learning, and the dynamic embedding of AI capabilities across all CIoT layers.

This journey from theoretical underpinnings to applied intelligence not only underscores the potential of integrating advanced analytics into CIoT but also charts a clear path forward for designing systems that continuously learn, adapt, and thrive amid the complexities of an ever-evolving, interconnected world.

2.4.1 Cognitive Computing

In recent years, cognitive computing has emerged as an evolution of conventional data analytics, integrating disciplines such as linguistics, psychology, artificial intelligence, neuroscience, anthropology, engineering, and computer science [165]. Similar to CDS, cognitive computing seeks to emulate and embed elements of human cognition into autonomous systems [166], enhancing machine intelligence through multimodal, adaptive data analysis. By combining traditional data analytics techniques, cognitive computing enables systems to ingest, analyze, and aggregate vast amounts of unstructured data, facilitating optimal action selection and policy decisions.

Cognitive computing can thus be defined as the adaptive integration of multiple machine learning and data analytics techniques to extract knowledge from the environment, enabling cognitive engines to enact corrective actions through actuators and machine-human interaction in dynamic and adaptable ways. Beyond being a mere machine learning technique, cognitive computing represents a comprehensive architecture, integrating multiple subsystems of machine learning and analytics [167]. Within a CIoT architecture, distributed intelligence is achieved through mechanisms such as distributed computing and federated learning, which have been further discussed in Section 2.3.3.2.

Moreover, cognitive systems leverage past interactions with the environment to maintain representations of entities through short- and long-term memories, encompassing the system's assumptions, motives, ideas, and knowledge [168]. The ultimate goal of human-centric cognitive computing is to process increasingly diverse data and deliver knowledge tailored to the needs of a specific individual in a given context [169]. DARPA defined cognitive computing in 2002 as the ability to accumulate knowledge, reason, use represented knowledge, learn from experience, follow directions, operate robustly under uncertainty, adapt to sudden events, and be aware of the system's behavior and its influence on the environment [169].

2.4.1.1 Reinforcement Learning and Reward-Based Adaptation

RL is a powerful class of machine learning algorithms that enable agents to learn optimal behaviors through iterative interactions with their environment. At the core of RL is the concept of learning from rewards and punishments: agents take actions in an environment, receive feedback in the form of rewards or penalties, and adjust their policies accordingly to maximize cumulative rewards over time. This process is formalized through the maximization of a specific cost function, often involving expected future rewards discounted over time. RL algorithms are particularly adept at handling problems where the optimal action is not immediately apparent and must be discovered through exploration and exploitation strategies.

This learning paradigm draws parallels with cognitive neuroscience, specifically the functioning of the striatal-dopaminergic system in the human brain [170]. The striatum and dopaminergic neurons play a critical role in motivation, reward processing, and motor control. Dopamine signals are believed to encode reward prediction errors (the difference between expected and received rewards), a fundamental concept in RL algorithms for updating value functions and policies. This neurobiological basis provides a

compelling connection between artificial learning systems and natural intelligence, suggesting that RL algorithms may capture essential aspects of how humans and animals learn from their environments.

Similarly, deep learning methods have been inspired by the hierarchical and interconnected nature of neural networks in the human brain. Deep learning utilizes artificial neural networks with multiple layers, comprising of input, hidden, and output layers, to model complex and non-linear relationships in data. These networks can automatically learn and extract high-level features from raw inputs, enabling breakthroughs in areas such as image recognition, natural language processing, and speech synthesis.

Combining the strengths of RL and deep learning has led to the emergence of DRL, a paradigm that leverages deep neural networks to approximate value functions, policies, or models of the environment [171]. DRL enables RL algorithms to handle high-dimensional state and action spaces, which were previously intractable with traditional RL methods. By integrating deep learning, DRL agents can process unstructured inputs like images or sound, allowing them to make decisions based on rich sensory data.

Generally, RL and DRL models are employed for dynamic and sequential decision-making and control problems. These problems are often formally modeled as MDPs [172], which provide a mathematical framework for modeling decision-making scenarios where outcomes are partly under the control of an agent and partly random. An MDP consists of a set of states, a set of actions, transition probabilities, and reward functions. The goal in an MDP is to find a policy (a mapping from states to actions) that maximizes the expected cumulative reward.

In recent years, hybrid DRL approaches have achieved remarkable success, outperforming humans in complex games such as Go and various Atari games [173, 174]. Notably, Google's DeepMind developed AlphaGo, which defeated world champion Go player Lee Sedol in 2016. AlphaGo combined deep neural networks with advanced search algorithms, demonstrating the potential of DRL in mastering tasks with vast search spaces and intricate strategic elements. Similarly, DRL agents have been trained to play Atari 2600 games directly from raw pixel inputs, achieving superhuman performance in many cases.

RL algorithms have also been successfully applied to a wide range of robotics problems, such as indoor navigation, manipulation tasks, and control-related challenges [171, 175]. In robotics, RL enables agents to learn control policies through interactions with either real or simulated environments, reducing the reliance on hand-crafted

controllers. For example, robots can learn to navigate complex environments, avoid obstacles, and perform tasks like object grasping by learning from trial and error.

As mentioned earlier in this paper, RL was the algorithm chosen by Feng and Haykin for developing cognitive control and cognitive risk control architectures. Their work highlights the suitability of RL for modeling adaptive decision-making processes in complex, uncertain environments. By leveraging RL, these architectures aim to mimic cognitive functions such as attention, learning, and risk assessment, which are essential for intelligent systems operating in real-world scenarios.

However, modern RL and DRL, despite their prowess in policy and action selection, face limitations in generalization and handling unstructured data. These models often require large amounts of training data and computational resources, and they tend to learn policies that are specific to the training environment. Consequently, their ability to ingest unstructured data, that does not have a predefined data model or is not organized in a pre-defined manner, is limited. This constraint hampers the application of RL and DRL in domains where data is heterogeneous and lacks clear structure.

Furthermore, the transferability of pre-trained RL and DRL algorithms to other applications is severely limited. Unlike humans, who can apply learned knowledge to new contexts with minimal adaptation, RL agents typically need to be retrained when the environment changes or when faced with new tasks. This lack of transfer learning capabilities reduces the practicality of deploying RL solutions across multiple domains, as it incurs significant time and resource costs for retraining.

Humans possess the remarkable ability to integrate prior knowledge with new information, a process formalized by the tendency to search for structures during interactions with the environment [170]. This cognitive process involves abstracting underlying patterns and relationships, even when no explicit structure is apparent. By forming structured representations of the environment, humans can generalize learning and apply it to novel situations, a strategy that proves advantageous for long-term learning and adaptation.

This tendency to abstract structured representations facilitates computational efficiency. By representing learned rules independently of specific sensory and motor outputs, humans can apply these rules flexibly across different contexts. This abstraction reduces the cognitive load and enables the reuse of learned behaviors, which is a significant computational gain. Policies derived from such structures are extensively generalizable and transferable, allowing humans to adapt quickly to new environments

or tasks. The ability to transfer and generalize learned knowledge and policies to other situations is a fundamental behavioral hallmark of human learning [170].

An illustrative example is how humans adjust their behavior according to situational context. For instance, one might communicate differently in a professional meeting compared to a casual gathering with friends. Humans can translate specific context-dependent policies, selecting appropriate actions based on cues from the environment. This context-aware decision-making showcases the flexibility and adaptability of human cognition.

The concept of state abstraction is crucial for lifelong RL, where the goal is to develop agents capable of learning and adapting over extended periods [176]. State abstraction involves simplifying the representation of the environment by focusing on relevant features and ignoring irrelevant details. By developing reusable policies and adapting previously created structures to new contexts, agents can generalize learning and improve efficiency. This approach mirrors human cognitive strategies, where abstract representations enable the application of knowledge across diverse situations.

2.4.1.2 IBM Watson: A Pioneer in Cognitive Systems

The first valid attempt to build a cognitive system covered in literature is IBM Watson and the Deep QA project, which can utilize a vast array of structured and unstructured data by adaptively integrating multiple analytics and machine learning techniques [177, 178]. This tool is commercially available as sub-packages within IBM Cloud services such as Watson Assistant, Watson Studio, Watson Discovery, and Watson Analytics [179]. The key feature that originally differentiated the Deep QA architecture from earlier analytics solutions offered by companies like Microsoft (Azure) and Amazon Web Services (AWS) is that Watson did not simply map questions to a database of predefined answers. Instead, it relied on principles such as massive parallelism, pervasive confidence estimation, and the integration of both shallow and deep knowledge [180]. While modern LLMs share some of these characteristics, the Deep QA architecture was among the pioneers in using such techniques. The high dimensionality of arrays produced by Watson in terms of the analysis and processing of hundreds of features or scores using various analytics tools and the ability to combine them into a statistical reference value inside an action space differentiates Watson. Deep QA can train and functionally adapt to domain-specific taxonomies and reasoning by detecting context and adapting specific content [181]. This architecture can be considered a first attempt to design a foundation model capable of utilizing knowledge extracted from large corpora of data and transferring the

learning to downstream tasks. The original depiction of the Deep QA architecture can be seen in Figure 2.11. This ability can be tied back to the five principles of cognition outlined by Haykin since it fits within the attention mechanism of cognition. It also possesses all other pillars of cognition, so it can be considered a cognitive engine. IBM Watson's development has mainly focused on the healthcare industry and the potential applications of the Deep QA architecture to aid doctors and researchers in making the most informed decisions [182]. This adds to the growing pool of research and new clinical studies published daily in medical journals, which constantly advance state of the art for treatments and procedures.

Through various techniques such as natural language processing (NLP), dynamic learning, and hypothesis generation, cognitive systems can effectively and intelligently parse through this massive amount of data to aid the coordination of care by healthcare professionals. Currently, Watson is being used to aid cancer treatment in partnership with New York's Memorial Sloan-Kettering Cancer Center, the MD Anderson Cancer Center, and the University of Texas [182]. IBM has also partnered with Apple, Johnson & Johnson, and Medtronic to further enhance patient monitoring, to better utilize information gathered by personal health, medical and fitness devices leaping into the future of healthcare IoT-based devices, to provide better patient monitoring and real-time feedback and recommendations to doctors in chronic and acute care [182].

IBM Watson can also be applied across domains, and it is capable of aggregating diverse data into a single repository called corpus or body, which is domain-specific. This flexibility allows Watson to be applied to law, medicine, engineering, finance, and more using a tailored corpus of information [183].

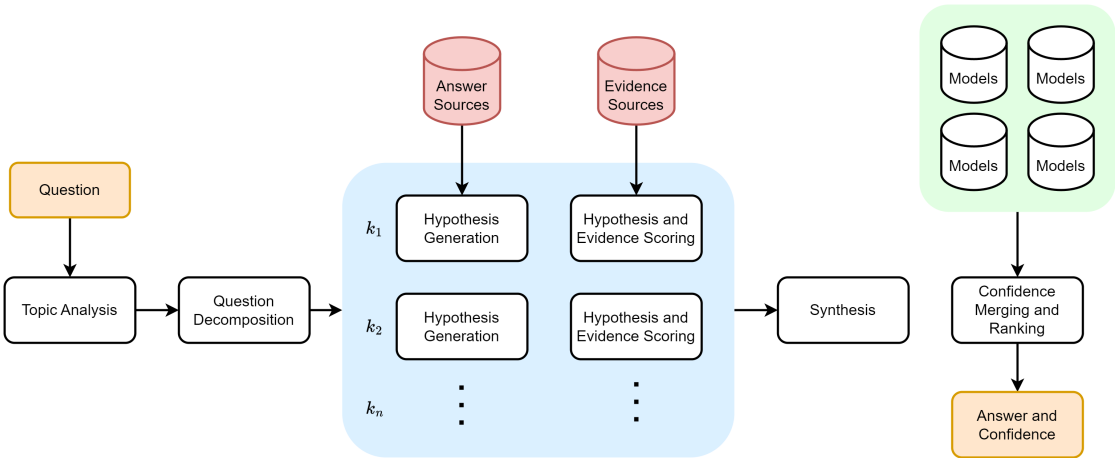


FIGURE 2.11: Watson Deep Question Answering Model Representation.

Moreover, as recent work in cognitive science shows, human learning capabilities can only be understood in the context of numerous independent, interacting memory systems rather than as a single, complex learner [170]. This characterization highlights how to create an artificial cognitive system, the parallel use of multiple analytics tools is needed. No general model can reproduce the context extrapolation, abstraction, and learning ability to efficiently adapt to new environments and sudden unexpected events or changes in the system's parameters. Although an aggregation model, biased by contextual encoded data at the edge could be used to optimize over an optimization problem of various objectives loss functions [147].

2.4.2 Transformers, Transfer Learning and LLMs

The transformer architecture was first proposed by Vaswani *et al.* in [184] and remains one of the most prominent machine learning models today. Transformers inherently possess strong generalization abilities, which enable them to excel across various tasks and modalities. In [185], Lu *et al.* tested the hypothesis that transformers pre-trained on data-rich modalities, such as massive natural language corpora, can be effectively utilized for tasks in different modalities, including image classification and protein folding prediction.

Table 2.3: Large pretrained language models based on transformer architectures.

Name	Model	Size	Citation
BERT-Large	Transformer	345 M	[186]
CLIP	Transformer	428 M	[187]
GPT-3	Transformer	175 B	[188]
GPT-4	Transformer	1.76 T (Unofficial)	[189]
PaLM	Transformer	540 B	[190]
PaLM 2	Transformer	340 B	[191]
LLaMA	Transformer	Up to 65 B	[192]
Megatron-Turing NLG	Transformer	530 B	[193]
BLOOM	Transformer	176 B	[194]
DeepSeek	Transformer	67 B	[195]

To test this hypothesis, the authors used a pre-trained language model termed the Frozen Pretrained Transformer (FPT); specifically, they utilized a version of GPT-2 [187]. After fine-tuning certain parameters, excluding self-attention or feedforward layers, the FPT demonstrated performance comparable to or better than long short-term memory networks (LSTM) and transformer models trained entirely from scratch on the given tasks and datasets [185]. This finding highlights the significant potential of language models for transfer learning and suggests that they might inherently possess

the capacity for universal data computation and structural learning for predictive tasks across different modalities [185].

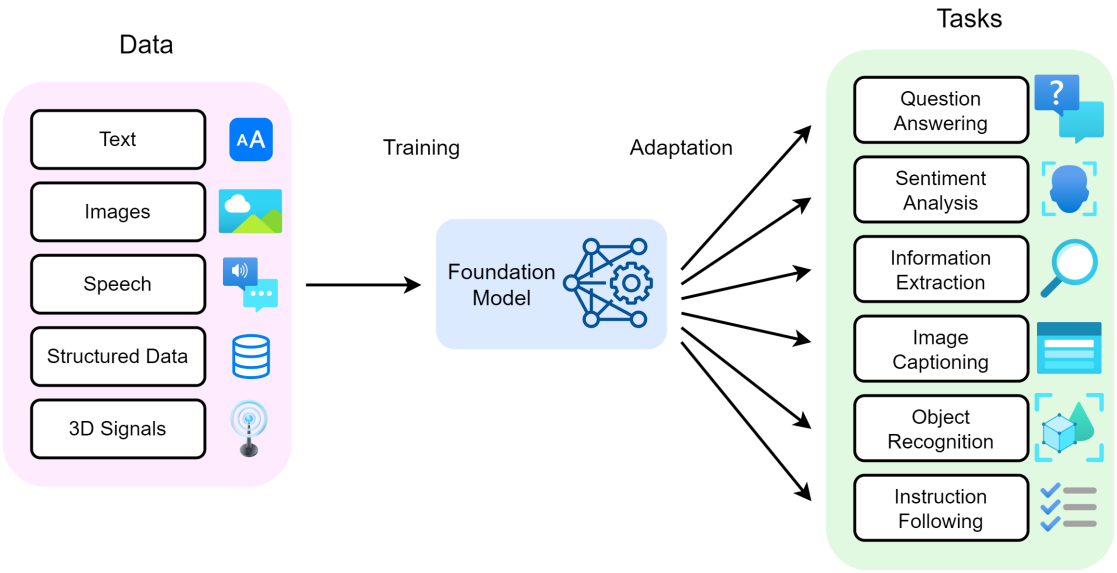


FIGURE 2.12: Foundation models visualization, adapted from [3]. Multi-modal data is fed to the foundation model for training, which is then able to utilize it and adapt to perform a multitude of tasks.

The term foundation models was introduced by researchers from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) at Stanford University [3]. A foundation model is defined as any model trained on a broad, diverse dataset at a scale that allows it to be efficiently adapted to perform a wide range of tasks. Examples include Bidirectional Encoder Representations from Transformers (BERT) [186], LLaMA [192], GPT-3 [188] and GPT-4 [189].

These models leverage self-supervised learning and various implementations of transformer architectures at an unprecedented parametric scale, enabling effective transfer learning across tasks. As a result, they can apply "knowledge" acquired from one task, such as object recognition in images, to other tasks not explicitly trained on, like activity recognition in videos [3]. Recent developments have expanded the capabilities of foundation models, with examples like PaLM [190], PaLM 2 [191], LLaMA [192], and BLOOM [194] pushing the boundaries of model size and performance. These advancements highlight the increasing versatility and generalization capabilities of foundation models across various modalities and complex tasks. A comparison of model size can be seen in 2.3.

Multimodal models such as CLIP [187] have further reinforced this trend by integrating visual and textual data, enabling zero-shot image classification without task-specific fine-tuning. The introduction of GPT-4 [189], which can process both text and image inputs, exemplifies the expanding versatility of transformer-based models across different modalities. Moreover, models like PaLM 2 [191] and LLaMA [192] have advanced language understanding and generation, signaling a trend toward increasingly generalized and multimodal AI systems.

Furthermore, transformers-based foundation models have led to unprecedented levels of homogenization as most state-of-the-art NLP models effectively adapt one of the foundation models [3]. This ability could potentially enhance the performance of models in domains where task-specific data is heavily limited [3]. The aggregation of diverse data types such as text, images, and speech as well as unstructured data and sensory feedback could be used in the training of foundation models, which then adapt to downstream tasks such as sentiment analysis, object recognition, question answering, and more through minor parameter fine-tuning, a depiction of the training and adaptation of foundation models can be seen in Figure 2.12.

The flexibility of foundation models comes from three primary characteristic abilities [3]:

- Generalization: the ability to generate suitable candidates in the action space of an optimal decision-making process. Foundation models can carry out this process entirely unconstrained, given their ability to model the output space as a sequence.
- Grounding: the ability to ingest and process diverse data types that may hold deep underlying semantic meanings, such as mathematical and symbolic language, and to use them in the correct context. Through inductive bias learning and pretraining, transformers have also been explicitly researched for mathematical reasoning, thanks to their flexibility. For example, in [196] the authors simulated three main features of primitive reasoning, deduction, induction, and abduction, embedding them within a classic transformer architecture for mathematical reasoning.
- Universality: the ability to transfer knowledge across tasks, either through the generalization of low-level techniques or by efficiently utilizing metadata techniques across domains [3]. Foundation models can leverage similarities between reasoning problems and latent structures through the use of meta-knowledge encoded in the models' weights. This ability is comparable to humans' ability to create structures and use them if applicable in different settings, as discussed earlier.

Moreover, self-supervised learning, which can be considered a subset of unsupervised learning, enables models to learn from diverse unstructured data without expensive labeled datasets [197]. Generally, self-supervised and supervised objectives are very similar, with the difference that the first one is only evaluated on a subset of the sequence. The challenge is to optimize the unsupervised objective to converge to the global minimum [187]. All major foundation models such as GPT-4 and LLama are trained using self-supervised learning, making them broadly task-agnostic by nature. Although to perform up to state-of-the-art, these models still need some fine-tuning for most applications, involving supervised learning but on a much smaller scale, especially for downstream tasks that the models were not explicitly trained for [188].

Tying back to CIoT, the challenge presented by heterogeneous data processing still presents significant obstacles to traditional models. The longstanding challenge of embedding IoT and robotics systems with the ability to efficiently handle the multitude of conditions and dynamicity of real-world scenarios could be potentially solved by the evolution of foundation models. Foundation models for robotics present opportunities for task specification, adding the ability to adjust structures to new contexts dynamically, adapting to downstream task learning, and taking the mathematical form of joint distributions over action and observation spaces [3]. To this end, a combination of transformers architectures and RL models could be a powerful combination for task learning and adaptation, through the aggregation of collected data. Some researchers have attempted to create hybrid architectures found in [198, 199, 200, 201] with promising results.

2.4.2.1 Data Fusion

The concept of combining sensory and contextual data of different modalities is central to the CIoT paradigm and is one the main enhancements from IoT systems. By virtue of utilizing diverse data sources that inform and complement each other it is possible to obtain a more comprehensive view of the environment, the system, and their interactions. A set of complimentary sensors that collect data on the same time frame to observe a common system or environment can provide different perspectives and enhance the overall information used in processing. To this end an ensemble of datasets and data types is more than the sum of its parts [202]. Furthermore, by linking together multimodal data a new form of diversity is introduced aiding the optimization problem to converge to a unique solution in techniques such as tensor decomposition [127].

The fusion of different sensors can be interpreted as cooperative, competitive, or complementary. Cooperative fusion relies in sensors that inform each other and is used for example in triangulation problems. Competitive fusion also referred to as redundant fusion, refers to the fusion of multiple sensors that are independent and provide the same type of information, used to improve reliability and accuracy. Complementary fusion, which is the focus of discussion in this article, can complement each other and provide a more complete picture of the system or environment [203].

The problem of data fusion as seen in cognitive risk control can be interpreted using Bayesian statistics. In fact, probabilistic data fusion methods were among the first proposed along with fuzzy logic methods for fusing multiple sensory measurements, of the same nature. Bayesian data fusion methods have found use in the past in tracking and position estimation problems. Using Bayesian inference prior knowledge of the parameters, parameter relationship, and environment uncertainty is integrated to obtain posterior probabilities or estimates [204]. A multisensory fusion model based on Bayesian statistics can be formulated as:

$$P\left(z_1|y\right) = \prod_{m=1}^{M} \prod_{n=1}^{N} N\left(y_{n,m}|z, \Sigma_m\right) \tag{2.1}$$

Where M is the number of sensors and N is the number of observations from sensors m, and $y = y_{1:N,1:M} \in \mathbb{R}^K$ [205]. Bayesian filtering techniques such as the Kalman filter (KF) have been also used to model temporal multivariate Gaussian distributions that aim to measure the same state such as in the case of GPS and inertial navigation system (INS) sensors [206, 207, 208]. Although these methods are harder to frame for heterogeneous data fusion.

Data fusion techniques that don't rely strictly on the Bayesian framework can be categorized into two main groups:

- Algorithms that focus on decomposition techniques such as tensor decomposition, principal component analysis (PCA), singular value decomposition (SVD), independent component analysis (ICA), eigenvalue decomposition (EVD), and factor analysis (FA)
- Machine learning algorithms [209].

Where the former can be used in maximum likelihood estimation (MLE) or tensor regression techniques, while the latter takes the form of various types of neural networks.

Decomposition techniques can be used in combination with machine learning and implemented at different stages of processing for example they can be used as a preprocessing step to a neural network architecture as seen in [210, 211], but also as part of the feature extraction process [212].

In general data fusion implementation levels can be categorized as follows and as shown in Figure 2.13:

- **Early fusion**: also known as feature fusion occurs as a preprocessing step. Common methods include concatenation, averaging or weighted combination of input data into a single matrix or tensor [213]. The concatenated features are then fed as a single input to a neural network. Decomposition techniques can be used to extract lower dimensional meaningful representation of the data to then be fused together by the same means, an example of this process is principal component regression (PCR) [209]. Machine learning models can also be creating feature maps that are concatenated to serve as a single modality input to the inference model. Variational autoencoders (VAE) and multimodal autoencoders (MVAE) are often used to learn a joint representation of the different inputs by means of signal reconstruction.

- **Intermediate fusion**: also known as joint fusion, combines latent representations of data obtained through neural networks. This method uses multiple neural networks to extract feature maps which are then jointly fused into an inference model. The key difference with early fusion lies in the backpropagation of the error to the feature extracting neural network [214]. Decomposition techniques do not fall under this category but can be complementary to the feature extracting networks in the preprocessing stages.

- **Late fusion**: also known as decision fusion combines the outputs of multiple machine learning models (trained separately) to make a final decision. The different model outputs are combined using aggregation functions such as average, majority voting, maximum value, Bayesian decision rule, metaclassifiers, and more [209, 214]. Late fusion can be particularly useful for multimodal heterogeneous applications since it doesn't need dimensionality reductions and other techniques to aggregate types of data which are diverse in nature and format [213].

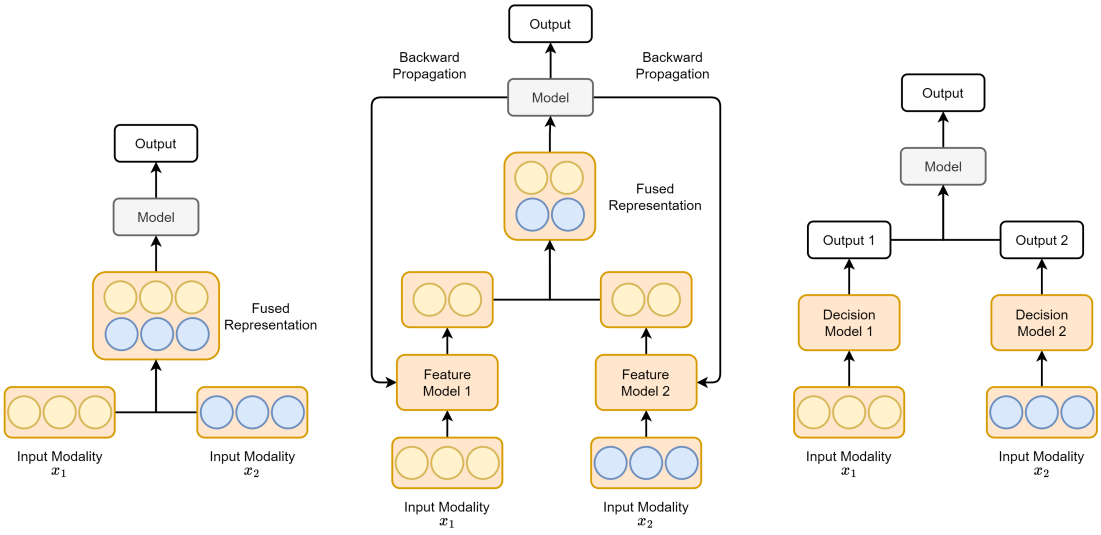


FIGURE 2.13: Data fusion levels in machine learning, schematic diagram. Early fusion (left), intermediate fusion (center), late fusion (right).

The use of data fusion for multimodal data utilization has shown promise in the fields of biomedical fusion, remote sensing, and autonomous vehicles. As already mentioned, IBM Watson is the first example of cognitive computing that has been leveraged to aid doctors in making diagnosis, as well as in smart patient monitoring and internet of medical things (IoMT) applications, making use of diverse multimodal data. In this context, data fusion techniques have also been used to fuse together different imaging modalities such as PET-MRI, MRI-CT scans [215, 216]. Data fusion techniques have also been proposed to fuse together time series data and imaging data in the biomedical context such as electroencephalograms (EEG) and functional magnetic resonance images (fMRI) which provide complementary information about brain functions [217, 202]. Moreover, data fusion techniques have been used in the diagnosis, prediction, and classification of several diseases. For the interested reader, more information is provided in [214].

Remote sensing refers to the acquisition of data about an object or phenomenon from a distance and typically refers to the use of sensors mounted on satellites and aircraft to collect data about the Earth surface, atmosphere, and oceans. Such data can provide information on the topography, land cover, vegetation, pollution levels, and climate patterns that can be used for a wide range of applications across many fields [218]. Examples include environmental monitoring, urban planning, disaster management, climate change research, and defense applications. The sensors used in these types of applications are usually complementary as they capture different aspects of the Earth's

surface, for example, optical sensors provide detailed color images while radar sensors can penetrate clouds and provide data on the surface topography. Data fusion has been widely used in this field in both homogeneous fusions, and the utilization of complementary optical imaginings such as pansharpening, hyperspectral (HS) pansharpening, and spatiotemporal fusion; but also, in heterogeneous fusion such as in LIDAR-optical and synthetic aperture radar (SAR)-optical fusion for applications such as identifying land use and land cover (LULC), object detection, change detection and terrain monitoring [219]. Many data fusion datasets openly available, apart from audio vision fusion, have been provided by the recurring IEEE GRS data fusion contest for land cover classification and semantic urban reconstruction among others [220, 221].

The application of CDS and cognitive control in the context of self-driving cars has already been covered in Section 2.3.1, as we discussed adaptation and cooperation for CAVs. In this section, a more concrete explanation of the state of the art in sensor fusion applied to smart vehicles is intended in noncooperative and not necessarily in a Bayesian framework. Autonomous vehicles have been the subject of extensive research both in the academic and industrial sectors to reduce car fatalities, reduce emissions and improve overall traffic efficiency. The design of a fully automated controller for automobiles is particularly challenging due to the highly complex environment and diversity the system may encounter during deployment [222]. The most common combinations of sensors for data fusion in autonomous vehicles are camera-LIDAR, camera-radar, and camera-LIDAR-radar [223]. These combinations provide complementary information regarding the surrounding environment and are commonly used to perceive objects and people in the surrounding environment for pedestrian, vehicle, and lane detection. Furthermore, a combination of GPS and inertial sensors have been used for navigation as well as in combination with vision for ego positioning [224].

In the context of CIoT the integration of multimodal data from the cyber-physical-social systems (CPSS) is of particular importance as already examined in section 2.3.1 and covered extensively by [33]. Wang *et al.* propose a series of tensor-based decomposition fusion methods as well as a comprehensive data fusion framework for CPSS data. The first tensor method aim to represent and fuse the data into a single unified representation named tensor-based unified fusion (TUF), which can be considered a form of early fusion to a Markov chain or decision process [225]. The second tensor method aims to integrate spatio-temporal elements in a probabilistic way employing a multivariate multi-step transition tensor named "M2T2". The last tensor method relied on multiple multivariate Markov chains interact and inform each other. This Cyber-Physical-Social

transition tensor is named the (CPST2) model and was proposed to fuse the CPSS data in a unified form [225].

Taking a step back and returning to machine learning-based fusion, attention-based architectures have emerged as a powerful tool in recent years, demonstrating remarkable results, particularly in generative models [226]. The core concept behind attention mechanisms is to assign different levels of importance, or "attention weights," to various parts of the input. By doing so, the model can selectively focus on the most relevant features or modalities, enhancing its ability to capture nuanced information that directly contributes to the task at hand. This approach is particularly advantageous in multi-modal fusion, where integrating diverse data types (e.g., audio, video, and text) requires identifying and combining the most critical aspects of each modality.

One notable example of this is the Video-Audio-Text Transformer (VATT) [227], a convolution-free model that leverages the transformer architecture's multi-head attention mechanism. Multi-head attention allows the model to attend to different parts of the input simultaneously, improving its ability to understand complex interactions between multiple data streams. In their work, Akbari *et al.* show that VATT achieves state-of-the-art performance across several tasks, including image recognition from video sequences and waveform-based audio event classification. This highlights the potential of attention-based models not only to process individual modalities effectively but also to fuse them in a way that enhances overall task performance.

The use of transformers to generate rich, high-dimensional latent spaces presents a unified approach to embedding heterogeneous contextual data. By leveraging these latent spaces, attention-based models can generalize well across various downstream tasks, regardless of whether the input consists of visual, auditory, or textual data. This flexibility is particularly valuable for tasks requiring the integration of multiple types of contextual information, such as multi-modal sentiment analysis, video understanding, or audio-visual scene recognition [228].

In the context of data fusion, transformers offer a scalable and adaptable framework for processing and combining diverse data types. As machine learning systems increasingly rely on multi-modal inputs, attention-based architectures stand out as a promising solution for creating embeddings that capture both the shared and distinct features of each modality. In the following section, we will delve deeper into how transformers and attention mechanisms can be harnessed to create effective latent representations for

multi-modal fusion and the role they play in improving downstream task performance [229].

Traditional solutions often require laborious data preprocessing and specialized pipelines to transform heterogeneous inputs such as numerical sensor readings, images, audio, and textual logs into a consistent format. LLMs streamline these processes by automatically harmonizing diverse data types through multimodal learning. By encoding and interpreting textual information, images, and other data modalities in a unified representation space, these models substantially simplify downstream tasks like anomaly detection, predictive maintenance [230], and context-aware decision-making.

2.4.2.2 Embedding LLMs Across CIoT Layers

One of the key advantages of LLMs is their flexibility to operate at various levels of a CIoT architecture. Nonetheless, it is important to highlight that edge deployments do not necessarily require full-scale LLMs. Simpler or specialized encoder models, rule-based agents, or smaller knowledge extraction frameworks may be more suitable for the computational and power constraints found in many edge scenarios.

- **Edge Devices**: At the edge, models (LLMs or otherwise) can perform lightweight inference for tasks such as real-time data preprocessing, encoding sensor readings, or detecting anomalies. By doing so, extensive upstream processing is reduced, and response times are faster [231]. However, deploying any form of AI (including LLMs) on edge devices introduces significant constraints related to power consumption, memory footprint, and compute capacity. Techniques such as quantization, pruning, or knowledge distillation may be employed to reduce model size and power usage without sacrificing too much accuracy [232]. Moreover, recent methods for reducing computational complexity and memory costs have shown promising results [233]. In scenarios where LLMs are too large or expensive to run locally, simpler encoder-based models or rule-based agents can still provide meaningful, context-aware insights while maintaining a small operational footprint.

- **Cloud Platforms**: In the cloud, LLMs can process complex queries, generate comprehensive reports, and handle large-scale analytics [234]. Their ability to unify data streams from diverse sources ensures seamless interoperability between devices and systems. Cloud-based LLM deployments also benefit from virtually unlimited compute resources, making them well-suited for more computationally demanding tasks, such as full-scale language generation or large-batch processing

of device logs. This division of labor between edge and cloud helps maximize efficiency across the CIoT ecosystem.

By embedding LLMs and other suitable AI components throughout the CIoT ecosystem, organizations can simplify data management pipelines and achieve greater scalability and resilience. Depending on power and compute budgets, different parts of the system can leverage different model architectures to balance performance with operational constraints.

2.4.2.3 LLM Architectures and Alternatives

LLMs with encoder-decoder architectures offer a universal solution for tasks requiring both language understanding and generation. These models are particularly well-suited for CIoT systems that need to handle tasks such as summarizing device logs, generating actionable insights, or translating complex error codes into human-readable formats. However, depending on the specific use case and device-level constraints, other architectures or smaller models may be preferable [235]:

- **Encoder-Only Models**: These models are ideal for classification and semantic representation tasks, such as detecting the type of error in device logs. They typically have lower computational overhead, making them more amenable to efficient edge deployments. Techniques like weight pruning or reduced-precision arithmetic can further shrink memory footprints and power usage.
- **Decoder-Only Models**: Decoder-based models are designed for generative tasks, such as creating detailed reports or synthesizing user commands. While they can be deployed on edge devices in simplified forms, they often benefit from offloading compute-intensive portions of the generation process to more powerful cloud resources [236].
- **Encoder-Decoder Models**: Encoder-decoder models (e.g., GPT-like) combine the strengths of both encoder and decoder components, enabling advanced tasks like context-aware summarization and complex query resolution [184]. Edge-friendly adaptations may incorporate smaller encoder-decoder architectures or efficient compression techniques to reduce power draw, while deferring heavier computations to cloud-based pipelines when necessary.
- **Smaller Models and Knowledge Extraction Agents**: In many cases, full-scale LLMs are not strictly necessary on edge devices. Simpler encoder-based

models, specialized neural networks, or agent-based frameworks can extract knowledge from sensor data or logs, focusing on specific tasks (e.g., anomaly detection, rule-based triggers) while minimizing power requirements [237].

This versatility ensures that CIoT systems equipped with the right mix of models can adapt to a wide range of operational demands and power constraints. Strategic partitioning of tasks between on-device intelligence and cloud-based computational capabilities optimizes performance while respecting the limitations inherent to edge deployments. The integration of LLMs into CIoT enhances both decision-making and automation across various industrial and consumer applications. Here are some real-world use cases that demonstrate the potential of LLMs in CIoT:

- **Smart Manufacturing:** LLMs can analyze data from machine sensors, production logs, and operator feedback to identify inefficiencies, predict maintenance needs, and even suggest optimal production schedules. For example, a factory could use an LLM to process sensor readings and historical failure data to predict machine breakdowns, minimizing unplanned downtime and improving overall equipment effectiveness [238].
- **Autonomous Vehicles and Traffic Systems:** LLMs can assist in enhancing autonomous vehicle systems by analyzing traffic data, road conditions, and sensor inputs in real-time [239]. In a connected transportation network, an LLM could interpret sensor data from vehicles and roadside units to predict traffic congestion and dynamically adjust traffic signal timings, reducing congestion and improving traffic flow [240].
- **Healthcare Monitoring Systems:** LLMs can be used to process data from wearable health devices, such as heart rate monitors, glucose sensors, or sleep trackers, to provide personalized health insights and early warnings of potential health issues [241]. For example, an LLM could analyze data from a diabetic patient's glucose monitor and historical trends to alert the patient or healthcare provider about a potential spike in blood sugar, suggesting proactive measures.
- **Supply Chain and Logistics Optimization:** LLMs can analyze data from IoT-enabled supply chain sensors such as RFID tags, GPS trackers, and inventory systems, to optimize stock levels, predict delays, and suggest routing adjustments [242]. In logistics, LLMs could process real-time data on weather conditions, vehicle performance, and traffic reports to optimize delivery routes and schedules, saving both time and fuel.

These applications show how LLMs transform raw data from IoT devices into actionable insights, improving efficiency, safety, and decision-making across industries. Their modularity and ability to handle multi-modal inputs make them powerful tools for enhancing Cognitive IoT systems in both consumer and industrial environments.

AI agents that incorporate LLMs excel at parsing and extracting semantic content from unstructured text, enabling CIoT systems to interpret nuanced information from device logs, user commands, and other textual inputs in real time. By embedding these agents at different layers, ranging from edge devices to centralized cloud platforms, organizations can manage a diverse range of data inputs without relying on overly complex or brittle pipelines. Through foundational pre-training, LLMs provide robust generalization across numerous tasks, reducing the need for siloed or task-specific modules in traditional CIoT architectures. Prompting mechanisms, including instruction-based queries and in-context learning, further enhance the adaptability of LLMs, allowing them to handle zero-shot and few-shot tasks when explicit training data may be limited or unavailable [184]. This unified processing layer (composed of multiple agents) not only segments, normalizes, and contextualizes text dynamically, but also lowers overhead by minimizing redundant preprocessing pipelines. As a result, CIoT deployments benefit from more efficient data integration, faster decision-making, and a broader capacity to accommodate evolving operational requirements.

2.4.2.4 Role of Fine-Tuning

Fine-tuning is a critical process that extends the capabilities of pre-trained models by adapting them to specific tasks or environments. While pre-training equips LLMs with general knowledge from diverse datasets, fine-tuning ensures these models can meet the nuanced requirements of real-world applications. In the context of RL with Human Feedback (RLHF), fine-tuning plays an essential role in aligning the behavior of LLMs with task-specific objectives, user preferences, and operational constraints.

Fine-tuning operates across two primary paradigms:

- Single-Consumer Scenarios: Tailored adaptation to a single user's feedback, enabling the model to specialize in narrowly defined tasks or domains.
- Multi-Consumer Scenarios: Balancing generality and user-specific needs across diverse applications, often requiring modular or parameter-efficient approaches.

By integrating fine-tuning with RLHF, LLMs achieve a balance between specialization and generalization. This synergy allows models to dynamically adapt to both highly

specific and broadly applicable tasks, offering solutions that are scalable, efficient, and responsive to user feedback [243].

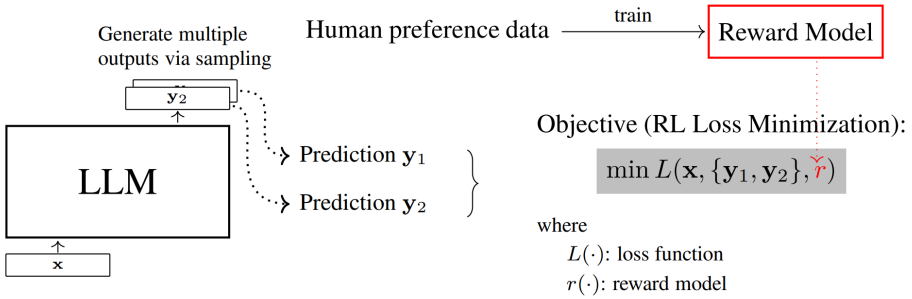


FIGURE 2.14: Schematic illustrating how an LLM is optimized via a reward model trained on human preference data.

RLHF for Single vs. Multi-Consumer Scenarios RLHF has emerged as a powerful method for aligning LLMs with user-defined objectives by incorporating explicit feedback into the training process. The distinction between single and multi-consumer scenarios poses unique challenges for generalization, world modeling, and fine-tuning [243].

Single-Consumer Scenarios In single-consumer settings, RLHF focuses on aligning the LLM's behavior with the specific preferences and objectives of a single user or system. Fine-tuning plays a crucial role here, allowing the model to specialize based on narrowly defined feedback loops.

- **Knowledge Specialization**:By fine-tuning on feedback data from a single consumer, the LLM develops a world model tailored to that consumer's requirements, such as domain-specific terminologies or unique operational patterns.
- **Efficiency and Optimization**: Since the model is optimized for one set of preferences, it can prioritize performance and efficiency for highly specific tasks, such as device-specific command synthesis or custom reporting formats.
- **Risk of Overfitting**: A potential downside is the risk of overfitting to the single consumer's preferences, limiting the model's ability to generalize to new tasks or contexts without further retraining.

Multi-Consumer Scenarios In multi-consumer contexts, RLHF is used to generalize the model's behavior across diverse user needs and preferences. This requires balancing competing objectives and maintaining broad applicability.

- **Generalization Across Preferences**: The model learns a more robust world model that captures commonalities across users while adapting to variations. For example, it can process device logs from various manufacturers while accounting for specific formatting differences.
- **Trade-Offs in Feedback Integration**: Feedback from multiple consumers may conflict, requiring sophisticated reward modeling to balance priorities effectively. RL techniques, such as preference aggregation, are critical here.
- **Scalability of Fine-Tuning**: Fine-tuning in multi-consumer scenarios often involves parameter-efficient techniques, such as LoRA (Low-Rank Adaptation) or adapter layers, to allow simultaneous adaptation without fully retraining the model.

2.4.2.5 Context of Generalization and World Models

The distinction between single- and multi-consumer scenarios highlights the role of RLHF in shaping a model's generalization capabilities and its underlying world model:

- **Single-Consumer World Models**: These are highly detailed and specialized, optimized for efficiency and performance in well-defined contexts.
- **Multi-Consumer World Models**: These models aim to be broader and more flexible, capturing diverse knowledge to support a wide array of applications.

Furthermore, multimodality in meta-learning could be the key to exploiting highly parameterized models for transfer learning to other domains' downstream tasks. Various approaches have been presented for meta-learning, such as optimization-based meta-learning, which includes within-task modality alignment and cross-modality alignment. Extensive surveys on these methods can be found in [244, 245, 246]. Additionally, for the readers interested in the state of the art in LLMs, the following highly influential surveys provide valuable insights: [243, 247, 248, 249, 250, 251].

Pre-trained large-scale models demonstrate strong adaptability to downstream tasks by generalizing previously learned knowledge. However, challenges remain in the data transmission capacity and the processing of heterogeneous data. If no single model can effectively aggregate raw data, it is possible to develop a system capable of utilizing latent spaces generated by lower-level, mode-specific models through data fusion. By

leveraging distributed intelligence, the transmission size and processing load on a central computing model can be significantly reduced, enabling efficient aggregation through the integration of latent space vectors.

2.5 Future Directions

2.5.1 Current Limitations

Despite the significant advancements in CIoT technologies, several critical limitations persist that warrant further exploration and resolution. These limitations include challenges in communication standardization, decentralization, evolving data processing methodologies, and the adaptability of foundation models to application-specific scenarios.

Standardized Communication Protocols: Some authors as detailed in Section 2.2.2.2 emphasize the necessity of standardized communication protocols to manage the massive data transmissions among devices. However, with the increasing prevalence of smart hosts and control interfaces, the requirement for distributed communication among end devices may not be as critical. Smart hosts can connect devices via various networks while managing data flow, effectively reducing the need for direct device-to-device communication. Furthermore, techniques like CTC enable heterogeneous devices using different standards and protocols to interact seamlessly, which may alleviate the urgency of establishing universal communication standards.

Decentralization vs. Edge Computing: Decentralized systems are frequently proposed as a solution for processing large datasets and handling computationally intensive tasks, such as neural network inference. However, the capabilities of modern end devices are evolving quickly, with some equipped with CPUs, GPUs, or specialized NPUs that offer limited on-device acceleration. While these advances enable certain AI tasks (e.g., real-time processing or local inference), overall computational capacity at the edge remains constrained compared to server-based infrastructure. Consequently, edge computing often finds a complementary role rather than replacing decentralized or cloud-based systems entirely. This highlights the ongoing balance between leveraging local, device-level processing for lower-latency tasks and distributing heavier workloads across more powerful, decentralized resources in CIoT environments.

Evolving Data Processing Techniques: While RL, IBM Watson, and data fusion techniques are central cornerstones to CIoT data processing, these methods have seen

reduced relevance in current applications. LLMs and foundation models have emerged as the dominant technology for handling multimodal data, predicting user behaviors, and facilitating natural language interactions. LLMs serve as powerful engines for extracting meaningful insights and generating creative outputs, making them indispensable in modern CIoT systems. Their ability to process diverse data types and provide intuitive user interactions positions them as superior alternatives to earlier methods. This is especially relevant when looking at the use of lower level AI systems or agents aiding the language model make sense of data [237].

2.5.2 Lessons Learned

The insights gained from this study underscore the critical importance of addressing these limitations to advance the CIoT paradigm effectively. While standardized protocols remain important, leveraging technologies like cross-technology communications and smart hosts can reduce the reliance on direct inter-device communication, streamlining the system's operation without compromising efficiency. The growing computational power of end devices highlights the potential of edge computing to process tasks locally, reducing dependency on centralized or decentralized systems. This shift can lead to more efficient and scalable solutions for CIoT applications. The dominance of LLMs in data processing showcases their versatility and capability in handling multimodal data and enabling intuitive interactions. Their integration into CIoT systems can revolutionize data processing, prediction, and decision-making processes. Developing foundation models that strike a balance between generalization and application-specific customization will be essential. Context-aware designs that cater to the unique requirements of different scenarios, such as households versus commercial spaces, will enhance the practical utility of CIoT systems. Addressing these challenges and leveraging the opportunities they present will guide the evolution of CIoT systems towards greater adaptability, efficiency, and user-centricity.

2.5.3 Forward Looking Statements

CIoT is poised to benefit from the rapid evolution of microcomputing and CPS. In the near future, technology will integrate more seamlessly into everyday life, demanding systems that can adapt to both human needs and the inherent uncertainties of dynamic environments. Although current implementations are limited—especially in terms of integrating robust cognitive functions—the trajectory is clear: CIoT architectures must evolve to accommodate flexible, distributed processing and the dynamic integration of foundation models.

Key areas for future research include:

- **Scalable Communication Frameworks**: Investigating how adaptive communication layers can be standardized to support diverse device ecosystems without sacrificing performance.
- **Hybrid Processing Architectures**: Exploring novel architectures that balance edge, fog, and cloud computing, including parallel processing strategies to overcome the limitations of current hardware.
- **Deployable Foundation Models on the Edge**: Developing methods for running large-scale models in resource-constrained environments, leveraging techniques from federated learning and TinyML to enable distributed intelligence.
- **Integrated Cognitive Frameworks**: Creating systems that seamlessly merge data abstraction with real-time decision-making, paving the way for multi-level intelligence in CIoT environments.

By addressing these limitations through targeted research and innovative design, future CIoT systems can achieve greater adaptability, efficiency, and intelligence—paving the way for a truly integrated technological symbiosis.

2.6 Conclusions

This paper has explored the emerging field of CIoT and its potential to revolutionize various domains such as smart homes, smart vehicles, and smart cities. The study has contributed to a deeper theoretical understanding of CIoT by examining its underlying mechanisms and exploring its applications across diverse IoT subdomains. Practically, CIoT holds the promise of transforming the technological landscape by enhancing the functionality, adaptability, and efficiency of IoT systems. By tracing the evolution of cognitive processes in engineering, from early theories by Haykin and Fuster to modern interpretations of IoT and CIoT, this work has positioned CIoT as a framework that addresses longstanding challenges in IoT systems, such as transmission limitations, scalability, and data fusion. The integration of foundation models, alongside multi-level deployment of machine learning algorithms through technologies like TinyML, federated learning, and cloud and edge computing, underscores a novel approach within the CIoT paradigm.

The findings presented here shed light on CIoT's capacity to address current issues in IoT data processing through cognitive computing and distributed architectures. Moreover, the study emphasizes the importance of standardization and interoperability for developing large-scale, holistic IoT and CIoT architectures. The potential of CIoT to facilitate intelligent decision-making suggests significant improvements in the performance of engineering systems, as its contextual understanding and higher-level abstraction capabilities can lead to more efficient and effective IoT architectures. Despite these promising developments, the research acknowledges that the field of CIoT is still in its infancy. Numerous challenges remain, and substantial work is needed to fully realize its potential. In response to these limitations, this study outlines promising research pathways that include the development of standardized communication protocols, the exploration of parallel processing strategies, the refinement of cognitive radios for IoT, and the advancement of foundational models and multi-level intelligence architectures.

By linking theoretical insights with practical applications, this work provides a foundation for future research and development in CIoT. The continued pursuit of these research avenues is essential to overcome current limitations and to harness the full potential of CIoT in shaping the future of intelligent, interconnected systems.

Chapter 3

How VAE Latent Spaces can be Utilized for Direct Integration

The content of this chapter is a reformatted version of the manuscript text published under the following citation:

A. Giuliano, S. Andrew Gadsden and J. Yawney, "Optimizing Satellite Image Analysis: Leveraging Variational Autoencoders Latent Representations for Direct Integration," in IEEE Transactions on Geoscience and Remote Sensing,
doi: 10.1109/TGRS.2024.3520879.

Optimizing Satellite Image Analysis: Leveraging Variational Autoencoders Latent Representations for Direct Integration

Alessandro Giuliano
Faculty of Engineering
McMaster University, Hamilton, ON, Canada
Email: giuliana@mcmaster.ca

S. Andrew Gadsden
Faculty of Engineering
McMaster University
Email: gadsdesa@mcmaster.ca

John Yawney
Adastra Corp.
Email: john.yawney@adastragrp.com

Abstract

Variational Autoencoders (VAEs) have emerged as powerful tools for data compression and representation learning. In this study, we explore the application of VAE-based neural compression models for compressing satellite images and leveraging the latent space directly for downstream machine learning tasks, such as classification. Traditional approaches to image compression require decoding the compressed format for subsequent analysis. However, we propose that the latent representation constructed by these models can be utilized directly by another machine learning model without explicit reconstruction, or inverse transform. We utilize latent spaces derived from neural compression model-encoded Sentinel-2 images for downstream classification tasks. We demonstrate the viability and flexibility of this approach, showcasing the impact of fine-tuning the neural compression models to further increase classification performance, achieving the same accuracy as state-of-the-art models at lower bitrates. By training these models to

compress satellite images into a low-dimensional latent space, we show that the latent representations capture meaningful information about the original images, facilitating accurate classification without the overhead of reconstruction. Our results highlight the potential of neural compression methods for direct satellite image analysis, offering a promising avenue for efficient data transmission and processing in remote sensing applications.

Keywords: *Remote Sensing, Variational Autoencoders, Neural Compression*

3.1 Introduction

Remote sensing plays a critical role in numerous applications such as environmental monitoring, disaster management, and agricultural planning. Satellites like Sentinel-1,2,3,5 provide high-resolution imagery essential for these tasks. However, the sheer volume of data generated by these satellites presents significant challenges for storage, transmission, and analysis.

Traditional image compression techniques aim to reduce data size while preserving visual quality. These methods typically involve quasi-lossless or lossless compression algorithms that require the compressed images to be decompressed before any analysis can be performed. While effective for reducing storage requirements, this two-step process of compression and decompression can be computationally expensive and inefficient, especially when real-time analysis is needed.

Compressed data is relayed back to Earth, preprocessed and analyzed to be used for a wide variety of tasks. In recent years deep learning based image analysis has grown in popularity and represents the state of the art for many image analysis tasks [252], including image fusion [253], image registration [254], scene classification [255] , object detection [256], land use and land cover (LULC) classification [257], segmentation [258], and object-based image analysis (OBIA) [259].

With the advancement of machine learning new compression methods also emerged. Neural compression techniques, particularly those based on Variational Autoencoders (VAEs), have shown great promise for compressing complex data while retaining meaningful features in the compressed representation. Neural compression has been shown to outperform conventional compression methods such as JPEG on a broad scale [260, 261, 262], and specifically in compressing satellite images as well [263]. VAEs are generative

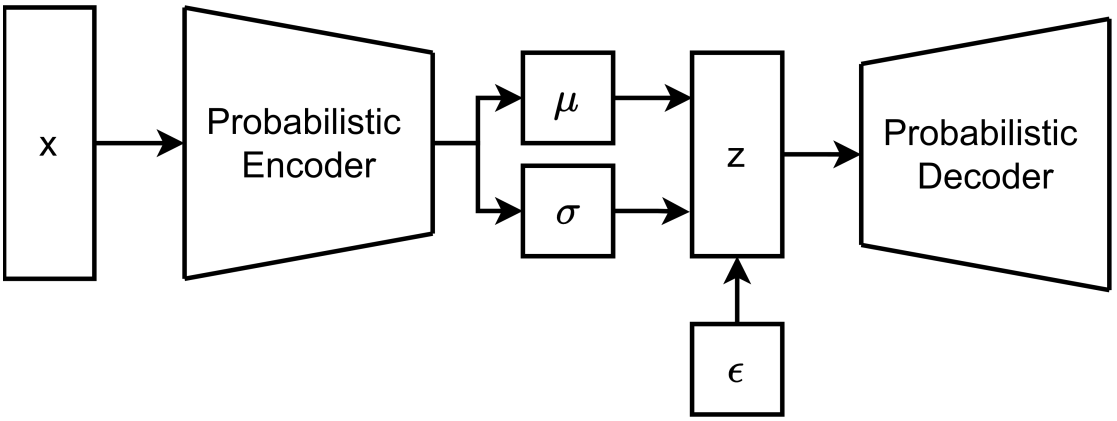


FIGURE 3.1: General architcture of a variational autoencoder (VAE).

models that learn to encode input data into a lower-dimensional latent space iteratively through composite optimization and then decode it back to the original format. The latent space, a compressed representation of the input data, captures essential information and underlying structures through multiple non linear projections.

A novel approach to neural compression utilization is to leverage the latent spaces generated by the transformation of the original data directly for downstream tasks, bypassing the need for reconstruction. This can potentially streamline the process, reducing computational overhead and enhancing efficiency. The hypothesis is that the latent representations produced by VAEs are rich enough to serve as inputs for machine learning models, such as classifiers, thereby facilitating direct and effective analysis of compressed data.

This study aims to investigate the viability of using neural compression derived latent spaces specifically for direct classification of Sentinel-2 satellite images. By fine-tuning these models, we seek to understand the impact on the quality of the latent space, classification accuracy and reconstruction quality performance. The approach is validated through extensive experiments on Sentinel-2 satellite images, demonstrating the validity of the approach and flexibility of using neural compression model-derived latent spaces for downstream machine learning tasks. Demonstrating the effectiveness of this approach could significantly advance the field of remote sensing, offering a more efficient method for handling large-scale satellite imagery.

This paper makes the following key contributions:

- The study proposes using the latent representation constructed by neural compression models directly for downstream machine learning tasks, such as classification. Without the need for explicit reconstruction or inverse transform, the learned latent space can be leveraged immediately for downstream tasks like classification, making the process more efficient. The results validate the possibility of using the lower dimensional representation of the data directly in contrast with conventional methods.
- It demonstrates the viability and flexibility of this approach through experiments with latent spaces derived from neural compression model-encoded Sentinel-2 images, specifically for classification tasks.
- The impact of fine-tuning state-of-the-art compression models is examined, showing how it can enhance downstream task accuracy while maintaining compression performance.
- A new metric is introduced for evaluating the Rate-Distortion-Accuracy tradeoff, providing a comprehensive measure that balances compression efficiency, reconstruction quality, and classification performance.

To the best of the authors' knowledge, this is the first attempt at leveraging neural compression latent spaces for direct use in downstream machine learning tasks. By eliminating the need for explicit decompression, this approach significantly reduces computational overhead, making it possible to develop more efficient and streamlined architectures. Furthermore, by transmitting the latent space representation of the original data, an additional security layer is implemented in terms of masking the original data through nonlinear transforms. Masking the data makes it more difficult for unauthorized parties to reconstruct the original data without access to the specific decoding model or to make use of it without models trained explicitly using clear data.

Unlike traditional image compression methods that require full reconstruction for analysis, our approach leverages the latent representations directly for classification tasks. This eliminates the need for the inverse transform, significantly reducing computational overhead and improving efficiency in real-time applications. The ability to use compressed representations directly could lead to advancements in various fields that rely on large-scale data analysis, such as remote sensing, environmental monitoring, and urban planning. This study sets a precedent for future research to explore and optimize the integration of neural compression models with machine learning frameworks, ultimately aiming to achieve higher performance and greater scalability.

The rest of the paper is structured as follows: Section 3.2 covers conventional data compression methods, neural compression and how neural compression has been applied for satellite imagery. Section 3.3 covers dataset used, neural compression and classification models, latent space representation techniques, fine tuning methods, as well as overall architecture, experimental setup and evaluation metrics. Section 3.5 covers qualitative and quantitative results, a comparison between baseline and fine tuning, latent space visualizations, and rate distortion accuracy plots. Section 3.6 covers insights, advantages, limitation and future work as well as security considerations. Finally Section 3.7 concludes paper remarking the findings.

3.2 Related Work

3.2.1 Deep Learning in Satellite Image Analysis

The use of deep learning in satellite image analysis has gained significant attention in recent years, driven by the increasing availability of high-resolution satellite imagery and advances in neural network architectures. These techniques have been applied to a wide range of tasks, including land cover classification, object detection, change detection, and image segmentation.

Recent advancements in hyperspectral imaging have further highlighted the significance of fusion-aware computational techniques for improving image quality, particularly in complex scenes. For instance, the CasFormer model proposed by Li et al. introduces a novel cascaded transformer architecture that effectively enhances hyperspectral imaging by integrating RGB and spectral data through spatial coherence alignment and spectral recovery. This approach demonstrates state-of-the-art performance by achieving high spatial consistency and spectral fidelity, which is crucial for applications in environmental monitoring, medical diagnosis, and remote sensing [264].

Interpretability and robustness remain critical challenges in hyperspectral anomaly detection, particularly in complex environments. To address these issues, Li et al. introduce a novel paradigm that integrates model-driven low-rank representation (LRR) methods with data-driven deep learning techniques. By leveraging disentangled priors (LDP), their approach effectively separates explicit from implicit priors, resulting in improved detection accuracy and enhanced generalization across a variety of hyperspectral datasets. This integration of explicit low-rank priors with implicitly learned features represents a significant advancement in hyperspectral anomaly detection [265].

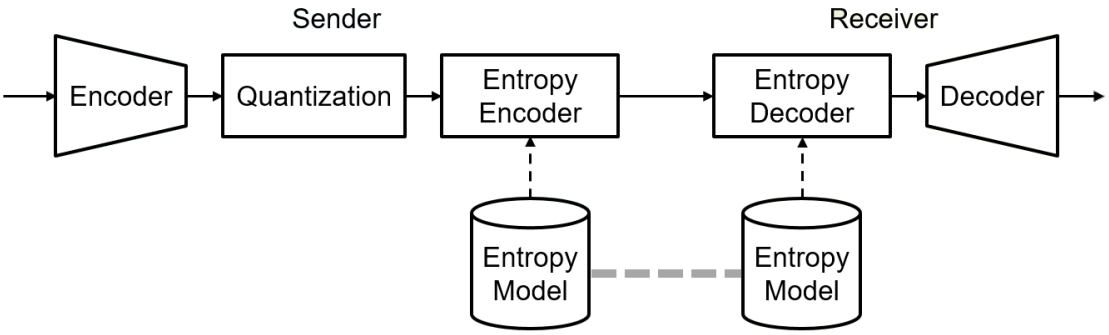


FIGURE 3.2: General neural compression model structure.

Additionally, the authors propose a deep unfolding solution, LRR-Net+, which fuses low-rank representation with deep learning to bolster anomaly detection while maintaining interpretability. This approach bridges the gap between data-driven and model-driven methods by incorporating the Alternating Direction Method of Multipliers (ADMM) optimizer, which strikes an effective balance between performance and explainability in hyperspectral anomaly detection tasks. The introduction of the AIR-HAD dataset further demonstrates the robustness of this method across diverse scenarios [266].

Addressing the challenges of generalization in remote sensing tasks across different urban environments, Hong et al. introduce the CrossCity multimodal dataset, which facilitates cross-city semantic segmentation through domain adaptation techniques. The authors present the HighDAN network, a high-resolution domain adaptation framework that improves the transferability of learned models between cities by minimizing domain shifts. This work is particularly relevant for tasks involving multimodal data integration and has set a new benchmark for semantic segmentation across diverse geographical regions [267].

The emergence of foundation models has also opened new avenues in remote sensing, particularly in spectral data analysis. The SpectralGPT model, introduced by Hong et al., represents a significant advancement by tailoring the generative pretrained transformer (GPT) architecture to spectral remote sensing. This model excels in handling large-scale spectral data through its innovative 3D masking strategy and multi-target reconstruction, leading to superior performance across various downstream tasks such as classification, segmentation, and change detection. SpectralGPT's ability to generalize across diverse datasets highlights its potential in advancing remote sensing applications [268].

3.2.2 Conventional Satellite Data Compression Techniques

Traditional methods for satellite data compression have relied heavily on both lossless and lossy compression techniques to reduce the volume of data while attempting to preserve the essential information needed for analysis. These conventional techniques include methods like JPEG, JPEG2000, and other domain-specific algorithms.

Lossless compression techniques, such as Run-Length Encoding (RLE) [269], Huffman Coding [270], and Lempel-Ziv-Welch (LZW) [271], aim to reduce data size without any loss of information. These methods are particularly valuable in applications where the integrity of the original data must be maintained, such as in scientific and technical imagery analysis. However, the compression ratios achieved by these methods are limited by Shannon theoretical lossless compression limit, equivalent to the corresponding entropy of the data being transmitted. Lossy compression techniques, such as JPEG and JPEG2000, provide higher compression ratios by allowing some loss of information [272]. These techniques are designed to exploit the human visual system's limitations, removing less noticeable details to achieve more significant data reduction. JPEG2000, in particular, has been widely used in satellite image compression due to its ability to offer higher compression ratios and better quality at lower bitrates compared to standard JPEG. Satellites also use several near-lossless image compression techniques, such as multi-stage vector quantization (SAMVQ) and cluster vector quantization (HSOCVQ). These methods aim to limit compression errors to levels comparable to the intrinsic noise of the original data, reducing the impact on remote sensing applications [273]. SAMVQ works by organizing 2D focal plane frames into regional datacubes, which are then split into subsets for parallel processing. Each subset is classified based on spectral similarity, enabling faster processing and better memory use through independent compression. Conversely, HSOCVQ classifies the entire datacube into clusters based on spectral similarity rather than splitting it into smaller sections. This technique enhances compression by grouping similar spectra, aligning clusters with specific scene targets. Predictive coding methods, such as Differential Pulse Code Modulation (DPCM) and the more advanced Context-based Adaptive Binary Arithmetic Coding (CABAC) used in H.264/AVC, predict pixel values based on neighboring pixels, encoding only the prediction errors [274]. These methods are effective in reducing redundancy in satellite images, which often contain large areas of homogeneous regions.

3.2.3 Neural Compression

Recent advancements in image compression using neural networks have shown significant improvements over traditional methods [260]. Various architectures and techniques have been proposed to address specific challenges in this field.

Neural compression employs neural networks and machine learning techniques to optimize data compression processes, specifically in the transform phase. This innovative approach utilizes deep generative models such as variational autoencoders (VAEs) [260], generative adversarial networks (GANs)[275], normalizing flows [276], and autoregressive models [277] to learn compression algorithms directly from data end to end. By capturing complex data distributions, neural compression can significantly enhance the efficiency and effectiveness of data reduction compared to traditional methods such as discrete cosine transform (DCT) used in JPEG coding [272]. Different entropy models, including fully factorized models [261], and hyperprior models [278], have been explored to improve the rate-distortion performance of learned image compression methods. The selection of an appropriate entropy model is essential for optimizing bit-rate and maintaining high-quality reconstructions.

The core idea behind neural compression is the replacement of linear transforms with neural network-based nonlinear transforms, enabling more flexible and adaptive data representations. The field has evolved rapidly since the introduction of deep generative models for data compression around 2016, when the parallels between variational inference and both lossless and lossy compression methods came to light. The introduction of hyperprior models [278], discretized Gaussian mixture likelihood models [262], hierarchical structures [279] and more [280, 281], has significantly advanced image compression performance since. These models estimate the likelihoods of the latent representations more accurately by capturing spatial dependencies, thereby enhancing the entropy coding process. The hyperprior approach, originally proposed for natural images, has been adapted and optimized for satellite imagery, showing considerable improvements in compression efficiency [263].

A typical neural compression pipeline involves transforming the input data into a lower-dimensional latent space using an encoder, followed by quantization and entropy coding to achieve a compressed representation, as seen in Figure 3.2. The decoder then reconstructs the data from this compressed form. This end-to-end learning process optimizes both the bit rate and the distortion, balancing compression efficiency with the quality of the reconstructed data. Additionally, advancements in neural compression

often integrate perceptual metrics and adversarial losses to enhance the realism and perceptual quality of the reconstructions.

Neural compression's flexibility is particularly beneficial for new and domain-specific data types where traditional codecs fall short. However, challenges remain, such as optimizing neural architectures, managing the trade-offs between distortion and realism, and addressing the specific requirements of various data types. Despite these challenges, neural compression holds significant promise for revolutionizing data compression through its data-driven, adaptive methodologies [260].

3.2.4 Neural Compression in Satellite Images

In recent years, neural compression has been applied to satellite imagery and is starting to become a topic of interest, although there are still a limited number of studies on the subject. The most relevant is the one by Oliveira et al. which propose a reduced-complexity VAE tailored for on-board satellite image compression [263], addressing time and memory constraints while maintaining performance. This approach simplifies the entropy model by leveraging a statistical analysis that shows most features follow a Laplacian distribution, thus replacing complex non-parametric models with a simpler parametric estimation. The proposed model outperforms the Consultative Committee for Space Data Systems (CCSDS) standard and remains competitive with state-of-the-art learned compression schemes. Also a few more have attempted to use neural compression in satellite images with promising results [282, 283]. Despite these advancements, the challenge of efficiently compressing and transmitting fast-increasing large volumes of satellite data remains. This study builds on the existing body of work by exploring the application of Variational Autoencoders (VAEs) and other advanced neural compression models for satellite image analysis, aiming to leverage the latent space directly for downstream machine learning tasks.

3.3 Methodology

3.3.1 Proposed Architecture

While neural compression techniques, such as those proposed in [261, 278, 279, 262, 280], have significantly advanced the field of image compression, their primary focus has been on optimizing the compression and reconstruction stages. These methods, although effective in minimizing data loss during compression, typically require decompression

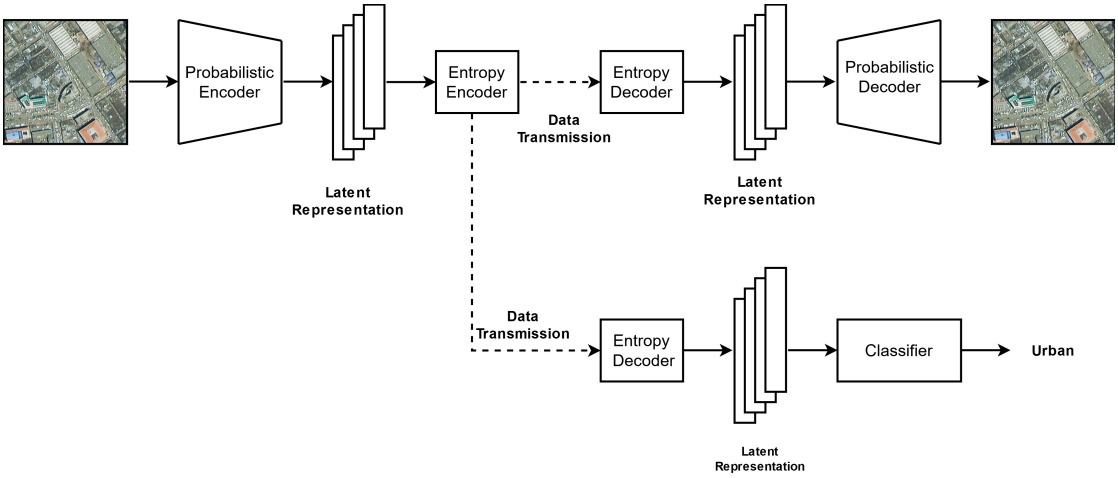


FIGURE 3.3: Pipeline diagram for direct utilization of neural compressed images.

before the data can be utilized for downstream tasks such as classification or segmentation. In our approach, we build upon these models and demonstrate that they can be adapted to function without the need for decompression. Specifically, we fine-tune these models to show that the latent spaces they generate can be directly exploited for downstream tasks. Our method thus diverges from the traditional workflow by utilizing the latent representations for classification tasks, thereby eliminating the need for reconstruction or inverse transformation and enhancing the overall efficiency of the process. In this study, we propose a novel architecture that integrates neural compression and classification within a unified framework.

To illustrate the difference between the proposed architecture and conventional methods, Figure 3.2 presents a typical pipeline for a traditional compression workflow. In this pipeline, the original image undergoes a transformation (e.g., Discrete Cosine Transform for JPEG or Discrete Wavelet Transform for JPEG2000), followed by quantization and encoding through an entropy encoder. The entropy-coded data is then transmitted through a channel, decoded by an entropy decoder, and transformed back into the original image via an inverse DCT or a trained decoder, as in neural compression. Only after this final reconstruction step can the image be used for downstream tasks, such as classification.

In contrast, as shown in Figure 3.3, we propose that the lower-dimensional transformation of the original image, produced by neural compression, can be directly used for downstream tasks such as classification, once entropy decoding has restored the data

to a tensor format. Here, the neural network itself performs the transformation, generating a compact, lower-dimensional 'latent representation'. This representation is then quantized, encoded using an entropy encoder, transmitted, and decoded by the entropy decoder on the other end of the transmission channel. Without the need for explicit reconstruction or inverse transformation, the learned latent space can be immediately leveraged for downstream tasks like classification, making the process more efficient.

The proposed architecture comprises two key components: a neural compression model and a classification model. The neural compression model, built on advanced aforementioned VAE architectures, compresses high-dimensional satellite images into compact latent representations that preserve essential features for classification. The classification models are trained directly on these latent representations, effectively distinguishing between different classes of satellite images. We employ various classification models, each suited to different characteristics of the latent space: MLPs were selected due to their straightforward architecture and basic feedforward neural network model, CNNs for spatial pattern recognition, and Transformers for handling complex dependencies. This diverse set of classifiers enables a comprehensive evaluation of our approach across multiple architectures. The full training process is detailed in Algorithm 1 below.

To classify satellite images based on the latent spaces created by neural compression models, we employed the described architecture where we first trained six advanced Variational Autoencoder (VAE) models for image compression: *bmshj2018_factorized*, *bmshj2018_hyperprior*, *mbt2018_mean*, *mbt2018*, *cheng2020_anchor*, and *cheng2020_attn*, from the CompressAI library (official port of Tensorflow neural compression library) [278, 262, 284]. The pretrained models weights were frozen at first to maintain the integrity of the learned latent spaces. Subsequently, we used these frozen models to generate latent representations of satellite images. These latent representations were then used as inputs to three different classification models: a Convolutional Neural Network (CNN), a Multi-Layer Perceptron (MLP), and a Transformer.

Mathematically, let x represent an input satellite image, and $E(\cdot)$ and $D(\cdot)$ denote the encoder and decoder of the compression models, respectively. The latent representation z is obtained as:

$$z = E(x) \tag{3.1}$$

For classification, we denote the classification model as $C(z; \theta_C)$, where θ_C are the trainable parameters of the classification model. The output class probabilities $\hat{y}$ are given by:

$$\hat{y} = C(z; \theta_C) \tag{3.2}$$

The training process involves minimizing a classification loss function $\mathcal{L}_{\text{class}}$ (e.g., cross-entropy loss) between the predicted class probabilities $\hat{y}$ and the true labels y:

$$\mathcal{L}_{\text{class}} = -\sum_i y_i \log(\hat{y}_i) \tag{3.3}$$

Algorithm 1 Training the Classifier on Latent Spaces

Require: Pre-trained and frozen VAE model with encoder E and decoder D
Require: Satellite image dataset $\{(x_i, y_i)\}_{i=1}^{N}$
Require: Classification model: Classifier
Ensure: Trained classification model: θ_C
Freeze weights of the VAE model
for each image x_i in dataset **do**
    Generate latent representation $z_i = E(x_i)$
Initialize classification model C with parameters θ_C
repeat
    Sample a batch of latent representations $\{z_b\}$ and labels $\{y_b\}$
    Compute class probabilities $\hat{y}_b = C(z_b; \theta_C)$
    Compute classification loss $\mathcal{L}_{\text{class}}$
    Update θ_C to minimize $\mathcal{L}_{\text{class}}$
until convergence
Save trained parameters θ_C

3.3.2 Fine Tuning

To further improve classification performance while maintaining low bit rates for reduced latency and transmission size, the models were fine-tuned to improve classification performance while retaining reconstruction capabilities and low bitrate. To do so the fine tuning loss function was a composite objective containing a term for each objective tempered by a multiplier, as shown in Algorithm 2. The reconstruction loss ($\mathcal{L}_{\text{rec}}$) ensures that the VAE retains the ability to accurately reconstruct the original satellite images from the latent space. The classification loss ($\mathcal{L}_{\text{class}}$) drives the VAE to create latent representations that are useful for predicting the correct class labels. The bit rate loss ($\mathcal{L}_{\text{bpp}}$) encourages the VAE to maintain an efficient compression, reducing the size of the latent representations. This composite loss is formulated using Lagrangian multipliers $\lambda_{\text{rec}}, \lambda_{\text{class}}, \lambda_{\text{bpp}}$ to balance the trade-offs between these objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{class}}\mathcal{L}_{\text{class}} + \lambda_{\text{bpp}}\mathcal{L}_{\text{bpp}} \tag{3.4}$$

Algorithm 2 Fine-tuning VAE and Training the Classifier on Latent Spaces

Require: Pre-trained VAE model with encoder E and decoder D
Require: Satellite image dataset $\{(x_i, y_i)\}_{i=1}^{N}$
Require: Classification model: Classifier
Require: Weights for loss terms: $\lambda_{\text{rec}}, \lambda_{\text{class}}, \lambda_{\text{bpp}}$
Ensure: Fine-tuned VAE and trained classification model: $\theta_E, \theta_D, \theta_C$
1: Initialize VAE model with parameters θ_E, θ_D
2: Initialize classification model C with parameters θ_C
3: **repeat**
4: Sample a batch of images $\{x_b\}$ and labels $\{y_b\}$
5: Generate latent representations $\{z_b\} = E(\{x_b\})$
6: Reconstruct images $\{\hat{x}_b\} = D(\{z_b\})$
7: Compute class probabilities $\hat{y}_b = C(\{z_b\}; \theta_C)$
8: Compute reconstruction loss $\mathcal{L}_{\text{rec}} = \|x_b - \hat{x}_b\|^2$
9: Compute classification loss $\mathcal{L}_{\text{class}} = -\sum_i y_{b,i} \log(\hat{y}_{b,i})$
10: Compute bit rate loss $\mathcal{L}_{\text{bpp}} = R(z_b)$
11: Compute total loss $\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{class}}\mathcal{L}_{\text{class}} + \lambda_{\text{bpp}}\mathcal{L}_{\text{bpp}}$
12: Update $\theta_E, \theta_D, \theta_C$ to minimize $\mathcal{L}_{\text{total}}$
13: **until** convergence
14: Save fine-tuned parameters θ_E, θ_D and trained parameters θ_C

where:

$$R(z_b) = \frac{\log(\text{latent}).\text{sum}()}{-\log(2) \cdot \text{num\_pixels}}$$

By incorporating a composite loss function that balances reconstruction accuracy, classification performance, and bit rate efficiency using Lagrangian multipliers, we ensure that the VAE is fine-tuned to produce latent spaces that are both efficient and highly discriminative, the change in latent space composition is shown in Section 3.5 through t-SNE projections. This integrated training approach allows for a more comprehensive optimization, where the VAE and classifier are jointly trained to maximize overall performance. As a result, the latent spaces generated by the fine-tuned VAE are optimized for multiple objectives, making them particularly effective for classification tasks while still benefiting from the compactness and efficiency of the VAE-based compression. This approach to temper composite loss functions has been seen, for example, in Beta-VAEs. By incorporating the Beta-VAE, the latent spaces generated are not only compact and efficient but also disentangled, which is particularly beneficial for classification tasks. The β parameter in the Beta-VAE acts similarly to a Lagrangian multiplier, balancing

the trade-off between reconstruction fidelity and latent space regularization [285].

3.3.3 Latent Space Visualization

To visualize and compare the latent spaces generated by the VAE, we utilized t-distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is a powerful dimensionality reduction technique that is particularly well-suited for visualizing high-dimensional data in a lower-dimensional space, typically two or three dimensions. It is similar to Stochastic Neighborhood Embedding but works better on high dimensional data composed several low-dimensional manifolds such as the ones produced by VAE for multiclass images [286].

t-SNE operates by converting the similarities between data points in the high-dimensional space into joint probabilities and then tries to optimize the low-dimensional representation to preserve these similarities. This is achieved through a probabilistic approach where a Gaussian distribution represents the similarity between two points in the high-dimensional space, while in the low-dimensional space, it is represented by a Student's t-distribution with one degree of freedom (a Cauchy distribution). This choice of distribution in the low-dimensional space helps to manage the so-called "crowding problem," where too many points are mapped too closely together.

Mathematically, t-SNE works as follows:

For each pair of points (i, j), t-SNE calculates the conditional probability $p_{j|i}$ that point j would be chosen as a neighbor of point i if neighbors were picked in proportion to their probability density under a Gaussian centered at i:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{3.5}$$

where x_i and x_j are the high-dimensional input data points, and σ_i is the variance of the Gaussian centered at x_i.

The joint probability p_{ij} is then symmetrized:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{3.6}$$

where N is the number of data points.

In the low-dimensional space, a similar joint probability q_{ij} is computed using a Student's t-distribution:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{3.7}$$

where y_i and y_j are the low-dimensional representations of the data points.

t-SNE aims to minimize the Kullback-Leibler (KL) divergence between the high-dimensional and low-dimensional joint probabilities:

$$\mathrm{KL}(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.8}$$

This is typically done using gradient descent, where the positions of the points in the low-dimensional space are iteratively adjusted to minimize the KL divergence.

t-SNE is particularly suitable for our application because it effectively captures the local structure of the data, making it easier to identify clusters and patterns within the latent space. By visualizing the latent spaces using t-SNE, we can gain insights into how well the VAE has learned to represent the data and how distinct the latent representations are for different classes. This visualization helps in evaluating the quality of the latent space and the effectiveness of the VAE model in capturing essential features of the satellite images.

In our study, t-SNE was used to project the high-dimensional latent representations into a 2D space using perplexity = 30, and 500 iterations, where each point represents a data sample's latent vector. The resulting 2D plot provides a visual comparison of the latent spaces, highlighting the clustering of data points based on their class labels, as shown in Figure 3.4.

3.3.4 The Rate Distortion Accuracy Index

To compare the efficacy of fine-tuning and assess the results of various neural compression and classification models on both reconstruction quality and classification performance, we introduce the Rate Distortion Accuracy Index (RDAI) inspired by the works of Luo et al. [287] on the rate distortion accuracy tradeoff in JPEG. This novel metric integrates the critical aspects of rate, distortion, and accuracy, providing a comprehensive evaluation framework for neural compression methods. The RDAI is defined as a weighted

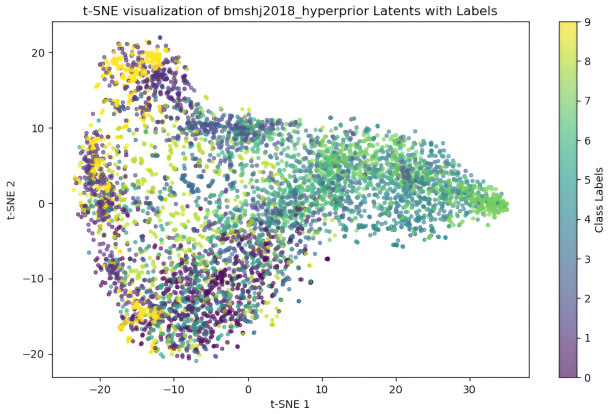


FIGURE 3.4: Sample t-SNE manifold of testing set of the EuroSAT dataset encoded using a VAE with a scale hyperprior, image patches from Sentinel-2 mission.

combination of rate, distortion, and accuracy. Given that BPP will be between 0 and 10, PSNR between 0 and 60, and F1 between 0 and 1, we normalize PSNR to the range [0, 1] as follows:

$$\text{PSNR}_{\text{normalized}} = \frac{\text{PSNR}}{60} \tag{3.9}$$

The formula for RDAI is then given by:

$$\text{RDAI} = \alpha \cdot \left(\frac{10 - \text{BPP}}{10}\right) + \beta \cdot \left(\frac{\text{PSNR}}{60}\right) + \gamma \cdot \text{F1} \tag{3.10}$$

where:

- α, β, and γ are the weights assigned to each component, such that $\alpha + \beta + \gamma = 1$. During this study, we gave all three components equal weight and importance, and therefore $\alpha = \beta = \gamma = \frac{1}{3}$.
- PSNR is the Peak Signal-to-Noise Ratio, measuring the distortion. Higher peak signal-to-noise ratio (PSNR) values indicate better preservation of image quality based on the mean squared error of all pixels.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right) \tag{3.11}$$

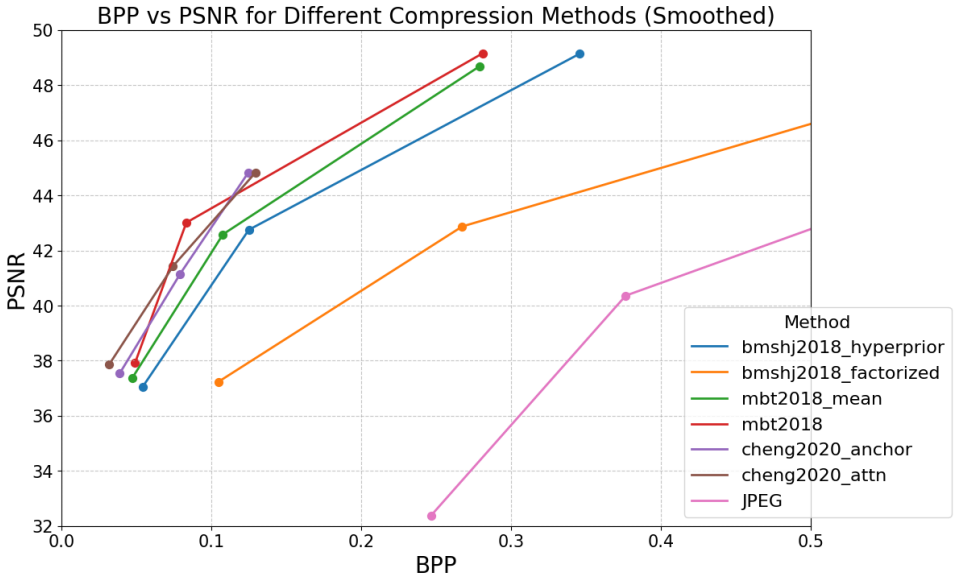


FIGURE 3.5: Models compression performance comparison performed on EuroSAT test set.

where MAX is the maximum possible pixel value of the image (e.g., 255 for an 8-bit image), and MSE is the mean squared error between the original and compressed image. In our case we use MAX = 1 since the images are transformed to float32 tensors during preprocessing.

- F1 is the F1 score of the neural network on the compressed images, used as our accuracy metric. Higher F1 scores indicate better classification performance.

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3.12}$$

The F1 score is preferred over accuracy for evaluating the performance of a neural network on compressed images due to its robustness in handling imbalanced classes. Accuracy measures the proportion of correct predictions out of all predictions, which can be misleading when class distribution is imbalanced. The F1 score, being the harmonic mean of precision and recall, balances false positives and false negatives, providing a more comprehensive assessment of the model's performance.

3.4 Experimental Setup

3.4.1 Hardware

All of the experiments were carried out with PyTorch and CUDA cores using GPU NVIDIA GeForce RTX 4070 Ti Super (16 Gb), CPU Intel(R) Core(TM) i7-14700KF, 3400 Mhz, 20 Core(s), 28 Logical Processor(s), and RAM DDR5 128 Gb at 6000 Mhz. Tensorboard was used to log all results including time series, scalar and image data.

3.4.2 Neural Compression Models Used

In this study, we employ a variety of advanced Variational Autoencoder (VAE) architectures and their derivatives to achieve efficient image compression and examine the utilization of the constructed latent space. The models examined include *bmshj2018_factorized*, *bmshj2018_hyperprior*, *mbt2018_mean*, *mbt2018*, *cheng2020_anchor*, and *cheng2020_attn* extensively pretrained to compress images. Each of these models leverages different techniques to encode images into compact latent representations while preserving essential information for accurate reconstruction and downstream tasks.

The *bmshj2018_factorized* model employs a fully factorized density model for latent variables. Each latent variable y_i is assumed to be independent and identically distributed (i.i.d.), such that

$$p(y) = \prod_i p(y_i) \tag{3.13}$$

This model uses a non-parametric piecewise linear density to approximate each factor of the prior. The density p is defined using its cumulative distribution function (CDF) c, where

$$p(x) = \frac{\partial c(x)}{\partial x} \tag{3.14}$$

By ensuring that the cumulative function c is monotonic and maps $\mathbb{R}$ to $[0, 1]$, a valid density function can be constructed as:

$$c = f_K \circ f_{K-1} \circ \cdots \circ f_1 \tag{3.15}$$

with

$$p = f'_K \cdot f'_{K-1} \cdot \cdots \cdot f'_1$$

where f_k are vector functions composed of matrices $H^{(k)}$, biases $b^{(k)}$, and element-wise nonlinearities g_k defined as:

$$g_k(x) = x + a^{(k)} \odot \tanh(x) \tag{3.16}$$

where $a^{(k)}$ are vectors controlling the expansion or contraction rate.

The *bmshj2018_hyperprior* model introduces a hyperprior to capture spatial dependencies among the elements of the latent representation y. This is achieved by an auxiliary autoencoder which models the latent scales using another set of latent variables z, defined as

$$z = h_a(y; \phi_h) \tag{3.17}$$

and

$$p(y|z) = \mathcal{N}(y; 0, \sigma^2(z; \theta_h)) \tag{3.18}$$

Here, h_a and h_s denote the analysis and synthesis transforms in the auxiliary autoencoder. The hyperprior model enhances the entropy model by conditioning on z, leading to more accurate and spatially adaptive entropy estimates.

The *mbt2018_mean* model extends the *bmshj2018_hyperprior* by incorporating a mean prediction in addition to the scale prediction for the Gaussian distribution of the latent variables:

$$p(y|z) = \mathcal{N}(y; \mu(z; \theta_h), \sigma^2(z; \theta_h)) \tag{3.19}$$

This allows the model to capture the mean shift in the latent space, providing a more flexible and accurate entropy model by predicting both mean μ and scale σ from the hyperprior.

The *mbt2018* model combines the hyperprior with an autoregressive context model to further refine the entropy estimation. The context model uses previously decoded latents to improve the prediction of the current latent:

$$p(y_i|y_{<i}, z) = \mathcal{N}(y_i; \mu(y_{<i}, z), \sigma^2(y_{<i}, z)) \tag{3.20}$$

This joint model leverages both the spatial dependencies captured by the hyperprior and the sequential dependencies captured by the autoregressive model, leading to a significant improvement in compression performance.

The *cheng2020_anchor* model introduces discretized Gaussian mixture likelihoods to model the distributions of latent codes more flexibly. This is formulated as

$$p(y|z) = \sum_{k=1}^{K} w^{(k)} \mathcal{N}(y; \mu^{(k)}(z), \sigma^{2(k)}(z)) \tag{3.21}$$

This mixture model is designed to capture complex distributions of latent variables by using multiple Gaussian components, each with its own mean and scale parameters conditioned on the hyperprior. This approach reduces spatial redundancy and improves the accuracy of entropy models, achieving better compression performance.

Finally, the *cheng2020_attn* model incorporates attention mechanisms to improve the performance of the image compression model. The attention modules help the network focus on complex regions of the image, enhancing the rate-distortion performance. The attention mechanism is integrated into the network architecture as

$$y = f(x; \theta, A) \tag{3.22}$$

where A represents the attention module parameters. The attention module modifies the convolutional features to prioritize information-rich regions, leading to better compression results.

It is important to note that each of the model was evaluated at 3 distinct quality levels. These levels represent directly the magnitude of the Lagrangian multiplier and channel size of the latent representation itself, which increases from 192 to 320 for the former and 128 to 192 for latter and are used to weight the compression vs distortion performance of the models. The levels go from 1 to 8 for *bmshj2018_factorized*, *bmshj2018_hyperprior*,

Neural Compression Model	Number of Parameters
bmshj2018_hyperprior_2	4,968,963
bmshj2018_hyperprior_5	4,968,963
bmshj2018_hyperprior_8	11,582,275
bmshj2018_factorized_2	2,887,363
bmshj2018_factorized_5	2,887,363
bmshj2018_factorized_8	6,788,675
mbt2018_mean_2	6,921,123
mbt2018_mean_5	17,327,651
mbt2018_mean_8	17,327,651
mbt2018_2	13,896,419
mbt2018_5	25,270,548
mbt2018_8	25,270,548
cheng2020_anchor_2	11,726,269
cheng2020_anchor_4	26,364,908
cheng2020_anchor_6	26,364,908
cheng2020_attn_2	13,076,413
cheng2020_attn_4	29,397,740
cheng2020_attn_6	29,397,740

TABLE 3.1: Number of parameters in various neural compression models.

mbt2018_mean, *mbt2018* and 1 to 6 for *cheng2020_anchor*, and *cheng2020_attn*. The former were evaluated at 2, 5 and 8 while the latter at 2, 4 and 6. A comparison of the size of of each neural compression model at different quality levels can be seen in Table 3.1. All networks were pretrained for 4-5M steps on 256x256 image patches randomly extracted and cropped from the Vimeo90K dataset [288]. In Section 3.5 the results are presented for each model underscoring the quality level of the pretrained models.

3.4.3 Classification Models Used

In this study, we utilized three different types of classification models: a Multi-Layer Perceptron (MLP), a Convolutional Neural Network (CNN), and a Transformer. Each model was specifically designed to leverage the latent spaces generated by the pre-trained and fine-tuned VAE for classifying satellite images.

3.4.3.1 Multi-Layer Perceptron (MLP)

The MLP model consists of three fully connected layers. The input dimension is determined by the size of the latent space produced by the VAE. The architecture of the MLP can be represented as follows:

$$\begin{aligned}
&\text{Layer 1:} \quad \mathbf{h}_1 = \sigma(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) \\
&\text{Layer 2:} \quad \mathbf{h}_2 = \sigma(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2) \\
&\text{Output Layer:} \quad \hat{\mathbf{y}} = \mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3
\end{aligned}$$

where $\mathbf{z}$ is the input latent representation, $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weights and biases of layer i, σ is the activation function (ReLU), and $\hat{\mathbf{y}}$ is the output class probabilities.

3.4.3.2 Convolutional Neural Network (CNN)

The CNN model is designed to handle the spatial dimensions of the latent representations. In this case, we use a custom ResNet architecture, which is adjusted to accommodate the specific input dimensions and number of output classes. The architecture can be summarized as follows:

$$\begin{aligned}
&\text{Convolution:} \quad \mathbf{h}_1 = \sigma(\text{conv1}(\mathbf{z})) \\
&\text{Residual Blocks:} \quad \mathbf{h}_2 = \text{ResNet blocks}(\mathbf{h}_1) \\
&\text{Pooling:} \quad \mathbf{h}_3 = \text{global average pooling}(\mathbf{h}_2) \\
&\text{Output Layer:} \quad \hat{\mathbf{y}} = \text{fully connected}(\mathbf{h}_3)
\end{aligned} \tag{3.23}$$

where $\mathbf{z}$ is the input latent representation, conv1 is the first convolutional layer tailored to the dimension of the relative latent space, σ is the activation function (ReLU), and the ResNet blocks represent the sequence of residual blocks in the ResNet architecture. The global average pooling and fully connected layer produce the final output class probabilities $\hat{\mathbf{y}}$.

3.4.3.3 Transformer

The Transformer model incorporates positional encoding and a multi-layer transformer encoder. This model is well-suited for capturing long-range dependencies in the latent space. The architecture includes an input projection layer, positional encoding, and a transformer encoder followed by a fully connected layer. The process can be described as follows:

Classification Model	**Number of Parameters**
MLP (128)	16,911,626
MLP (192)	25,300,234
MLP (320)	42,077,450
ResNet50 (128)	23,920,522
ResNet50 (192)	24,121,226
ResNet50 (320)	24,522,634
Transformer (128)	35,697,162
Transformer (192)	44,085,770
Transformer (320)	60,862,986

TABLE 3.2: Number of parameters in various classification models.

$$\text{Input Projection:} \quad \mathbf{h}_0 = \text{proj}(\mathbf{z})$$
$$\text{Positional Encoding:} \quad \mathbf{h}_1 = \mathbf{h}_0 + \text{pos\_enc}(\mathbf{h}_0)$$
$$\text{Transformer Encoding:} \quad \mathbf{h}_2 = \text{transformer\_encoder}(\mathbf{h}_1)$$
$$\text{Global Average Pooling:} \quad \mathbf{h}_3 = \text{mean}(\mathbf{h}_2, \text{dim} = 1)$$
$$\text{Output Layer:} \quad \hat{\mathbf{y}} = \mathbf{W}\mathbf{h}_3 + \mathbf{b}$$

where $\mathbf{z}$ is the input latent representation, proj denotes the input projection, pos_enc is the positional encoding, transformer_encoder represents the transformer encoder layers, mean denotes global average pooling, and $\hat{\mathbf{y}}$ is the output class probabilities.

These models were selected for their distinct architectures and capabilities in handling different aspects of the latent representations generated by the VAE. The MLP is straightforward and efficient for general-purpose classification, the CNN leverages spatial information, and the Transformer is adept at capturing complex dependencies within the data. Pretrained models for image classification, such as those trained on large-scale datasets like ImageNet, were not suitable for our application. The latent representations generated by the VAE are inherently different from raw image data, as they are a compressed and abstract representation of the original images. Consequently, using pretrained models directly on these latent spaces would not effectively capture the specific features encoded by the VAE. Therefore, we opted to design and train custom classification models tailored to the latent space characteristics. The dimensions, in terms of parameters, for each classification model can be seen in Table 3.2.

3.4.4 Dataset

Three different datasets were used to test the proposed architecture: EuroSAT [289], RSI-CB256 [290], and PatternNet [291]. The EuroSAT dataset consists of Sentinel-2 satellite images covering 13 spectral bands, with 27,000 labeled images divided into 10 classes: Annual Crop, Forest, Herbaceous Vegetation, Highway, Industrial, Pasture, Permanent Crop, Residential, River, and Sea/Lake. Each image has a spatial resolution of 10 meters per pixel and is provided in 64x64 pixel patches. EuroSAT is a well-regarded benchmark for satellite image classification tasks due to its diverse range of land cover classes and high-quality imagery.

In addition to the EuroSAT dataset, we utilized the RSI-CB-256 dataset, which consists of 24,000 images spanning 35 land-use classes, each with a resolution of 256x256 pixels and a spatial resolution of 0.3 meters. This dataset covers a diverse range of land-use categories, including woodlands (e.g., forest, sapling, shrubwood) and transportation scenes (e.g., crossroads, highways, marinas, river bridges), making it a valuable resource for evaluating the differentiation of various land-use types. The comprehensive labeling, large size, and varied environmental conditions make it an excellent choice for evaluating the performance of our variational autoencoder-based image compression and classification approach.

We also integrated the PatternNet dataset, another well-regarded dataset, comprising 30,400 images across 38 classes that represent various urban and suburban environments, such as airports, harbors, and stadiums. Each image in PatternNet has a resolution of 256x256 pixels and captures intricate patterns characteristic of human-made structures, collected from Google Earth. The inclusion of PatternNet enables us to assess the model's robustness in identifying and distinguishing complex, highly structured man-made environments.

To ensure compatibility with our model, the original image patches from the EuroSAT dataset were preprocessed and resized to 256x256 pixels using bilinear interpolation, preserving the integrity of the images while adapting them to the required dimensions. For all datasets, we restricted our usage to the RGB bands (Red, Green, Blue) to ensure compatibility with standard image processing pipelines. Although we explored normalization techniques to standardize the datasets, normalization proved ineffective in enhancing model performance and was therefore not utilized. The images were used in their raw form without any normalization, ensuring that the intrinsic characteristics of the data were maintained.

We also explored normalization techniques to standardize the dataset during the preprocessing phase. However, normalization proved to be ineffective in enhancing model performance, likely due to the reduction in variance of the dataset, and was therefore not utilized. The images were used in their raw form without any normalization, ensuring that the intrinsic characteristics of the data were maintained.

To maintain conciseness, the results presented in the main section of the paper are based exclusively on the EuroSAT dataset. Corresponding results for the other two datasets are provided in Appendices A and B.

3.5 Results

3.5.1 Baseline

To establish a baseline comparison between the proposed architecture and standard compression models, we tested JPEG compression at various quality levels. We compared these models not only in terms of compression capabilities but also in terms of classification accuracy using pretrained classifiers similar to the custom ones built for our experiments. We applied JPEG compression to the Sentinel-2 images at various quality levels (e.g., 10%, 30%, 50%, 70%, and 90%) to assess the impact of compression on image quality and size, especially in terms of loss of information and accuracy at lower levels. As shown in Figure 3.5 each method's performance demonstrates how image quality, as measured by PSNR, improves with an increase in BPP. The JPEG method (pink) shows a consistent but less efficient performance compared to neural methods, especially at higher BPP values, indicating poorer compression efficiency, also shown in Figure 3.7. In contrast, methods like *bmshj2018_hyperprior*r (blue) and *mbt2018_mean* (green) exhibit significantly higher PSNR values at equivalent BPPs, showcasing their superior ability to retain image quality at lower bit rates as expected.

As JPEG reconstruction quality levels increases, the F1 scores for all classifiers generally improve, indicating better classification performance with higher image quality, as shown in Figure 3.6 as well as better reconstruction quality as shown in Figure 3.7. The ResNet classifier shows the most significant improvement, achieving an F1 score above 0.95 at higher JPEG quality levels. The vision transformer (ViT) classifier also demonstrates strong performance, with F1 scores approaching 0.9. The MLP classifier, while improving with higher quality levels, shows a more modest increase in F1 scores compared to the other two classifiers, probably due to the simplicity of its architecture. The average BPP, depicted by the red line in Figure 3.6, increases significantly with

Classifier	Quality	Average BPP ↓	PSNR ↑	F1 ↑	RDAI ↑
MLP	10	0.2467	32.14	0.3967	0.6353
MLP	40	0.3766	39.86	0.4450	0.6899
MLP	70	0.5933	43.39	0.4586	0.7068
MLP	100	2.8203	42.85	0.4606	0.6303
ResNet	10	0.2467	32.14	0.4398	0.6496
ResNet	40	0.3766	39.86	0.8854	0.8365
ResNet	70	0.5933	42.85	0.9361	**0.8628**
ResNet	100	2.8203	46.56	**0.9520**	0.8145
ViT	10	0.2467	32.14	0.4852	0.6647
ViT	40	0.3766	39.86	0.8345	0.8196
ViT	70	0.5933	42.85	0.8931	0.8485
ViT	100	2.8203	46.56	0.9169	0.8028

TABLE 3.3: Evaluating pre trained classifiers performance on multiple JPEG compression quality levels on EuroSAT dataset; multi layer perceptron (MLP), convolutional neural network with residual connections (ResNet), and vision transformer (ViT).

higher JPEG quality levels, reflecting the trade-off between compression efficiency and image quality. The equally weighted RDAI metric peaks at quality level 70 for both ResNet and ViT, indicating an optimal trade-off between compression, reconstruction quality, and classification accuracy at this level as shown in Table 3.3.

3.5.2 Frozen Weights

To assess the performance of the neural compression models in creating meaningful latent spaces that are utilizable from other machine learning models, we used quantitative and qualitative metrics. Quantitatively, metrics such as F1 score, PSNR, BPP, and RDAI were employed to measure the classification performance, reconstruction quality, compression efficiency, and the trade-offs between these factors, respectively. These metrics provide a comprehensive evaluation of how well the neural compression models balance the competing requirements of high compression rates and high classification accuracy.

Qualitatively, t-SNE visualizations were used to assess the structure of the latent spaces generated by the neural compression models. The t-SNE plots, as shown in Figure 3.8, illustrate how different models organize the latent representations of the EuroSAT test set. These visualizations highlight the clustering and separability of the latent features, which are crucial for downstream classification tasks.

We performed a first round of experiments using the logic explained previously in Algorithm 1. The pretrained neural compression models' parameters were frozen and

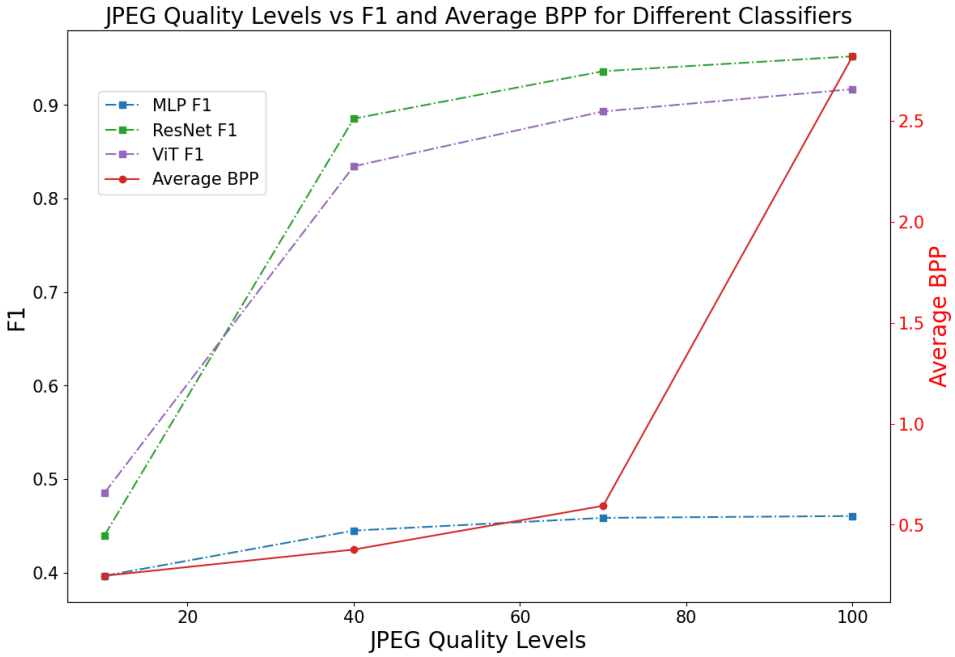


FIGURE 3.6: Reconstruction F1 and BPP vs JPEG quality levels.

then used to compress images from the EuroSAT train dataset. The latent spaces created were used to train three separate classifiers (MLP, CNN, Transformer), results can be seen in Table 3.4.

When comparing the performance with JPEG compression at quality level 100, which serves as a high-quality baseline, it is evident that neural compression models significantly enhance classification performance while maintaining lower BPP values. For instance, the MLP classifier using *bmshj2018_hyperprior_8* achieves an F1 score of 0.6856 and a PSNR of 49.14 with a BPP of 0.3459, substantially better than the JPEG's F1 score of 0.4606 and PSNR of 42.85, and at a fraction of the BPP for a similar classifier. Similarly, the ResNet50 classifier with *bmshj2018_hyperprior_8* shows an F1 score of 0.5011 and a PSNR of 49.14 at a BPP of 0.3389, surpassing the JPEG's performance. The Transformer classifier exhibits the highest performance gains, with the *bmshj2018_hyperprior_8* configuration achieving an F1 score of 0.794, a PSNR of 49.14, and an RDAI of 0.8608, all at a BPP of 0.2816. This trend is consistent across other neural compression models; for instance, *mbt2018_mean_8* and *mbt2018_8* configurations show significant improvements in F1 scores and PSNR values while maintaining lower BPP compared to JPEG. These results highlight the superiority of neural compression techniques in balancing compression efficiency and classification accuracy.

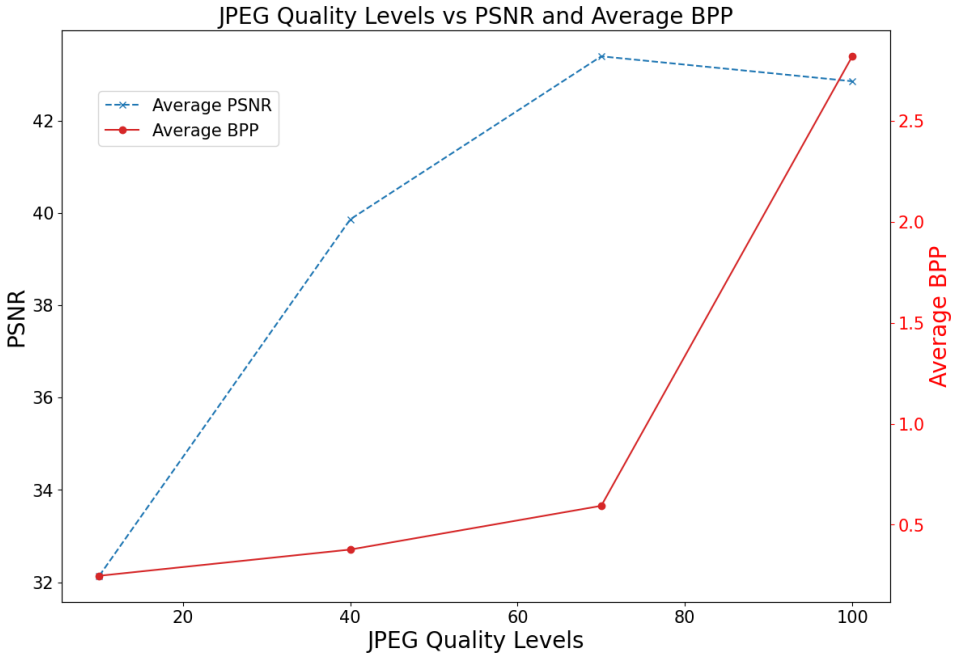
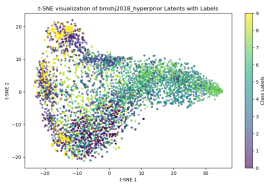


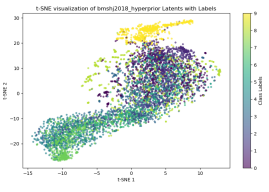
FIGURE 3.7: Reconstruction PSNR and BPP vs JPEG quality levels.

From these results we can also extrapolate that by interpreting directly the latent spaces some of the information that would be lost during decompression is preserved. Additionally, the RDAI metric, which provides a comprehensive measure of the trade-offs between rate, distortion, and accuracy, consistently shows higher values for neural compression models across all classifiers, further validating their effectiveness over traditional JPEG compression. In particular, the Transformer classifier, when paired with *bmshj2018_hyperprior_8*, achieves an RDAI of 0.8608, indicating an optimal balance and best performance for both neural compression and JPEG values.

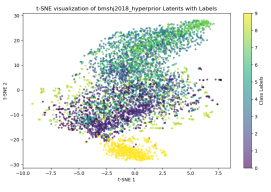
The t-SNE plots in Fig. 8 illustrate how different models, such as *cheng2020_anchor*, *mbt2018_mean*, *bmshj2018_hyperprior*, *bmshj2018_factorized*, and *cheng2020_attn*, organize the latent representations of the EuroSAT test set. For example, the *bmshj2018_hyperprior* models show better-defined clusters at higher bitrates, indicating that the latent space captures meaningful structures



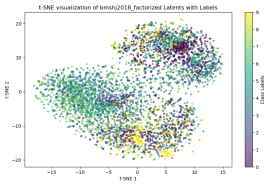
(A) bmshj2018_hyperprior 2



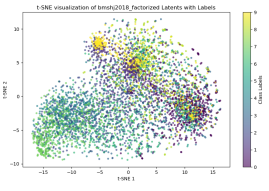
(B) bmshj2018_hyperprior 5



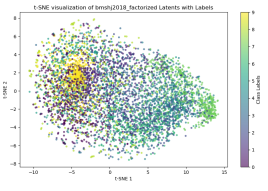
(C) bmshj2018_hyperprior 8



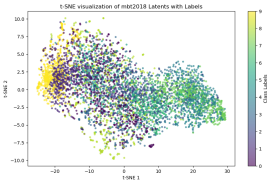
(D) bmshj2018_factorized 2



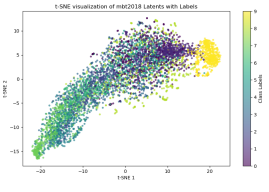
(E) bmshj2018_factorized 5



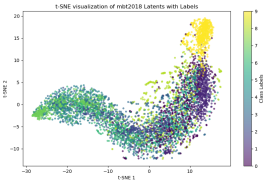
(F) bmshj2018_factorized 8



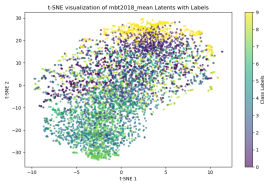
(G) mbt2018 2



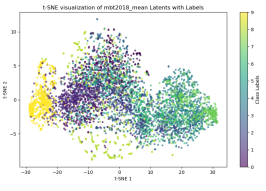
(H) mbt2018 5



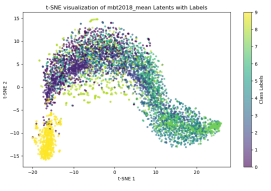
(I) mbt2018 8



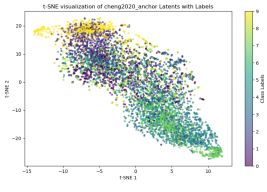
(J) mbt2018_mean 2



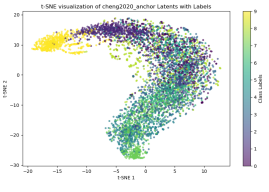
(K) mbt2018_mean 5



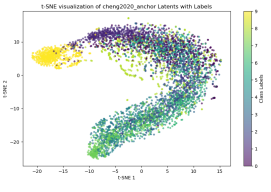
(L) mbt2018_mean 8



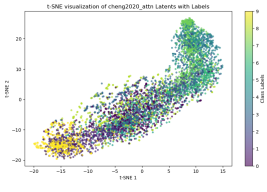
(M) cheng2020_anchor 2



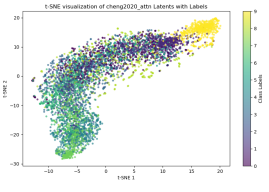
(N) cheng2020_anchor 4



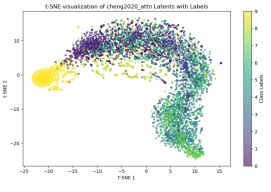
(O) cheng2020_anchor 6



(P) cheng2020_attn 2



(Q) cheng2020_attn 4



(R) cheng2020_attn 6

FIGURE 3.8: t-SNE visualizations of models constructed latents of EuroSAT test set, with labels.

Method	**MLP**					**ResNet50**					**Transformer**				
	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑
bmshj2018_hyperprior 2	0.6619	0.6561	0.0541	37.04	0.755	0.5067	0.4744	0.07	37.04	0.694	0.7481	0.7477	0.0584	37.04	0.786
bmshj2018_hyperprior 5	0.6478	0.6247	0.1251	42.75	0.774	0.3785	0.2942	0.1288	42.75	0.664	0.7907	0.7898	0.127	42.75	0.829
bmshj2018_hyperprior 8	0.6931	0.6856	0.3459	49.14	0.8225	0.5144	0.5011	0.3389	49.14	0.761	**0.7939**	**0.794**	0.2816	49.14	**0.861**
bmshj2018_factorized 2	0.3228	0.2613	0.1047	37.23	0.623	0.313	0.2422	0.1012	37.23	0.616	0.6711	0.6721	0.1061	37.23	0.760
bmshj2018_factorized 5	0.7135	0.7068	0.2675	42.87	0.797	0.2941	0.2172	0.2801	42.87	0.634	0.7581	0.7574	0.2637	42.87	0.814
bmshj2018_factorized 8	0.6587	0.6451	0.6893	49.62	0.800	0.3789	0.274	0.6893	49.62	0.677	0.7369	0.7309	0.6566	**49.62**	0.830
mbt2018_mean 2	0.6019	0.5881	0.0474	37.38	0.735	0.3937	0.3156	0.0384	37.38	0.644	0.6983	0.6992	0.0332	37.38	0.772
mbt2018_mean 5	0.5372	0.5167	0.1077	42.58	0.734	0.4528	0.3775	0.0959	42.58	0.691	0.7806	0.7784	0.0948	42.58	0.826
mbt2018_mean 8	0.7476	0.7479	0.279	48.68	0.843	0.3536	0.3119	0.3053	48.68	0.697	0.7928	0.7919	0.252	48.68	0.858
mbt2018 2	0.5317	0.5243	0.0491	37.91	0.716	0.3369	0.2929	0.0371	37.91	0.640	0.6128	0.6002	0.0448	37.91	0.741
mbt2018 5	0.722	0.7161	0.0834	43.01	0.807	0.1957	0.1168	0.1015	43.01	0.607	0.7715	0.7694	0.1065	43.01	0.824
mbt2018 8	0.75	0.7471	0.281	49.15	0.845	0.3441	0.3105	0.2935	49.15	0.699	0.7848	0.7864	0.2709	49.15	0.859
cheng2020_anchor 2	0.5709	0.5655	0.0388	37.53	0.728	0.2785	0.1584	0.0495	37.53	0.592	0.6793	0.6783	0.0276	37.53	0.766
cheng2020_anchor 4	0.63	0.6288	0.079	41.13	0.768	0.332	0.271	0.0555	41.13	0.650	0.7393	0.737	0.0817	41.13	0.804
cheng2020_anchor 6	0.7374	0.7375	0.1248	44.83	0.823	0.4539	0.391	0.1598	44.83	0.707	0.7635	0.7635	0.0876	44.83	0.833
cheng2020_attn 2	0.5863	0.5696	**0.032**	37.87	0.732	0.2974	0.2209	0.042	37.87	0.615	0.6969	0.6968	0.0495	37.87	0.773
cheng2020_attn 4	0.6798	0.6748	0.0745	41.43	0.786	0.4489	0.3939	0.0781	41.43	0.6915	0.7444	0.7423	0.0624	41.43	0.808
cheng2020_attn 6	0.7119	0.7127	0.1296	44.83	0.815	0.3891	0.3537	0.1369	44.83	0.695	0.7563	0.7562	0.0994	44.83	0.83

TABLE 3.4: Performance comparison of MLP, CNN, and Transformer models, using Algorithm 1 and frozen weights for neural compression models. Accuracy and F1 represent classification results of models using neural compressed latent representations.

even at lower compression rates. The *bmshj2018_hyperprior* and *cheng2020_attn* models, particularly at higher bitrates, demonstrate clear separations between clusters, suggesting that these models retain significant feature information necessary for high classification. It also shows that there is a direct correlation between better separation in the latent space and higher classification accuracy, which makes sense. The visualizations also reveal that models like *bmshj2018_factorized_8*, while achieving high PSNR and F1 scores, might have more overlapping clusters, which could impact classification performance in more challenging scenarios.

3.5.3 Fine Tuning

The second set of experiments was carried out following the logic outlined in Algorithm 2 for fine-tuning the VAE along with the training of the classifiers. To test this, we selected the best-performing classifier from the previous experiments (the Transformer architecture) and trained it jointly using the composite loss shown in Equation 3.4. This composite loss function was designed to optimize both the reconstruction quality and classification performance simultaneously, allowing the VAE to adapt its latent space to

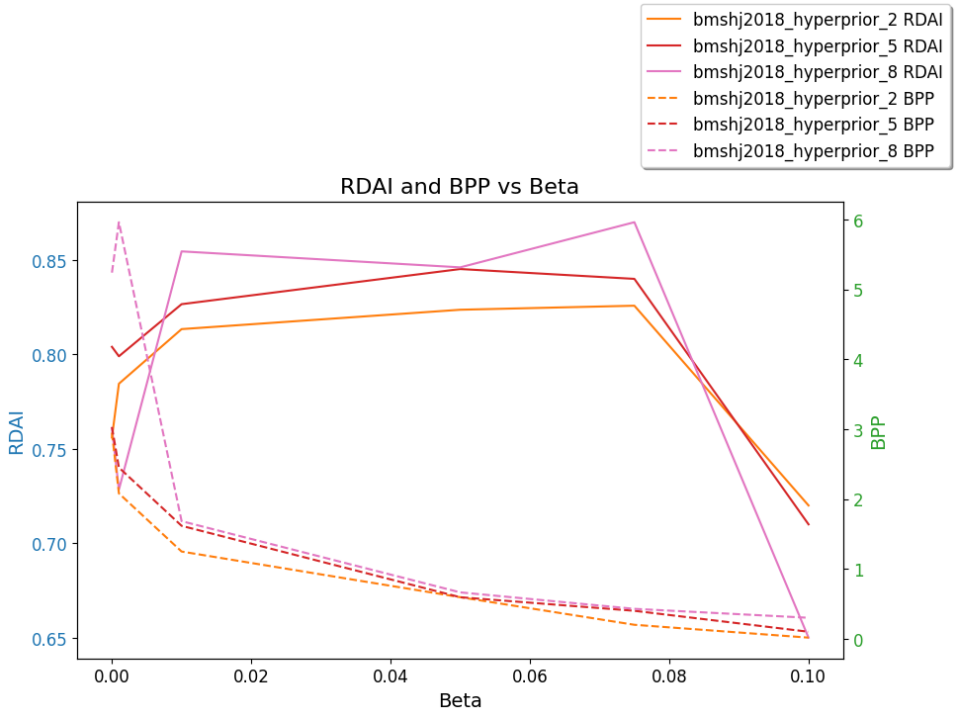
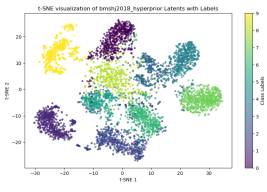


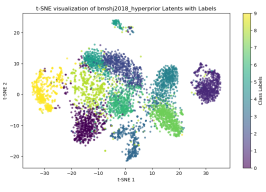
FIGURE 3.9: Effects of tuning the λ_{bpp} multiplier on RDAI (left axis) and BPP (right axis) values.

better suit the classification task. To temper and balance the various elements of the composite loss function, the Lagrangian multipliers were fine-tuned using a grid search on various models, which resulted in the best values to be $\lambda_{rec} = 10$, $\lambda_{class} = 1$, $\lambda_{bpp} = 0.075$. The grid search spanned values between $[0.01, 1, 10]$ for all multipliers. It was then narrowed down using halfway points $[0.05, 0.075]$ to further improve performance. Choosing the right multiplier parameters turned to be crucial to improve the overall performance of the models. An example of the effects of choosing the λ_{bpp} parameter can be seen in Figure 3.9. Choosing a λ_{bpp} that is too low results in high bitrate as not enough importance is assigned to compression during the multi objective optimization process. While if λ_{bpp} is too large, then the data is compressed excessively and information is lost in the process lowering the reconstruction and classification performance, ultimately hurting RDAI scores.

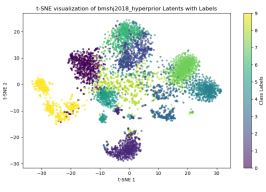
When comparing the performance of the fine tuned combinations of neural compression models and transformer classifier we see a substantial increase in performance. The fine-tuned Transformer model consistently outperformed the frozen weight counterpart and the JPEG baseline. We can see in Table 3.5 that the fine tuned *cheng2020_attn_6*



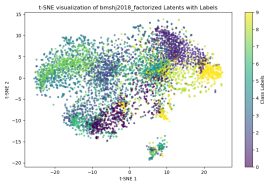
(A) bmshj2018_hyperprior 2



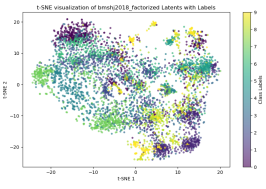
(B) bmshj2018_hyperprior 5



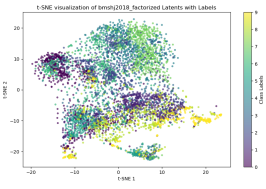
(C) bmshj2018_hyperprior 8



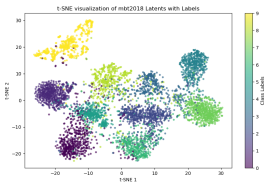
(D) bmshj2018_factorized 2



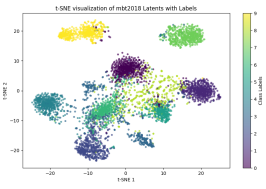
(E) bmshj2018_factorized 4



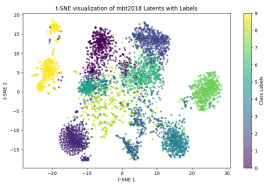
(F) bmshj2018_factorized 8



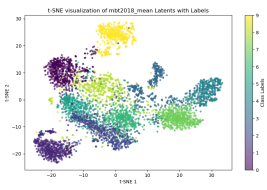
(G) mbt2018 2



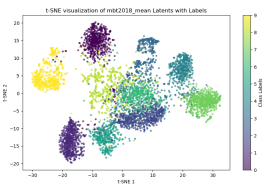
(H) mbt2018 5



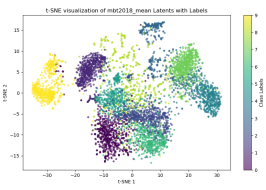
(I) mbt2018 8



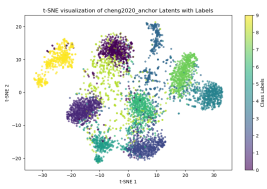
(J) mbt2018_mean 2



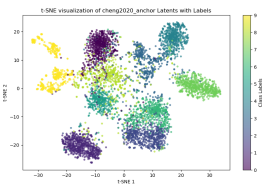
(K) mbt2018_mean 5



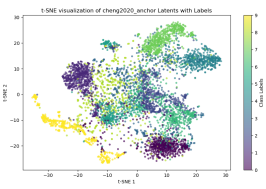
(L) mbt2018_mean 8



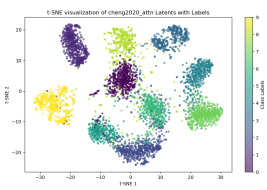
(M) cheng2020_anchor 2



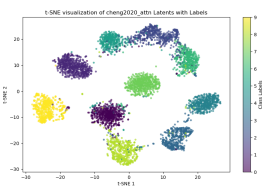
(N) cheng2020_anchor 4



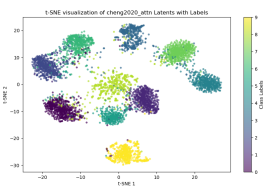
(O) cheng2020_anchor 6



(P) cheng2020_attn 2



(Q) cheng2020_attn 4



(R) cheng2020_attn 6

FIGURE 3.10: t-SNE visualizations of fine tuned models constructed latents of EuroSAT test set, with labels.

Method	Accuracy	F1	BPP	PSNR	RDAI
bmshj2018_hyperprior 2	0.9365	0.9363	0.1996	33.81	0.826
bmshj2018_hyperprior 5	0.9456	0.9456	0.3998	37.15	0.841
bmshj2018_hyperprior 8	0.937	0.9371	0.4278	43.1	0.87
bmshj2018_factorized 2	0.8533	0.8525	0.3257	41.04	0.834
bmshj2018_factorized 5	0.8609	0.8605	0.7613	42.02	0.827
bmshj2018_factorized 8	0.8924	0.8916	1.262	**46.97**	0.849
mbt2018_mean 2	0.9209	0.9209	0.2632	41.17	0.859
mbt2018_mean 5	0.937	0.937	0.254	42.88	0.875
mbt2018_mean 8	0.9387	0.9387	0.4037	45.4	0.884
mbt2018 2	0.9044	0.9044	0.2171	40.75	0.853
mbt2018 5	0.9269	0.9265	0.2351	42.69	0.871
mbt2018 8	0.9304	0.9302	0.4179	46.37	0.886
cheng2020_anchor 2	0.9424	0.9427	0.2181	41.16	0.868
cheng2020_anchor 4	0.9006	0.9006	0.1776	42.29	0.862
cheng2020_anchor 6	0.892	0.8922	0.5173	44.06	0.857
cheng2020_attn 2	0.9439	0.9439	**0.163**	41.65	0.873
cheng2020_attn 4	0.9491	0.949	0.1845	41.67	0.874
cheng2020_attn 6	**0.9487**	**0.9487**	0.2615	44.94	**0.890**

TABLE 3.5: Fine tuning performance using various methods and transformer classifier.

achieves and RDAI of 0.890, the highest of all experiments and a substantial increase from the performance using pretrained frozen weights with an RDAI of 0.83. When comparing this result with the JPEG baseline we can see better reconstruction performance 44.94 at a bit rate of 0.2615 where JPEG is only able to achieve 32.14 and much better classification performance with an F1 score of 94.87 vs JPEG ViT F1 score of 66.47. Therefore the lossy neural compression model is able to achieve better performance at lower bitrates in both clasification and reconstruction. If we compare the performance of the various neural compression models qualitatively on top of quantitatively we can see from Figure 3.10 that the best performing neural compression models learn to induce separation in the latent spaces between classes, effectively clustering similar images together. This proves particularly beneficial for classification with the *cheng2020_attn* models performing best and showing the greatest separation between clusters, while the *bmshj2018_factorized* performing the worse and showing a lot more overlapping between clusters and poor separation.

3.5.4 Ablation Study

During parameter selection for the Lagrange multipliers through grid search, various scenarios were explored, including the absence of some components from the composite objective function to study the effects. The aim was to understand how each component

influenced the overall performance of the model. For instance, by omitting certain terms, we could observe changes in the model's ability to balance different objectives, such as compression efficiency and reconstruction quality.

Since the neural compression models were pretrained, there was no point in experimenting with λ_{class} being zero. This would have simply meant trying to fine-tune the neural compression model to the dataset, which would have reduced generalizability and generally shown to reduce performance. Instead, we focused on testing three specific scenarios that allowed us to investigate the trade-offs between different components of the objective function. These scenarios are illustrated in Figure 3.11 and are detailed below:

- **No Constraint on Bit-Rate**:

$$\lambda_{\text{class}} = 1, \lambda_{\text{bpp}} = 0, \lambda_{\text{rec}} = 10$$

- **No Constraint on Bit-Rate and Reconstruction**:

$$\lambda_{\text{class}} = 1, \lambda_{\text{bpp}} = 0, \lambda_{\text{rec}} = 0$$

- **No Constraint on Reconstruction**:

$$\lambda_{\text{class}} = 1, \lambda_{\text{bpp}} = 1, \lambda_{\text{rec}} = 0$$

As seen in Figure 3.11, both scenarios that cancel out the "bpp" term and do not optimize for bitrate result in poor compression performance, with bitrates all higher than 3 BPP. The "No Constraint on Bit-Rate and Reconstruction" scenario performs the worst, achieving low reconstruction and compression performance. Additionally, it does not significantly increase classification performance, remaining comparable to the better-performing fine-tuned models shown earlier. The "No Constraint on Reconstruction" scenario results in lower bitrates and similar classification performance, but very poor reconstruction performance as expected. Therefore, while it could be a viable solution if the only goal is known classification, it might not be possible to reconstruct the original image.

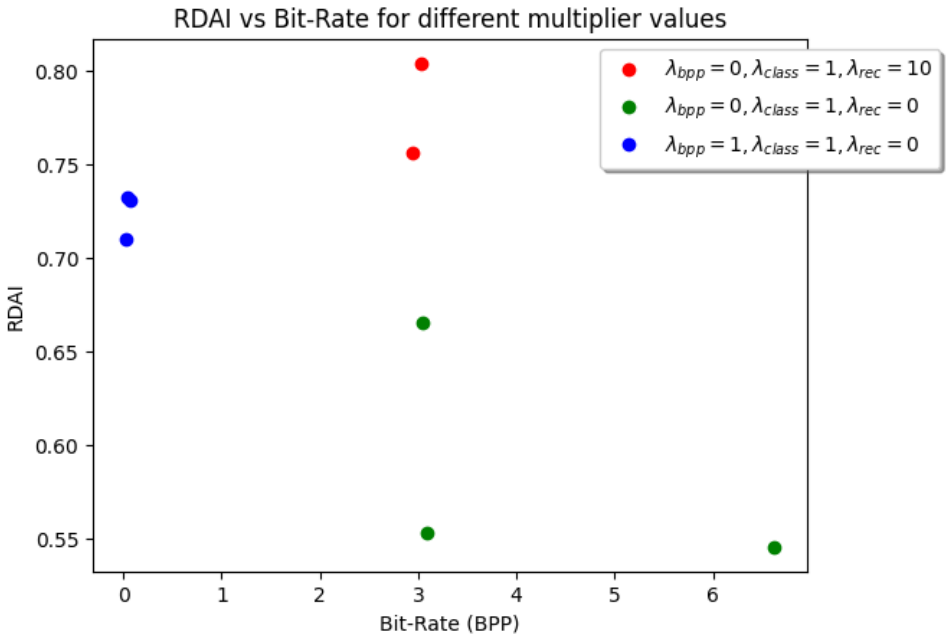


FIGURE 3.11: Fine tuning performance in terms of RDAI vs BPP for *bmshj2018_hyperprior [2,5,8]* with no bit rate loss constraint (red), with no reconstruction loss constraint (blue), with no reconstruction and bit rate loss constraint (green).

3.6 Discussion

The findings from this study demonstrate the potential of using neural compression models, specifically Variational Autoencoders (VAEs), to enhance satellite image analysis by leveraging latent representations directly for classification tasks. This approach offers a significant improvement over traditional methods, in terms of the comperssion, reconstruction and accuracy tradeoff. The results highlight several key aspects. First, in terms of compression efficiency and reconstruction quality, neural compression models, such as bmshj2018 hyperprior and cheng2020 attn, achieve higher compression ratios while maintaining excellent reconstruction quality. The PSNR values for these models are significantly higher than those obtained with JPEG compression, even at lower bit rates, indicating better preservation of image quality. This improvement is crucial for applications where both storage and transmission efficiency are paramount, such as remote sensing and environmental monitoring. The higher compression efficiency means that more data can be stored and transmitted without significant loss of quality, making it feasible to handle the large volumes of data generated by satellite imagery.

Moreover, the latent representations generated by these neural compression models

are highly effective for classification tasks, in future studies the utilization of such latent spaces for other tasks such as segmentation and object detection will be explored. The Transformer classifier, when fine-tuned with these latent spaces, achieves remarkable F1 scores, demonstrating that the latent spaces retain essential features for accurate classification as well as the viability of using the proposed architecture. This performance indicates that neural compression models do not just compress the data but also preserve critical information necessary for downstream tasks. By using the latent spaces directly, the approach eliminates the need for decompression, reducing computational overhead and loss of information.

The visualizations of latent spaces using t-SNE further support these findings, showing well-defined clusters corresponding to different classes, particularly for models like cheng2020 attn. This clear separability in the latent space is crucial for high classification performance, as it indicates that the model has effectively learned to distinguish between different types of images. The t-SNE plots illustrate that the best-performing models induce a structured and meaningful organization in the latent space, which is directly correlated with their classification performance. This proves that neural compression models can successfully be fine tuned for specific tasks without sacrificing reconstruction quality and compression ratios.

3.6.1 Limitations and Future Work

While our study demonstrates the potential of variational autoencoder (VAE) models with Gaussian and discretized Gaussian mixture likelihoods for neural compression and classification tasks, it does not include more recent neural compression models due to several limitations that must be acknowledged. A key constraint is our reliance on pretrained models available through the CompressAI library. These models were trained using specific and undisclosed configurations, techniques, and hyperparameters, such as the application of an exponential moving average for weight updates and the implementation of a gradually decaying learning rate schedule, both of which are commonly used in VAE training to mitigate instability. The lack of transparency and control over the pretraining process introduces challenges in maintaining consistency across models. Attempting to incorporate more recent methods, such as hierarchical VAEs or diffusion models, without consistent pretraining conditions could lead to biased comparisons and potentially compromise the validity of our results.

Moreover, integrating these other advanced models would have required a significant shift in the focus of the study. Hierarchical VAEs [279], diffusion models [280], and other

hybrid architectures [281] represent distinct classes of neural compression techniques that would necessitate different experimental setups, comparisons, and analyses. This paper is specifically focused on exploring the capabilities of Gaussian-based VAE models for classification tasks. Including these more recent models would extend the scope of the paper beyond the primary objective and would require a different research approach to thoroughly evaluate and compare the diverse set of models. Therefore, while these advanced models offer promising avenues for future research, they were intentionally excluded from this work to maintain clarity and focus in addressing our specific research goals. Future work should focus on experimenting with new neural compression architectures, such as hierarchical VAEs and diffusion models, as well as expanding the scope to other tasks beyond classification, including segmentation and object detection.

3.6.2 Security considerations

In addition to the benefits of compression efficiency and classification performance, the use of neural compression models for satellite image analysis also offers notable security advantages. The process of transforming satellite images into latent representations inherently applies a form of data masking. This transformation makes it significantly more difficult for unauthorized parties to reconstruct the original images without access to the specific neural compression model used for encoding and decoding. This added layer of security is particularly important for sensitive applications, such as military or confidential environmental monitoring, where the protection of raw data is paramount. By transmitting only the latent representations instead of the raw images, we reduce the risk of data interception and misuse during transmission. Furthermore, the neural compression models can be designed to incorporate additional security measures, such as encryption of the latent space, further enhancing data protection. This approach can be integrated seamlessly into existing data processing pipelines, ensuring that security does not come at the expense of efficiency or accuracy. The robustness of these neural compression models against potential attacks aimed at extracting sensitive information from the latent space is an area worthy of further research. Investigating the resilience of different neural architectures and training methodologies to adversarial attacks will be crucial in ensuring the security and reliability of these systems in practical applications. Overall, leveraging neural compression models for satellite image analysis not only optimizes data handling but also provides an enhanced security framework, addressing one of the critical concerns in the transmission and storage of satellite imagery.

3.7 Conclusion

This study demonstrates the effectiveness of neural compression models for satellite image analysis within the proposed architecture. By utilizing the latent spaces generated by these models, we achieve significant improvements in both compression efficiency and classification accuracy, with additional benefits for data transmission and analysis. The compact latent representations not only enable efficient storage and transmission but also enhance the speed and accuracy of downstream tasks, while providing a degree of data security by obfuscating the original images.

Future research could build on this work by integrating neural compression models with other machine learning frameworks, potentially expanding their applicability and improving performance across a broader range of scenarios. Further advancements in the architecture could optimize compression and classification tasks beyond current limitations. Additionally, applying these latent spaces to tasks such as segmentation, object detection, and anomaly detection would help generalize and validate this approach across diverse remote sensing applications, offering new insights and capabilities for satellite image analysis.

Furthermore, expanding the use of these compressed representations in multi-modal learning scenarios, combining satellite imagery with other data sources like sensor data or textual information, could significantly enhance the ability to extract meaningful patterns across complex datasets. Embracing these advancements could revolutionize the way we process and analyze satellite imagery in the future.

3.8 PatternNet Dataset Results

Method	**MLP**					**ResNet50**					**Transformer**				
	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑
bmshj2018_hyperprior_2	0.7173	0.7135	0.2128	28.7236	0.7229	0.2530	0.1960	0.2414	28.6717	0.5494	0.7742	0.7717	0.2065	28.7275	0.7425
bmshj2018_hyperprior_5	0.7939	0.7875	0.7109	33.7323	0.7588	0.6237	0.6251	0.6703	33.6864	0.7058	0.8372	0.8367	0.7991	33.7265	0.7722
bmshj2018_hyperprior_8	0.8413	0.8392	1.6141	39.9900	0.7806	0.4266	0.3955	1.5120	39.9649	0.6362	**0.8604**	**0.8598**	2.0504	39.9937	0.7730
bmshj2018_factorized_2	0.3794	0.3665	0.2023	27.5458	0.6012	0.4987	0.4705	0.1962	27.4860	0.6357	0.6809	0.6795	0.2080	27.5481	0.7052
bmshj2018_factorized_5	0.4725	0.4601	0.6981	32.1816	0.6416	0.5025	0.5007	0.6327	32.1118	0.6569	0.7839	0.7834	0.5954	32.1878	0.7527
bmshj2018_factorized_8	0.4816	0.4465	2.0057	39.1262	0.6320	0.2997	0.2822	1.7339	39.0857	0.5862	0.7785	0.7838	2.3702	39.1238	0.7322
mbt2018_mean_2	0.7036	0.7042	0.2093	28.8692	0.7208	0.0735	0.0476	0.1918	28.8160	0.5024	0.7595	0.7564	0.2165	28.8752	0.7379
mbt2018_mean_5	0.8146	0.8137	0.7283	34.3152	0.7702	0.5408	0.5522	0.7717	34.2862	0.6815	0.8362	0.8372	0.7690	34.2781	0.7764
mbt2018_mean_8	0.8512	0.8509	1.3572	40.1849	0.7942	0.4651	0.4573	1.4883	40.1599	0.6586	0.8497	0.8485	1.7180	40.1566	0.7812
mbt2018_2	0.5972	0.5977	0.2854	29.0184	0.6836	0.0839	0.0682	0.2310	28.9651	0.5088	0.6806	0.6687	**0.1735**	29.0129	0.6940
mbt2018_5	0.8345	0.8360	0.6473	34.5062	0.7813	0.5276	0.5157	0.7395	34.4741	0.6714	0.8653	0.8649	0.5551	34.5023	0.7837
mbt2018_8	0.7895	0.7899	1.6005	40.2685	0.7662	0.6416	0.6356	1.6126	40.2361	0.7143	0.8561	0.8551	1.6308	**40.2697**	**0.8127**
cheng2020_anchor_2	0.7937	0.7906	0.2165	29.0947	0.7505	0.2240	0.1910	0.1805	29.0311	0.5517	0.8237	0.8231	0.2301	29.0942	0.7609
cheng2020_anchor_4	0.8173	0.8190	0.3848	32.4883	0.7732	0.2577	0.2311	0.4854	32.4447	0.5739	0.8464	0.8461	0.3634	32.4882	0.7830
cheng2020_anchor_6	0.8225	0.8196	0.8840	35.8766	0.7756	0.6798	0.6917	0.8331	35.8375	0.7345	0.8317	0.8309	1.0572	35.7953	0.7731
cheng2020_attn_2	0.7164	0.7134	0.2125	29.1918	0.7255	0.1840	0.1406	0.2036	29.1313	0.5347	0.8206	0.8186	0.1796	29.1379	0.7613
cheng2020_attn_4	0.8141	0.8129	0.4378	32.5416	0.7697	0.3620	0.3300	0.4811	32.4898	0.6072	0.8479	0.8477	0.4563	32.4909	0.7804
cheng2020_attn_6	0.8258	0.8263	1.0323	35.8179	0.7726	0.6646	0.6618	0.7824	35.7788	0.7259	0.8352	0.8342	0.8692	35.7796	0.7804

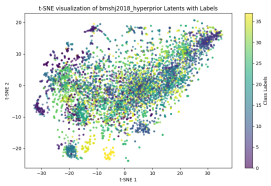
TABLE 3.6: Performance comparison using frozen weights for neural compression models applied to the PatternNet dataset.

JPEG Quality	**Classifier**	**Average BPP ↓**	**PSNR ↑**	**F1 ↑**	**RDAI ↑**
JPEG_10	MLP	**0.4349**	27.91	0.38	0.5999
JPEG_40	MLP	0.8985	33.53	0.5144	0.7861
JPEG_70	MLP	1.3458	36.32	0.5279	0.7244
JPEG_100	MLP	6.3135	**48.4**	0.5036	0.5591
JPEG_10	ResNet	**0.4349**	27.91	0.8918	0.7704
JPEG_40	ResNet	0.8985	33.53	0.9458	**0.8041**
JPEG_70	ResNet	1.3458	36.32	0.9324	0.8003
JPEG_100	ResNet	6.3135	**48.4**	**0.9743**	0.7158
JPEG_10	ViT	**0.4349**	27.91	0.7046	0.7081
JPEG_40	ViT	0.8985	33.53	0.8902	0.7856
JPEG_70	ViT	1.3458	36.32	0.9177	0.7954
JPEG_100	ViT	6.3135	**48.4**	0.9084	0.6939

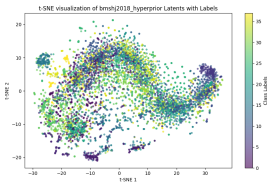
TABLE 3.7: Performance of pre-trained classifiers on different JPEG compression levels applied to the PatternNet dataset.

Method	**Accuracy**	**F1**	**BPP**	**PSNR**	**RDAI**
bmshj2018_hyperprior 2	0.9064	0.9073	0.5093	28.5868	0.777
bmshj2018_hyperprior 5	0.9439	0.9441	0.6741	30.3419	0.793
bmshj2018_hyperprior 8	0.9475	0.9471	0.6571	30.1022	0.794
bmshj2018_factorized 2	0.8653	0.8642	0.3673	27.1	0.759
bmshj2018_factorized 5	0.9094	0.9092	0.7469	30.874	0.782
bmshj2018_factorized 8	0.8715	0.8739	1.03	**32.5343**	0.770
mbt2018_mean 2	0.9211	0.9203	**0.4002**	29.2967	0.789
mbt2018_mean 5	0.9365	0.9357	0.5167	30.3864	0.796
mbt2018_mean 8	0.9474	0.947	0.5482	30.8431	0.801
mbt2018 2	0.7959	0.7919	0.4152	29.3315	0.746
mbt2018 5	0.9117	0.9106	0.5086	30.4802	0.788
mbt2018 8	0.9434	0.9433	0.5561	30.9452	0.800
cheng2020_anchor 2	0.9401	0.9397	0.4331	29.2683	0.794
cheng2020_anchor 4	0.952	0.952	0.3504	30.243	0.806
cheng2020_anchor 6	0.9319	0.9317	0.6106	32.4809	0.803
cheng2020_attn 2	0.9424	0.9425	0.4912	29.3601	0.793
cheng2020_attn 4	0.9548	0.9544	0.4449	29.6791	0.801
cheng2020_attn 6	**0.9638**	**0.9638**	0.476	29.8219	**0.804**

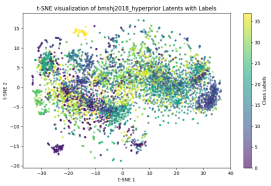
TABLE 3.8: Fine-tuning performance using various methods and Transformer classifier applied to the PatternNet dataset.



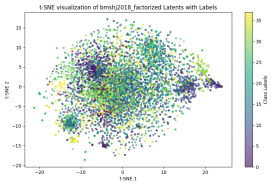
(A) bmshj2018_hyperprior 2



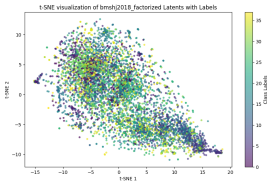
(B) bmshj2018_hyperprior 5



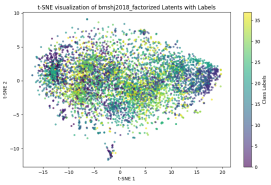
(C) bmshj2018_hyperprior 8



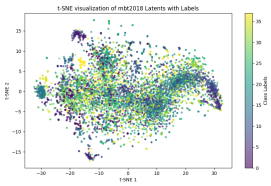
(D) bmshj2018_factorized 2



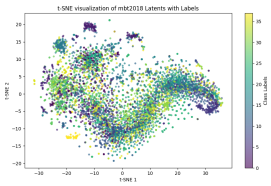
(E) bmshj2018_factorized 4



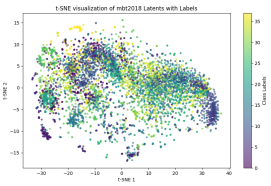
(F) bmshj2018_factorized 8



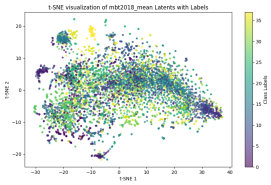
(G) mbt2018 2



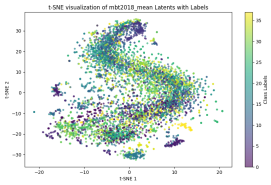
(H) mbt2018 5



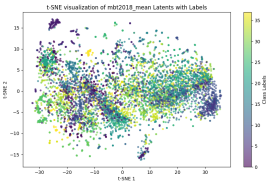
(I) mbt2018 8



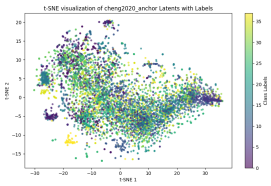
(J) mbt2018_mean 2



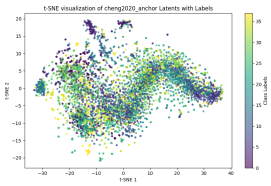
(K) mbt2018_mean 5



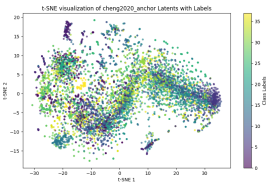
(L) mbt2018_mean 8



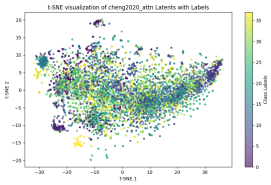
(M) cheng2020_anchor 2



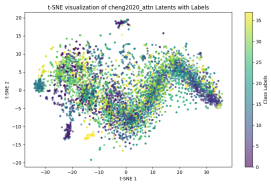
(N) cheng2020_anchor 4



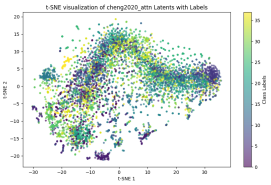
(O) cheng2020_anchor 6



(P) cheng2020_attn 2

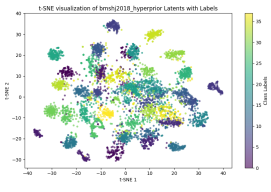


(Q) cheng2020_attn 4

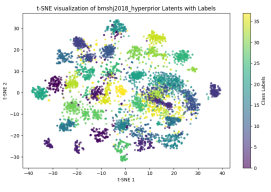


(R) cheng2020_attn 6

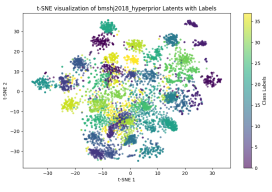
FIGURE 3.12: t-SNE visualizations of models constructed latents of PatternNet test set, with labels.



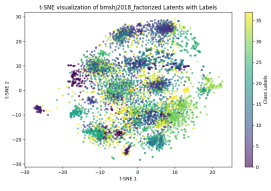
(A) bmshj2018_hyperprior 2



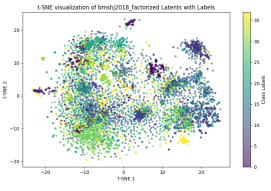
(B) bmshj2018_hyperprior 5



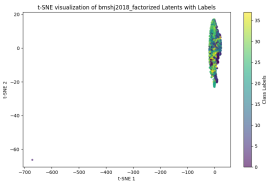
(C) bmshj2018_hyperprior 8



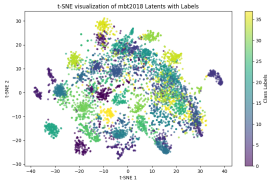
(D) bmshj2018_factorized 2



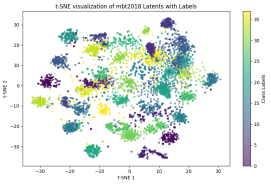
(E) bmshj2018_factorized 4



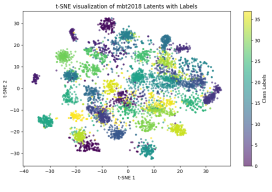
(F) bmshj2018_factorized 8



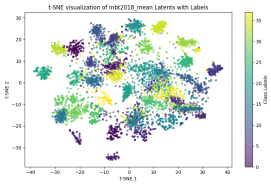
(G) mbt2018 2



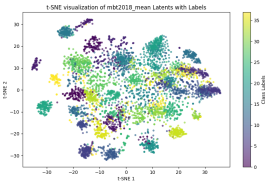
(H) mbt2018 5



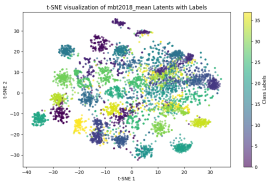
(I) mbt2018 8



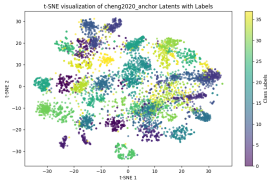
(J) mbt2018_mean 2



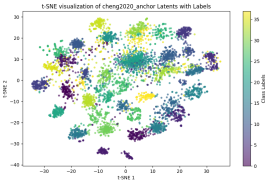
(K) mbt2018_mean 5



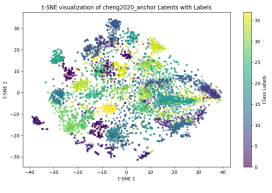
(L) mbt2018_mean 8



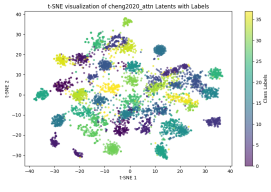
(M) cheng2020_anchor 2



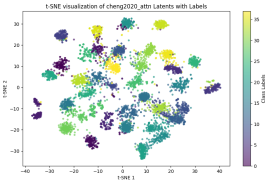
(N) cheng2020_anchor 4



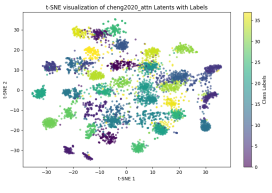
(O) cheng2020_anchor 6



(P) cheng2020_attn 2



(Q) cheng2020_attn 4



(R) cheng2020_attn 6

FIGURE 3.13: t-SNE visualizations of models constructed latents of PatternNet test set, with labels.

3.9 RSI-CB256 Dataset Results

JPEG Quality	**Classifier**	**Average BPP ↓**	**PSNR ↑**	**F1 ↑**	**RDAI ↑**
JPEG_10	MLP	**0.4337**	27.94	0.312	0.5775
JPEG_40	MLP	0.8949	33.58	0.4622	0.7537
JPEG_70	MLP	1.3401	36.37	0.4993	0.6952
JPEG_100	MLP	6.2916	**48.45**	0.4793	0.5520
JPEG_10	ResNet	**0.4337**	27.94	0.7932	0.7378
JPEG_40	ResNet	0.8949	33.58	0.9819	0.8165
JPEG_70	ResNet	1.3401	36.37	**0.9860**	**0.8186**
JPEG_100	ResNet	6.2916	**48.45**	0.9837	0.7200
JPEG_10	ViT	**0.4337**	27.94	0.6154	0.6786
JPEG_40	ViT	0.8949	33.58	0.9012	0.7897
JPEG_70	ViT	1.3401	36.37	0.9286	0.7995
JPEG_100	ViT	6.2916	**48.45**	0.9323	0.7028

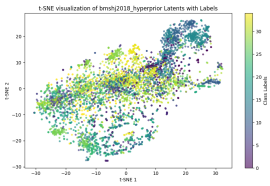
TABLE 3.9: Performance of pre-trained classifiers on different JPEG compression levels. The table shows the values of Average BPP, PSNR, F1, and RDAI for MLP, ResNet, and ViT classifiers applied to the RSI-CB256 dataset.

Method	**MLP**					**ResNet50**					**Transformer**				
	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑	Accuracy ↑	F1 ↑	BPP ↓	PSNR ↑	RDAI ↑
bmshj2018_hyperprior_2	0.6194	0.5903	0.1787	29.3387	0.6864	0.319	0.2531	0.1597	29.3309	0.5748	0.8271	0.8261	0.2060	29.3277	0.7640
bmshj2018_hyperprior_5	0.7598	0.7535	0.5149	34.6062	0.7588	0.6368	0.6384	0.6018	34.6086	0.7176	0.8814	0.8797	0.7021	34.6110	0.7947
bmshj2018_hyperprior_8	0.7851	0.7705	1.725	41.2658	0.7612	0.6947	0.6839	1.3649	41.2805	0.7444	**0.876**	**0.8737**	1.307	41.2819	0.8095
bmshj2018_factorized_2	0.3432	0.3038	0.2076	28.3281	0.5845	0.5467	0.5344	0.1951	28.2901	0.6615	0.7147	0.7147	0.2094	28.3179	0.7212
bmshj2018_factorized_5	0.5087	0.6840	0.6094	33.2297	0.7249	0.4356	0.4326	0.6577	33.2264	0.6396	0.7913	0.7891	0.4583	33.2080	0.7648
bmshj2018_factorized_8	0.5578	0.5124	1.2091	40.3815	0.6875	0.4374	0.4318	1.4207	40.3835	0.6536	0.818	0.8145	1.6786	40.3854	0.7725
mbt2018_mean_2	0.616	0.5864	0.2281	29.4829	0.6843	0.3465	0.3164	0.1212	29.4894	0.5980	0.7754	0.7700	**0.1354**	29.4935	0.7486
mbt2018_mean_5	0.7281	0.7165	0.5921	35.1885	0.7472	0.5081	0.4922	0.5941	35.2007	0.6725	0.8263	0.8258	0.3066	35.2019	0.7932
mbt2018_mean_8	0.7875	0.7764	1.4715	41.5130	0.7729	0.6368	0.6204	1.0462	41.5313	0.7353	0.8669	0.8671	1.4205	41.5296	0.8049
mbt2018_2	0.5489	0.5252	0.2295	29.5786	0.6644	0.218	0.1907	0.1677	29.5674	0.5550	0.7327	0.7267	0.1549	29.5779	0.7147
mbt2018_5	0.7362	0.7252	0.6394	35.3631	0.7495	0.3349	0.3467	0.6500	35.3693	0.6231	0.857	0.8563	0.4874	35.3656	0.7859
mbt2018_8	0.8032	0.7932	1.4109	41.6113	0.7811	0.6778	0.6911	1.5225	41.6259	0.7435	0.8715	0.8692	1.4960	**41.6188**	**0.8244**
cheng2020_anchor_2	0.6648	0.6478	0.2136	29.6633	0.7062	0.3192	0.3124	0.2051	29.6627	0.5948	0.8034	0.7993	0.2329	29.6612	0.7560
cheng2020_anchor_4	0.7388	0.7270	0.3470	33.3501	0.7486	0.5683	0.5803	0.3607	33.3510	0.6993	0.8499	0.8484	0.5859	33.3534	0.7811
cheng2020_anchor_6	0.7911	0.7805	1.1125	36.8395	0.7603	0.7491	0.7521	0.6999	36.8243	0.7645	0.8634	0.8626	0.7805	36.8401	0.7987
cheng2020_attn_2	0.6527	0.6360	0.2434	29.8007	0.7021	0.402	0.3807	0.1533	29.7939	0.6200	0.8012	0.7968	0.2016	29.7883	0.7569
cheng2020_attn_4	0.7172	0.7035	0.3629	33.4527	0.7408	0.5634	0.5491	0.5153	33.4597	0.6844	0.8347	0.8325	0.3945	33.4562	0.7828
cheng2020_attn_6	0.7594	0.7454	0.9044	36.8170	0.7554	0.6638	0.6667	0.7924	36.8132	0.7329	0.8602	0.8573	0.9029	36.8076	0.7927

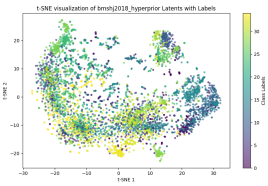
TABLE 3.10: Performance comparison using frozen weights for neural compression models applied to the RSI-CB256 dataset.

Method	**Accuracy**	**F1**	**BPP**	**PSNR**	**RDAI**
bmshj2018_hyperprior_2	0.9473	0.9465	0.3575	28.393	0.794
bmshj2018_hyperprior_5	0.9592	0.9587	0.6019	31.2434	0.806
bmshj2018_hyperprior_8	0.9608	0.9605	0.7183	31.2724	0.802
bmshj2018_factorized_2	0.9034	0.9020	**0.3524**	28.0888	0.778
bmshj2018_factorized_5	0.9275	0.9263	0.8842	32.3250	0.791
bmshj2018_factorized_8	0.9214	0.9205	1.1057	**35.5023**	0.800
mbt2018_mean_2	0.9293	0.9282	0.4197	30.0056	0.795
mbt2018_mean_5	0.9576	0.9573	0.5344	31.1483	0.807
mbt2018_mean_8	0.9620	0.9617	0.5955	32.3721	0.813
mbt2018_2	0.9416	0.9408	0.4445	29.6112	0.796
mbt2018_5	0.9576	0.9574	0.5408	31.1927	0.807
mbt2018_8	0.9642	0.9639	0.7441	31.9874	0.807
cheng2020_anchor_2	0.9618	0.9617	0.3905	29.4765	0.804
cheng2020_anchor_4	0.9485	0.9475	0.3779	30.7447	0.807
cheng2020_anchor_6	0.9576	0.9573	0.4891	32.5162	**0.816**
cheng2020_attn_2	0.9541	0.9536	0.4009	29.9058	0.803
cheng2020_attn_4	0.9644	0.9646	0.3258	29.9594	0.810
cheng2020_attn_6	**0.9715**	**0.9715**	0.4300	30.3372	0.811

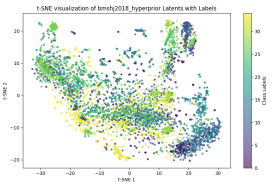
TABLE 3.11: Fine-tuning performance using various methods and Transformer classifier applied to the RSI-CB256 dataset.



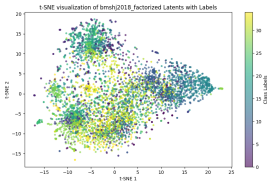
(A) bmshj2018_hyperprior 2



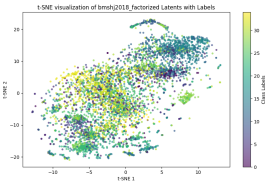
(B) bmshj2018_hyperprior 5



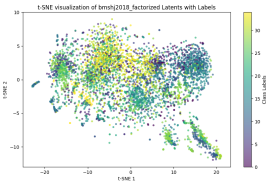
(C) bmshj2018_hyperprior 8



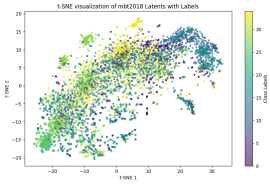
(D) bmshj2018_factorized 2



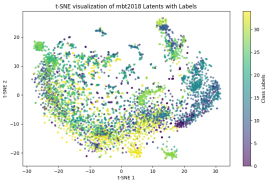
(E) bmshj2018_factorized 4



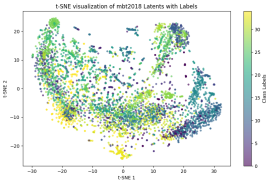
(F) bmshj2018_factorized 8



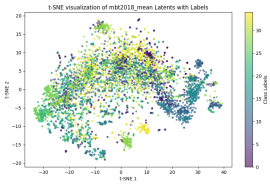
(G) mbt2018 2



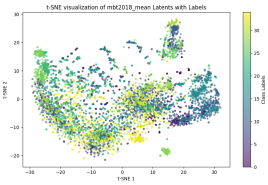
(H) mbt2018 5



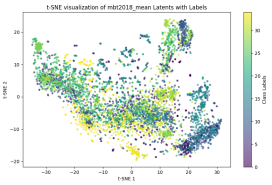
(I) mbt2018 8



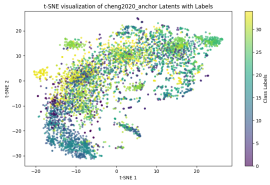
(J) mbt2018_mean 2



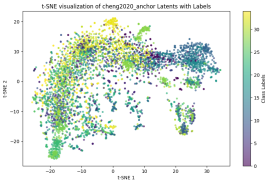
(K) mbt2018_mean 5



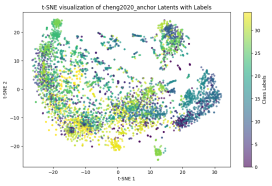
(L) mbt2018_mean 8



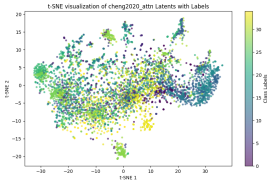
(M) cheng2020_anchor 2



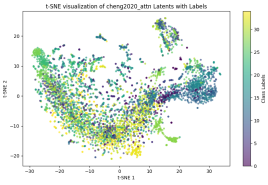
(N) cheng2020_anchor 4



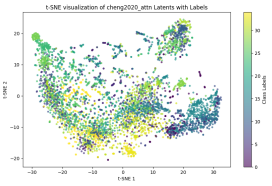
(O) cheng2020_anchor 6



(P) cheng2020_attn 2

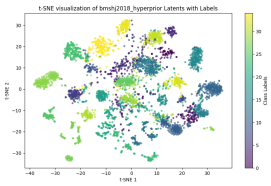


(Q) cheng2020_attn 4

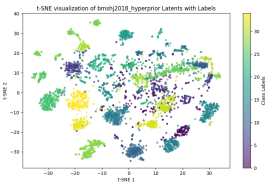


(R) cheng2020_attn 6

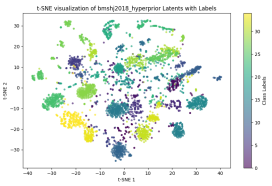
FIGURE 3.14: t-SNE visualizations of models constructed latents of RSICB-256 test set, with labels.



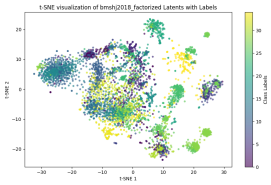
(A) bmshj2018_hyperprior 2



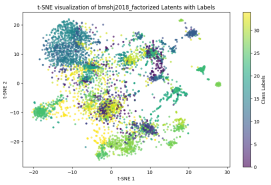
(B) bmshj2018_hyperprior 5



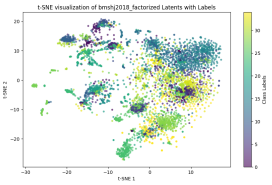
(C) bmshj2018_hyperprior 8



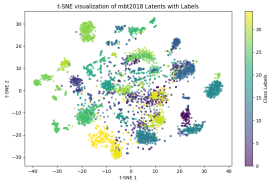
(D) bmshj2018_factorized 2



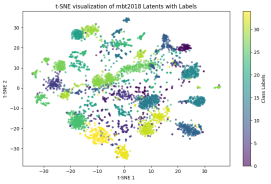
(E) bmshj2018_factorized 4



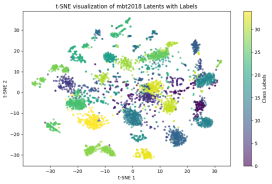
(F) bmshj2018_factorized 8



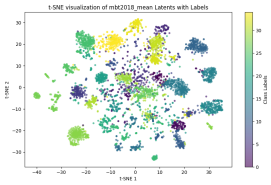
(G) mbt2018 2



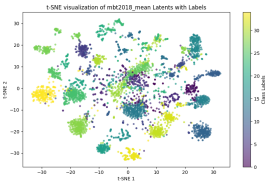
(H) mbt2018 5



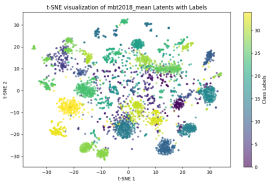
(I) mbt2018 8



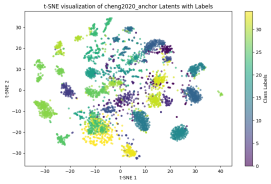
(J) mbt2018_mean 2



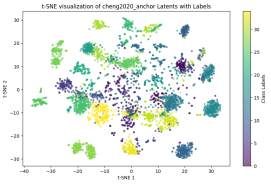
(K) mbt2018_mean 5



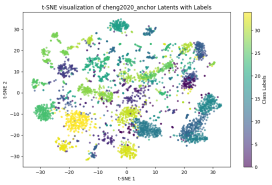
(L) mbt2018_mean 8



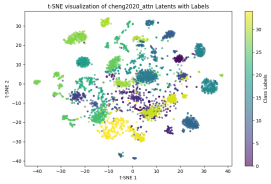
(M) cheng2020_anchor 2



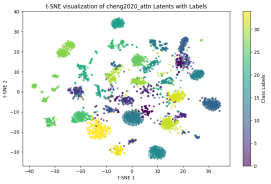
(N) cheng2020_anchor 4



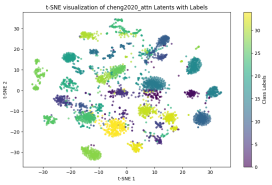
(O) cheng2020_anchor 6



(P) cheng2020_attn 2



(Q) cheng2020_attn 4



(R) cheng2020_attn 6

FIGURE 3.15: t-SNE visualizations of models constructed latents of RSICB-256 test set, with labels.

Chapter 4

Data Fusion Using VAE Latent Representations

The content of this chapter is a first version of the manuscript text for publication under the following citation:

Giuliano, A. (2024). Enhancing Data Fusion and Classification of Sentinel-1 and Sentinel-2 Imagery Using Neural Compression. Informatin Fusion.

Enhancing Data Fusion and Classification of Sentinel-1 and Sentinel-2 Imagery Using Neural Compression

Alessandro Giuliano
Faculty of Engineering
McMaster University, Hamilton, ON, Canada
Email: giuliana@mcmaster.ca

S. Andrew Gadsden
Faculty of Engineering
McMaster University
Email: gadsdesa@mcmaster.ca

John Yawney
Adastra Corp.
Email: john.yawney@adastragrp.com

Abstract

This study introduces a novel approach to data fusion and classification by integrating Synthetic Aperture Radar (SAR) and RGB images from the Sentinel-1 and Sentinel-2 satellites using compressive neural networks. By leveraging neural compression, we fuse the diverse data sources into a unified latent space, enhancing information content and enabling direct classification without the need for explicit decompression. The proposed method is evaluated against several advanced data fusion techniques, including PCA, DWT, and SA, as well as traditional JPEG compression, with a comprehensive analysis of classification accuracy, compression efficiency, and reconstruction quality. Our findings demonstrate that the proposed approach significantly outperforms conventional methods in classification performance through multimodal fusion. This work highlights the potential of neural compression models for latent data fusion in remote sensing applications.

Keywords: *Data Fusion, Neural Compression, Satellite SAR-Image Compression, Remote Sensing*

4.1 Introduction

The advent of advanced satellite technologies has led to an exponential increase in the volume and variety of remote-sensing data. Among the various types of sensors deployed, Synthetic Aperture Radar (SAR) and optical sensors from platforms such as Sentinel-1 and Sentinel-2 have proven very valuable due to their complementary capabilities. SAR provides all-weather, day-and-night imaging capabilities with excellent penetration through clouds and vegetation. At the same time, optical sensors offer high-resolution images that capture the visual characteristics of the Earth's surface. However, integrating these diverse data sources into a coherent and informative representation remains a significant challenge [292, 293, 294].

Traditional data fusion techniques, such as Principal Component Analysis (PCA) [295], Discrete Wavelet Transform (DWT) [296], and spectral analysis (SA) through fast Fourier transform [293], have long been utilized to integrate SAR and optical imagery. These methods aim to merge the complementary strengths of different data sources to create a more comprehensive representation of the observed scenes. PCA reduces dimensionality by transforming the data into a set of orthogonal components, DWT decomposes the data into distinct frequency components for more detailed analysis, and SA uses Fourier transform to analyze the spectral properties of the data. While these approaches have been valuable in various applications, they often fall short in several key aspects. One significant limitation of these traditional methods is their inability to maximize the information content from the diverse data sources fully. PCA, for instance, might discard subtle yet important details during the dimensionality reduction process. Similarly, DWT can sometimes lead to loss of spatial resolution due to its focus on frequency components. Consequently, these techniques may not capture the intricate relationships and complementary information inherent in SAR and optical data. Moreover, conventional compression techniques like JPEG, designed primarily for visual data, are not well-suited for preserving the critical information contained in remote sensing imagery. JPEG compression reduces the file size through lossy compression, which can eliminate fine details and introduce artifacts that degrade the data quality. This loss of information is especially problematic in remote sensing applications where precision and accuracy are paramount. For example, subtle changes in terrain or vegetation, which

are crucial for environmental monitoring and disaster management, may be obscured or lost entirely due to compression artifacts.

Recent advancements in machine learning, particularly in neural networks, have opened new avenues for addressing these challenges [297]. Neural compression and multimodal fusion techniques leverage the power of deep learning to create unified latent representations that capture the essential features of diverse data sources [261]. By employing these techniques, it is possible to enhance the information content and enable direct classification from the latent space without the need for explicit decompression. This approach improves classification accuracy and offers better compression efficiency and reconstruction quality compared to conventional methods.

This study introduces a novel approach that integrates SAR and RGB images from Sentinel-1 and Sentinel-2 satellites using compressive neural networks. The proposed method employs VAE-based neural compression to fuse diverse data sources into a unified latent space, thereby enhancing the information content and enabling direct classification without explicit decompression. The key contributions of this work are as follows:

- The introduction of a neural compression framework for efficient multimodal data fusion, offers a new way to process SAR and optical imagery together.
- Fusion of SAR and RGB data into a unified latent space, leading to an enriched and more informative data representation compared to traditional methods.
- Direct classification from the latent space without requiring decompression, significantly improving computational efficiency in the classification process.

The effectiveness of this approach is evaluated against advanced data fusion techniques such as PCA, DWT, and SA, as well as traditional JPEG compression. Comprehensive analyses of classification accuracy, compression efficiency, and reconstruction quality demonstrate that the proposed method significantly outperforms conventional techniques. These results highlight the potential of neural compression models in remote sensing data fusion and classification, offering enhanced performance and practicality for future applications.

4.2 Related Work

Remote sensing involves the collection of data about objects or phenomena from a distance, typically using sensors mounted on satellites and aircraft to gather information

about the Earth's surface, atmosphere, and oceans. This data provides insights into the topography, land cover, vegetation, pollution levels, and climate patterns, supporting a diverse range of applications across numerous fields [218]. These applications include environmental monitoring, urban planning, disaster management, climate change research, and defense operations. The sensors employed in these applications are often complementary, capturing various aspects of the Earth's surface. Data fusion is extensively utilized in remote sensing, encompassing both homogeneous fusion techniques, such as pansharpening, hyperspectral (HS) pansharpening, and spatiotemporal fusion, as well as heterogeneous fusion approaches, like LIDAR-optical and synthetic aperture radar (SAR)-optical fusion. These methods are employed for various applications, including land use and land cover (LULC) classification, object detection, change detection, and terrain monitoring [298]. Many publicly available data fusion datasets, beyond the scope of audiovisual fusion, have been contributed by the recurring IEEE GRS Data Fusion Contest, which focuses on tasks such as land cover classification and semantic urban reconstruction [299, 300].

4.2.1 Data Fusion in Remote Sensing

Multimodal data fusion is a critical field in modern remote sensing and data analysis, aiming to combine information from different sources to provide a more comprehensive understanding of complex phenomena. This approach is driven by the limitations of individual sensors and the complementary nature of the data they provide. For instance, while optical sensors capture detailed spectral information essential for land cover classification and material identification, their performance is hampered by cloud cover and varying illumination conditions. Conversely, Synthetic Aperture Radar (SAR) provides consistent spatial data unaffected by weather, but often suffers from speckle noise and lower interpretability. By integrating data from multiple sources, multimodal data fusion seeks to overcome these individual limitations, creating a richer dataset that enhances the accuracy and reliability of remote sensing applications [301]. The use of diverse data sources has significantly improved decision-making in fields ranging from environmental monitoring to disaster management.

Pixel-level fusion techniques represent one of the foundational approaches in multimodal data fusion, directly combining the pixel values from different sources to generate a single fused image. These techniques include Intensity-Hue-Saturation (IHS), and Brovey Transform, which aim to enhance spatial resolution while preserving spectral characteristics. Such methods are particularly useful in applications where high spatial resolution is crucial, such as urban mapping and agricultural monitoring. However,

pixel-level fusion often requires precise geometric registration of input images and can be computationally intensive, especially when dealing with large datasets or high-resolution images. Recent advancements have seen the integration of more sophisticated models, such as multi-scale decomposition and hybrid techniques, which combine several fusion methods to balance spatial and spectral fidelity effectively [293]. Despite these advancements, challenges remain in mitigating artifacts and ensuring that the fused images retain meaningful information for subsequent analysis.

4.2.2 Traditional Data Fusion Methods

Traditional multimodal fusion methods have been foundational in integrating data from multiple sources to enhance the accuracy and interpretability of the resulting information. These methods typically involve three levels of fusion: pixel-level, feature-level, and decision-level fusion, as illustrated in Figure 4.1. Pixel level fusion, while straightforward, often requires precise geometric alignment of the datasets and can be computationally intensive. Feature-level fusion, on the other hand, involves extracting and combining features from each modality before performing the analysis. This method allows for the integration of more abstract representations of the data, reducing the impact of registration errors and enhancing the robustness of the analysis. Decision-level fusion combines the outputs of individual classifiers or decision-making processes, which is useful in scenarios where the data sources are significantly heterogeneous or when the individual decisions are more reliable than the fused data itself. Multimodal fusion methods, such as multiset canonical correlation analysis (CCA), parallel factor analysis (PARAFAC), and various tensor decomposition techniques, have been extensively employed in remote sensing to exploit complementary information from diverse modalities.

Multiset CCA extends canonical correlation analysis to handle multiple datasets simultaneously, identifying shared structures across different modalities. This method has been used in integrating multispectral and SAR data for enhanced land cover classification and environmental monitoring, addressing challenges like varying spatial resolutions and sensor-specific noise [301, 302, 303].

PARAFAC is a well known technique for decomposing multi-way data into a sum of rank-1 tensors, which allows for a structured extraction of latent features from complex datasets. In remote sensing, it has been applied to fuse spatial and spectral information from hyperspectral images, thereby improving the robustness of data interpretation and reducing the impact of noise and misalignment. This method has proven effective for

applications such as vegetation mapping and soil analysis, where maintaining spectral integrity is essential [304].

Tensor decomposition methods, including canonical polyadic decomposition (CPD) and Tucker decomposition, extend these capabilities by modeling interactions in multi-way arrays common in high-dimensional datasets like those used in SAR-optical fusion. These techniques enable the integration of multimodal data while preserving the unique features of each modality, thereby improving the accuracy of applications such as terrain mapping and urban monitoring [301]. Recent advancements in SAR-optical fusion techniques have highlighted the effectiveness of hybrid and multi-scale decomposition approaches. For example, wavelet-based methods have been combined with component substitution techniques to overcome limitations such as spectral distortion and low spatial resolution, as demonstrated in recent studies on SAR-optical fusion [293]. These methods provide robust solutions to the challenges posed by multimodal data fusion, ensuring high-quality outputs that are essential for decision-making in diverse remote sensing applications [305].

Pixel-level fusion techniques for SAR and optical images include component substitution methods, multiscale decomposition methods, hybrid methods, and model-based methods [293]. Component substitution (CS) methods, such as Principal Component Analysis (PCA), Intensity-Hue-Saturation (IHS) transformation, and Brovey Transform, involve projecting multispectral data into another space where spatial and spectral information is separated. The spatial component is then substituted with high-resolution SAR data to enhance spatial details while maintaining spectral characteristics. However, these methods often result in spectral distortions due to differences in data characteristics between SAR and optical images. Multiscale decomposition methods, including wavelet and pyramid transforms, aim to overcome these limitations by decomposing the images into multiple scales and fusing them at different resolution levels, providing better localization in both spatial and spectral domains. Hybrid methods combine the strengths of CS and multiscale decomposition to reduce both spatial and spectral distortions. Model-based methods, such as sparse representation and variational models, treat fusion as a restoration problem and employ advanced mathematical models to achieve high-quality fusion results

4.2.3 Multimodal Data Fusion with Neural Networks

Deep learning has significantly advanced the field of multimodal data fusion by providing tools to capture complex, non-linear relationships between diverse data sources. Instead

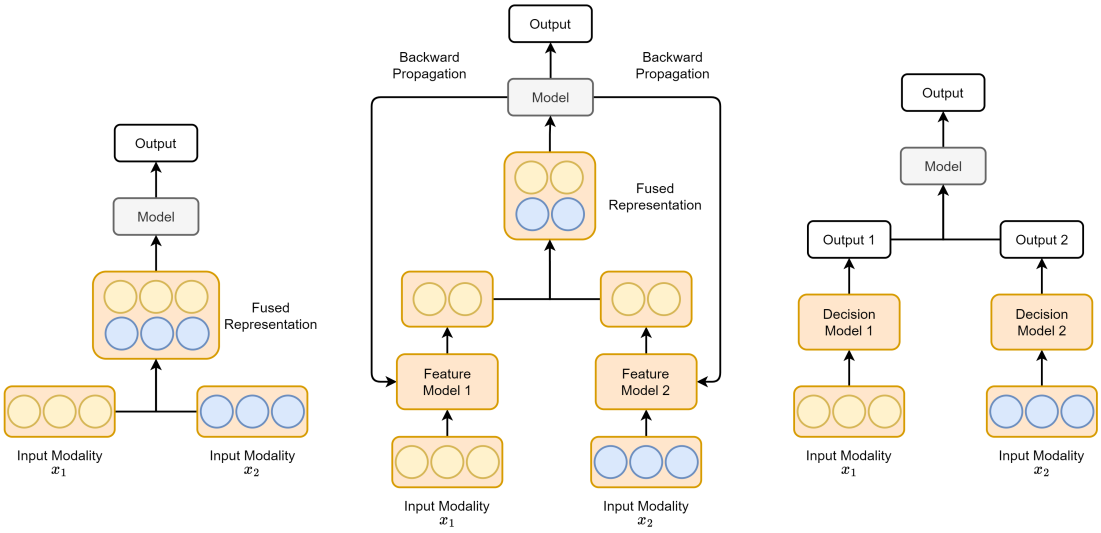


FIGURE 4.1: Illustration of different levels of multimodal data fusion: *Pixel-level fusion* (left): Input modalities are combined directly at the raw data level to create a fused representation, which is then processed by a model for output. *Feature-level fusion* (center): Features are extracted independently from each modality, then combined into a fused representation, which is processed by a model. *Decision-level fusion* (right): Each modality is processed separately, and their individual outputs are fused at the decision level to generate the final output.

of merely combining raw data, these methods focus on extracting and integrating features from each modality to build robust models that enhance performance across a range of applications. For example, Convolutional Neural Networks (CNNs) have been effectively employed to capture spatial patterns from Synthetic Aperture Radar (SAR) images and rich spectral information from optical images, leading to superior performance in tasks such as object detection and land cover classification. Generative Adversarial Networks (GANs) have been used to generate high-quality fused images, synthesizing the strengths of each modality while reducing the noise and artifacts often associated with traditional fusion techniques [306].

Recent research has explored various deep learning architectures to model complex intermodal relationships. Gao et al. [307] reviewed the use of Deep Belief Networks (DBN) and Stacked Autoencoders (SAE) for multimodal data fusion, demonstrating their effectiveness in applications ranging from image annotation to medical diagnosis. Similarly, Zhang et al. [308] emphasized the importance of tailored fusion strategies—early, late, or hybrid—specifically for semantic image segmentation, to fully exploit the complementary strengths of different data sources. These approaches highlight the need for innovative methods that can effectively integrate diverse data types to improve classification and interpretation.

The use of Transformers in multimodal data fusion has recently gained traction due to their ability to handle multiple modalities with minimal architectural modifications. Akbari et al. [309] introduced a convolution-free Transformer model for self-supervised learning from raw video, audio, and text data, achieving state-of-the-art results across various downstream tasks. This modality-agnostic approach shows potential for enhancing feature extraction and fusion in remote sensing applications. Xu et al. [310] provide a comprehensive review of multimodal learning with Transformers, underscoring their advantages in achieving effective cross-modal interactions and modality-agnostic processing.

Additionally, Shi et al. [311] proposed a Variational Mixture-of-Experts Autoencoder, which facilitates coherent joint and cross-generation of multimodal data, enhancing the flexibility and robustness of fusion models.

In the context of remote sensing, traditional methods such as PCA, DWT and SA have been widely used for data fusion, but they often fall short in preserving the rich information contained in multimodal datasets.

The continuous development of advanced deep learning techniques in multimodal data fusion offers promising prospects for remote sensing, as evidenced by the significant improvements in classification accuracy, compression efficiency, and reconstruction quality achieved through neural compression models. This progress suggests a future direction for leveraging these sophisticated architectures to enhance the accuracy and interpretability of satellite image analysis.

4.2.4 Compressive Neural Networks in Remote Sensing

In recent years, neural compression has gained attention for its application to satellite imagery, although research in this area remains limited. A notable contribution is the work by Oliveira et al., which introduces a low-complexity VAE specifically designed for on-board satellite image compression [263]. This model addresses constraints related to time and memory while preserving compression performance. The approach simplifies the entropy model by demonstrating that most features follow a Laplacian distribution, replacing complex non-parametric techniques with a straightforward parametric estimation. The proposed model surpasses the Consultative Committee for Space Data Systems (CCSDS) standard and holds its own against cutting-edge learned compression methods.

Additionally, a few other studies have explored neural compression for satellite images with encouraging outcomes [282, 283]. However, efficiently compressing and transmitting the rapidly growing volumes of satellite data remains a significant challenge. This research extends the current work by investigating the use of Variational Autoencoders (VAEs) and other sophisticated neural compression models for satellite image analysis, with the goal of utilizing the latent space directly for subsequent machine learning tasks and data fusion.

4.3 Methodology

4.3.1 Proposed Architecture

In this work, we introduce a novel architecture for satellite image classification that leverages latent space representations generated by neural compression models. Our approach is motivated by the observation that latent representations, which are compact and informative versions of the original data, can be directly utilized for downstream tasks like classification without the need for full image reconstruction. By fusing the latent spaces of different data modalities—specifically, optical and Synthetic Aperture

Radar (SAR) images—our method effectively combines complementary information from both sources, leading to enhanced classification performance.

The proposed architecture consists of three primary stages: individual compression of optical and SAR images, fusion of their latent representations, and classification based on the fused latent space.

At the core of our approach is the use of Variational Autoencoders (VAEs) for neural compression. Both optical and SAR images are independently compressed using a set of pre-trained VAEs from the CompressAI library. These VAEs, including models such as *bmshj2018_factorized*, *bmshj2018_hyperprior*, and *cheng2020_attn*, were trained on large image datasets (e.g., Vimeo90K) to learn compact, lower-dimensional representations of the input images.

Each VAE consists of an encoder $E(\cdot)$ that transforms the input image into a latent space, and a decoder $D(\cdot)$ that reconstructs the image from this representation. The encoder is probabilistic, meaning it outputs a distribution over the latent space instead of a deterministic point. This design allows the model to capture uncertainty and variability in the data, which is crucial for efficient compression.

Let x_{optical} and x_{sar} denote the input optical and SAR images, respectively. The encoder $E(\cdot)$ maps each image to its corresponding latent representation:

$$z_{\text{optical}} = E(x_{\text{optical}}), \quad z_{\text{sar}} = E(x_{\text{sar}})$$

The latent spaces z_{optical} and z_{sar} are compact representations of the original images, containing the most essential features while reducing the dimensionality significantly. These latent spaces are further quantized and entropy encoded to reduce their size for transmission over communication channels, ensuring efficient use of bandwidth. At the receiving end, the encoded representations are entropy decoded back into the latent spaces for further processing.

One of the key contributions of our architecture is the fusion of latent representations from multiple modalities—optical and SAR images. Rather than relying on the individual characteristics of each image modality for classification, we combine their latent spaces to create a unified representation that integrates the strengths of both data sources.

The fusion process is performed element-wise, where the latent representation z_{optical} of the optical image and the latent representation z_{sar} of the SAR image are added together to form the fused latent space z_{fused}:

$$z_{\text{fused}} = z_{\text{optical}} + z_{\text{sar}}$$

This element-wise addition serves to merge complementary information from the two modalities. Optical images provide detailed visual information that is affected by weather and lighting conditions, while SAR images capture structural details independent of these factors. By fusing the two, we create a representation that is both robust and rich in features, suitable for more accurate classification.

The fused latent space z_{fused} is then entropy encoded and transmitted through the communication channel, similar to the individual latent spaces. Once decoded, it serves as the input for both image reconstruction and classification.

After decoding the fused latent representation, our architecture proceeds with classification. The fused latent space, z_{fused}, is used as input to a Multi-Layer Perceptron (MLP) classification model, which is tasked with predicting the class label of the input images. This classification model, denoted $C(\cdot; \theta_C)$, is trained to learn decision boundaries that separate the different classes based on the fused features.

The classification process can be described mathematically as follows: Given the fused latent representation z_{fused}, the classifier outputs the predicted class probabilities $\hat{y}$:

$$\hat{y} = C(z_{\text{fused}}; \theta_C)$$

The model is trained using a standard cross-entropy loss function, which compares the predicted class probabilities $\hat{y}$ with the ground truth labels y:

$$\mathcal{L}_{\text{class}} = -\sum_i y_i \log(\hat{y}_i)$$

Here, y_i is the true label for class i, and $\hat{y}_i$ is the predicted probability for that class. By minimizing this loss function, the model learns to improve its classification accuracy.

An important aspect of our architecture is the separation of tasks between the VAE models and the classifier. The VAE models are responsible for feature extraction, transforming the high-dimensional input images into compact latent spaces that capture essential information. Once the latent spaces are generated, the VAE weights are frozen, ensuring that the learned representations remain stable throughout the training process. This allows the classifier to focus solely on learning the decision boundaries required for accurate classification.

Freezing the VAE weights also reduces the computational complexity of training. Rather than training the entire architecture end-to-end, we can focus on optimizing the classifier, which is a much simpler model compared to the VAE. This division of labor simplifies the training process and accelerates convergence.

In addition to classification, our architecture allows for the reconstruction of the original images from the fused latent space. Although classification is the primary task, the ability to reconstruct images is a useful secondary feature that demonstrates the quality of the latent space encoding. Using the probabilistic decoders of the VAE models, the fused latent space can be decoded back into its original optical and SAR components, providing visual confirmation of the compressed and fused representations' effectiveness.

The proposed architecture offers several key advantages:

- **Compact and Efficient Representations**: By using VAEs for neural compression, the architecture generates highly compact latent spaces that preserve essential image features while significantly reducing the dimensionality.
- **Fused Representation**: The fusion of optical and SAR latent spaces creates a unified representation that leverages complementary information from both modalities, improving classification accuracy.
- **Modular Design**: The separation of feature extraction and classification tasks simplifies training, reduces computational complexity, and speeds up convergence.
- **Dual Functionality**: In addition to classification, the architecture allows for the reconstruction of original images from the fused latent space, ensuring that the latent representations are informative and compact.

4.3.2 Neural Compression Models

To focus on the fusion of the latent manifolds from different data modalities, we exclusively selected VAE-based neural compression models. Specifically, we utilized models

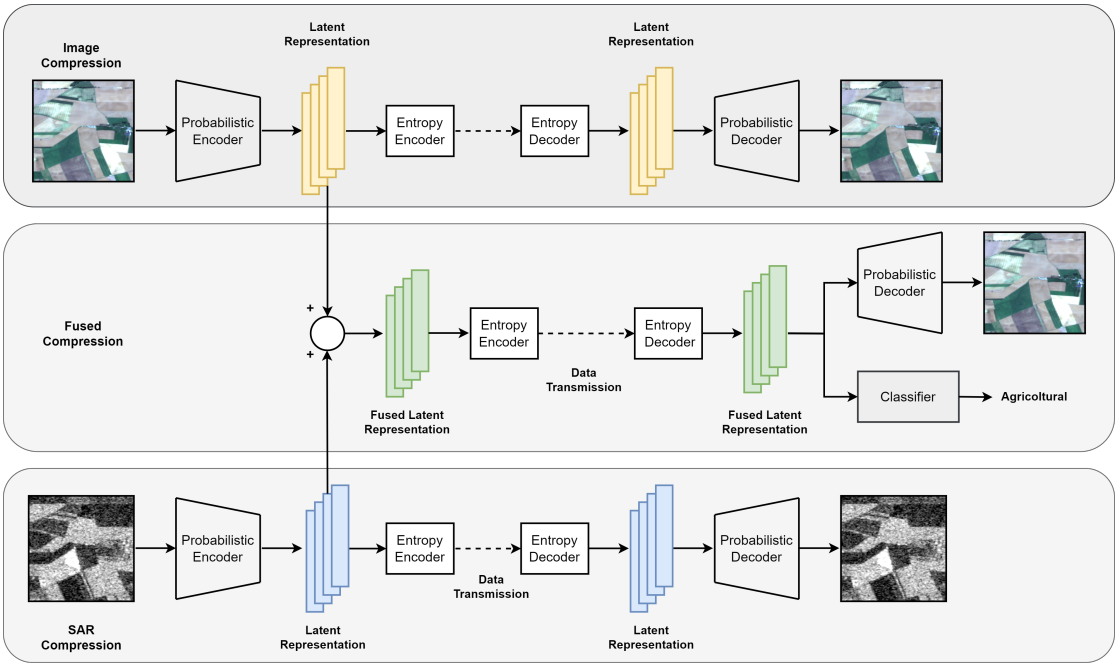


FIGURE 4.2: Proposed architecture: The top box illustrates conventional image transmission using neural compression; the bottom box shows conventional SAR image transmission using neural compression. The center represents the fused latent space transmission, with options for transforming it back to an image or using the latent space directly for classification.

that employ Gaussian latent variables and discretized Gaussian mixture models, as they provide flexible and expressive latent representations. The Gaussian latent variable models capture the continuous underlying structure of the data, while the discretized Gaussian mixture models enhance the model's ability to capture more complex data distributions. This choice was motivated by the need to fuse latent spaces in a way that preserves the essential features of both image modalities, optical and SAR, while maintaining the probabilistic interpretability inherent to VAEs. By leveraging these specific latent space formulations, we ensure that the fused representations retain both the compactness and informativeness required for downstream classification tasks, improving overall model performance.

Neural Compression Model	**Number of Parameters**
bmshj2018_hyperprior_2	4,968,963
bmshj2018_hyperprior_5	4,968,963
bmshj2018_hyperprior_8	11,582,275
bmshj2018_factorized_2	2,887,363
bmshj2018_factorized_5	2,887,363
bmshj2018_factorized_8	6,788,675
mbt2018_mean_2	6,921,123
mbt2018_mean_5	17,327,651
mbt2018_mean_8	17,327,651
mbt2018_2	13,896,419
mbt2018_5	25,270,548
mbt2018_8	25,270,548
cheng2020_anchor_2	11,726,269
cheng2020_anchor_4	26,364,908
cheng2020_anchor_6	26,364,908
cheng2020_attn_2	13,076,413
cheng2020_attn_4	29,397,740
cheng2020_attn_6	29,397,740

TABLE 4.1: Number of parameters in various neural compression models.

4.3.2.1 Gaussian Latent Variable Models

The *bmshj2018_factorized* model employs a fully factorized density model for the latent variables. Each latent variable y_i is assumed to be independent and identically distributed (i.i.d.), allowing the prior distribution over the latent variables to be factorized as:

$$p(y) = \prod_i p(y_i) \tag{4.1}$$

This model uses a non-parametric piecewise linear density to approximate each factor of the prior. The density p is derived from its cumulative distribution function (CDF) c, where:

$$p(x) = \frac{\partial c(x)}{\partial x} \tag{4.2}$$

The cumulative function c is designed to be monotonic, mapping $\mathbb{R}$ to $[0, 1]$, ensuring a valid density function. This is constructed through a series of transformations:

$$c = f_K \circ f_{K-1} \circ \cdots \circ f_1 \tag{4.3}$$

with the density derived from the derivatives of these transformations:

$$p = f'_K \cdot f'_{K-1} \cdot \cdots \cdot f'_1$$

Here, f_k are vector functions composed of matrices $H^{(k)}$, biases $b^{(k)}$, and element-wise nonlinearities g_k, defined as:

$$g_k(x) = x + a^{(k)} \odot \tanh(x) \tag{4.4}$$

where $a^{(k)}$ are vectors controlling the expansion or contraction rate. This setup efficiently approximates the latent distribution using piecewise linear segments, which is useful for compression.

The *bmshj2018_hyperprior* model builds upon this by introducing a hyperprior to capture spatial dependencies among the elements of the latent representation y. This is achieved by using an auxiliary autoencoder that models the latent scales with another set of latent variables z, such that:

$$z = h_a(y; \phi_h) \tag{4.5}$$

and the conditional distribution of the latent variables given the hyperprior is modeled as a Gaussian distribution:

$$p(y|z) = \mathcal{N}(y; 0, \sigma^2(z; \theta_h)) \tag{4.6}$$

Here, h_a and h_s represent the analysis and synthesis transforms in the auxiliary autoencoder, respectively. The hyperprior enhances the entropy model by conditioning the latent variable y on the auxiliary latent variable z, leading to more accurate, spatially adaptive entropy estimates.

The *mbt2018_mean* model further extends the *bmshj2018_hyperprior* by incorporating mean prediction alongside the scale prediction for the Gaussian distribution of the latent variables. The latent variable distribution is modeled as:

$$p(y|z) = \mathcal{N}(y; \mu(z; \theta_h), \sigma^2(z; \theta_h)) \tag{4.7}$$

This addition allows the model to capture not only the scale but also the mean shift in the latent space, leading to a more flexible and accurate entropy model. The model predicts both the mean μ and scale σ from the hyperprior, improving the ability to compress complex data distributions.

The *mbt2018* model combines the hyperprior with an autoregressive context model, further refining the entropy estimation by incorporating dependencies between latent variables. The context model conditions the current latent variable on previously decoded latents, using the following autoregressive distribution:

$$p(y_i|y_{<i}, z) = \mathcal{N}(y_i; \mu(y_{<i}, z), \sigma^2(y_{<i}, z)) \tag{4.8}$$

By using previously decoded latents $y_{<i}$ in combination with the hyperprior z, this model improves the prediction of each latent variable, leading to a more accurate and efficient entropy model.

4.3.2.2 Discretized Gaussian Mixture Models

In addition to Gaussian latent variable models, we employed models based on discretized Gaussian mixture models, which provide a more flexible and expressive representation

of the latent spaces. These models are effective at capturing complex distributions in the latent space, improving the accuracy of the compression and fusion processes.

The *cheng2020_anchor* model introduces discretized Gaussian mixture likelihoods to model the distribution of the latent variables more effectively. Instead of assuming a single Gaussian distribution for each latent variable, this model uses a mixture of K Gaussian components, each with its own mean and variance, to provide a more expressive and flexible prior. The conditional probability of the latent variable y given the hyperprior z is formulated as:

$$p(y|z) = \sum_{k=1}^{K} w^{(k)} \mathcal{N}(y; \mu^{(k)}(z), \sigma^{2(k)}(z)) \tag{4.9}$$

Here, $w^{(k)}$, $\mu^{(k)}(z)$, and $\sigma^{2(k)}(z)$ represent the mixture weights, means, and variances of the k-th Gaussian component, respectively, all of which are conditioned on the hyperprior z. This mixture model is good at capturing the complex, multimodal nature of the latent space, allowing the model to handle a wider variety of spatial structures and features in the data. By using multiple Gaussian components, the model can better approximate the true distribution of the latent codes, reducing spatial redundancy and significantly improving the compression performance.

The flexibility of the discretized Gaussian mixture model allows it to model intricate variations in the data more accurately, especially when there are discontinuities or complex dependencies within the latent space. This improvement in entropy modeling leads to better compression, especially when combined with the spatial dependencies captured by the hyperprior.

Finally, the *cheng2020_attn* model extends the performance of the image compression system by incorporating attention mechanisms. Attention modules help the model focus on specific regions of the image that contain more complex or important information, enhancing the overall rate-distortion performance. The attention mechanism is integrated directly into the model's architecture, modifying the convolutional features to prioritize information-rich areas. Mathematically, this is expressed as:

$$y = f(x; \theta, A) \tag{4.10}$$

where A represents the parameters of the attention module, and f is the function parameterized by θ that includes the attention-based transformations. The attention module dynamically adjusts the focus of the network during compression, leading to a more efficient allocation of the available bits and better compression performance. This mechanism is beneficial for complex image regions where simple spatial dependencies are insufficient, further boosting both compression efficiency and the quality of the latent representations.

4.3.3 Classification Model

The MLP model used in our approach is designed to process the latent space representations produced by the VAE models. These latent representations serve as compact, informative summaries of the input satellite and SAR images, and the role of the MLP is to map these representations to the corresponding class labels.

The MLP architecture consists of three fully connected layers, each responsible for progressively refining the extracted features from the latent space. The input dimension of the MLP is determined by the size of the latent space produced by the VAE, which varies depending on the complexity of the data and the architecture of the compression model. Between each fully connected layer, non-linear activation functions are applied to introduce non-linearity, allowing the model to learn complex decision boundaries, structured as follows:

$$\begin{aligned}
&\text{Layer 1:} \quad \mathbf{h}_1 = \sigma(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) \\
&\text{Layer 2:} \quad \mathbf{h}_2 = \sigma(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2) \\
&\text{Output Layer:} \quad \hat{\mathbf{y}} = \mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3
\end{aligned}$$

where $\mathbf{z}$ is the input latent representation, $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weights and biases of layer i, σ is the activation function (ReLU), and $\hat{\mathbf{y}}$ is the output class probabilities.

In addition to non-linearity, dropout regularization is employed between layers to prevent overfitting by randomly deactivating a fraction of neurons during training. The final layer of the MLP is a softmax output layer, which outputs class probabilities for classification.

4.3.4 Dataset

The dataset used in this study originates from the SEN1-2 dataset, introduced by Schmitt et al. [312]. The SEN1-2 dataset consists of synthetic aperture radar (SAR) and optical (RGB) image pairs collected by the Sentinel-1 and Sentinel-2 satellites, part of the European Space Agency's Copernicus program. This dataset contains 282,384 pairs of SAR and optical image patches, sampled from locations across the globe and covering all four meteorological seasons. The dataset is designed to support research in SAR-optical data fusion, providing co-registered multi-modal images with diverse geographic and environmental coverage.

For our study, we specifically curated a subset of image pairs. These images were manually selected from the SEN1-2 dataset to represent four distinct land cover classes: barren land, grassland, agricultural land, and urban areas. This selection was made to ensure that each class reflects a wide range of visual and environmental variability. Representative optical images from each class are included in subsequent sections to illustrate this diversity.

This dataset is well-suited for training deep learning models, such as Conditional Generative Adversarial Networks (Conditional GANs) and Variational Autoencoders (VAEs). The complexity of the SAR and optical images, characterized by their irregular spatial patterns, lack of geometric consistency, and varied orientations, makes it an ideal benchmark for evaluating model robustness across diverse tasks. The dataset's complexity also provides an excellent testbed for generative modeling, data fusion, and classification challenges in remote sensing.

By leveraging the rich variability in the SEN1-2 dataset, this study aims to explore the performance of deep learning models under non-ideal conditions, testing their adaptability and generalization to complex, real-world scenarios.

4.3.5 Quality Metrics

- **Rate Distortion Accuracy Index (RDAI)** To compare the efficacy of fine-tuning and assess the results of various neural compression and classification models on both reconstruction quality and classification performance, we introduce the Rate Distortion Accuracy Index (RDAI) inspired by the works of Luo et al. [287] on the rate distortion accuracy tradeoff in JPEG. This novel metric integrates the critical aspects of rate, distortion, and accuracy, providing a comprehensive evaluation framework for neural compression methods. The RDAI is defined as a

weighted combination of rate, distortion, and accuracy. Given that Bits Per Pixel (BPP) will be between 0 and 10, PSNR between 0 and 60, and F1 between 0 and 1, we normalize PSNR to the range [0, 1] as follows:

$$\text{PSNR}_{\text{normalized}} = \frac{\text{PSNR}}{60} \tag{4.11}$$

The formula for RDAI is then given by:

$$\text{RDAI} = \alpha \cdot \left(\frac{10 - \text{BPP}}{10}\right) + \beta \cdot \left(\frac{\text{PSNR}}{60}\right) + \gamma \cdot \text{F1} \tag{4.12}$$

where α, β, and γ are the weights assigned to each component, such that $\alpha+\beta+\gamma = 1$. During this study, we gave all three components equal weight and importance, and therefore $\alpha = \beta = \gamma = \frac{1}{3}$.

- **Relative Bias**

$$\text{Relative Bias} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{y}_i - y_i}{y_i}\right) \tag{4.13}$$

This metric measures the average deviation of the estimated values ($\hat{y}_i$) from the true values (y_i) relative to the true values. It provides insight into whether the estimates are systematically over or under the true values.

- **Relative Variance**

$$\text{Relative Variance} = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} \tag{4.14}$$

This metric compares the variance of the estimated values ($\hat{y}$) to the variance of the true values (y). It indicates how the estimates' dispersion matches the true values' dispersion.

- **Relative Standard Deviation**

$$\text{Relative Standard Deviation} = \frac{\text{Std}(\hat{y})}{\text{Std}(y)} \tag{4.15}$$

This metric measures the ratio of the standard deviation of the estimated values ($\hat{y}$) to the standard deviation of the true values (y). It helps to assess the relative spread of the estimates compared to the true values.

- **Correlation Coefficient (Pearson)**

$$r = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{4.16}$$

  This metric measures the linear correlation between the true values (y_i) and the estimated values ($\hat{y}_i$). A value of r close to 1 indicates a strong positive correlation, while a value close to -1 indicates a strong negative correlation.

- **Universal Quality Index (UQI)**

$$\text{UQI} = \frac{4\mu_x\mu_y\sigma_{xy}}{(\mu_x^2 + \mu_y^2)(\sigma_x^2 + \sigma_y^2)} \tag{4.17}$$

  The UQI evaluates the quality of an image by considering the mean (μ), variance (σ^2), and covariance (σ_{xy}) of the true and estimated images. It is designed to be a more comprehensive measure of image quality than simple error metrics.

- **Spectral Angle Mapper (SAM)**

$$\text{SAM} = \frac{1}{N}\sum_{i=1}^{N}\arccos\left(\frac{\mathbf{y}_i \cdot \hat{\mathbf{y}}_i}{\|\mathbf{y}_i\|\|\hat{\mathbf{y}}_i\| + \epsilon}\right) \tag{4.18}$$

  SAM measures the spectral similarity between the true ($\mathbf{y}_i$) and estimated ($\hat{\mathbf{y}}_i$) spectral vectors. It is commonly used in remote sensing to compare spectral signatures.

- **Entropy**

$$H(X) = -\sum_{i} p(x_i)\log p(x_i) \tag{4.19}$$

  Entropy measures the amount of information or randomness in an image. A higher entropy value indicates a more complex image with more information content.

- **Standard Deviation**

$$\text{Std}(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{4.20}$$

  The standard deviation measures the amount of variation or dispersion of pixel values in an image. It indicates how much the pixel values deviate from the mean value of the image.

- **Spectral Frequency**

$$\text{Spectral Frequency} = \frac{1}{N}\sum_{i=1}^{N}|F(x_i)| \tag{4.21}$$

  Spectral frequency measures the average magnitude of the Fourier transform of an image. It provides insight into the frequency content of the image, which can be useful in analyzing texture and other patterns.

- **SSIM (Structural Similarity Index)**

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.22}$$

  The SSIM evaluates the visual similarity between two images by considering luminance, contrast, and structure. It is designed to mimic human perception of image quality.

- **PSNR (Peak Signal-to-Noise Ratio)**

$$\text{PSNR} = 10\log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right) \tag{4.23}$$

  PSNR measures the ratio between a signal's maximum possible power and the corrupting noise's power. It is commonly used to assess the quality of reconstruction in image compression.

4.3.6 Latent Space Visualization

To visualize and compare the latent spaces generated by the Variational Autoencoder (VAE), we utilized t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). Both are dimensionality reduction techniques well-suited for visualizing high-dimensional data in lower-dimensional spaces, typically two or three dimensions.

t-SNE operates by converting the similarities between data points in the high-dimensional space into joint probabilities and then tries to optimize the low-dimensional representation to preserve these similarities. This is achieved through a probabilistic approach where the similarity between two points in the high-dimensional space is represented by a Gaussian distribution, while in the low-dimensional space, it is represented by a Student's t-distribution with one degree of freedom (a Cauchy distribution). This choice

of distribution in the low-dimensional space helps to manage the so-called "crowding problem," where too many points are mapped too closely together.

Mathematically, t-SNE works as follows:

For each pair of points (i, j), t-SNE calculates the conditional probability $p_{j|i}$ that point j would be chosen as a neighbor of point i if neighbors were picked in proportion to their probability density under a Gaussian centered at i:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{4.24}$$

where x_i and x_j are the high-dimensional input data points, and σ_i is the variance of the Gaussian centered at x_i.

The joint probability p_{ij} is then symmetrized:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{4.25}$$

where N is the number of data points.

In the low-dimensional space, a similar joint probability q_{ij} is computed using a Student's t-distribution:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{4.26}$$

where y_i and y_j are the low-dimensional representations of the data points.

t-SNE aims to minimize the Kullback-Leibler (KL) divergence between the high-dimensional and low-dimensional joint probabilities:

$$\mathrm{KL}(P\|Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{4.27}$$

This is typically done using gradient descent, where the positions of the points in the low-dimensional space are iteratively adjusted to minimize the KL divergence.

t-SNE is particularly suitable for our application because it effectively captures the local structure of the data, making it easier to identify clusters and patterns within the latent space. By visualizing the latent spaces using t-SNE, we can gain insights into how well the VAE has learned to represent the data and how distinct the latent representations are for different classes. This visualization helps in evaluating the quality of the latent space and the effectiveness of the VAE model in capturing essential features of the satellite images.

T-SNE was used to project the high-dimensional latent representations into a 2D space using perplexity = 30, and 500 iterations, where each point represents a data sample's latent vector. The resulting 2D plot provides a visual comparison of the latent spaces, highlighting the clustering of data points based on their class labels, as shown in Figures **??**.

UMAP is another technique used for visualizing high-dimensional data. It is based on manifold learning techniques and is good at preserving the global structure of the data. UMAP works by constructing a high-dimensional graph of the data and then optimizing a low-dimensional graph to be as structurally similar as possible to the high-dimensional one. This makes UMAP an excellent complementary technique to t-SNE, as it can provide a different perspective on the data structure.

Mathematically, UMAP optimizes the following objective function:

$$L = \text{cross-entropy}(X, Y) = -\sum_{i \neq j} \Big[A_{ij} \log(B_{ij}) + (1 - A_{ij}) \log(1 - B_{ij}) \Big] \tag{4.28}$$

where A is the adjacency matrix of the high-dimensional data and B is the adjacency matrix of the low-dimensional representation.

UMAP was used to project the same high-dimensional latent representations into a 2D space, similarly to t-SNE. The resulting 2D plot provides another visual comparison of the latent spaces, highlighting the clustering of data points based on their class labels as shown in Figures 4.5.

Both t-SNE and UMAP are crucial for understanding the latent spaces produced by the VAE and for assessing the effectiveness of the model in capturing and representing the

underlying structure of the data. These techniques allow us to observe the distribution and separability of different classes in the latent space, which is essential for evaluating the performance of the model in classification tasks.

4.3.7 Experiments

The experiments were conducted using PyTorch and CUDA on an NVIDIA GeForce RTX 4070 Ti Super GPU (16 GB), paired with an Intel(R) Core(TM) i7-14700KF CPU running at 3400 MHz, with 20 cores and 28 logical processors. The system also included 128 GB of DDR5 RAM at 6000 MHz. Tensorboard was utilized to log all results, including time series, scalar, and image data.

4.4 Results

In this section, we present the results of various experiments designed to evaluate the performance of neural compression models in comparison to singular modalities and traditional data fusion techniques. The primary goal is to demonstrate the superiority of fused representations, where Sentinel-1 SAR and Sentinel-2 optical data are combined in the latent space, as compared to singular modality approaches. Additionally, we compare the neural compression-based fusion technique with other established fusion methods, such as Discrete Wavelet Transform (DWT), Spectral analysis, and Principal Component Analysis (PCA), to assess the advantages of latent space fusion as shown in other related publications as well [293].

4.4.1 Classification and Compression Performance of Neural Compression

The performance of neural compression models using fused representations of Sentinel-1 SAR and Sentinel-2 optical data consistently exceeded that of singular modalities (either SAR or optical) and JPEG compression across all quality metrics. This superiority is evident in terms of classification accuracy, compression efficiency, and image quality, as highlighted in Tables 4.2 and 4.3.

4.4.1.1 Baseline Comparison with JPEG

The results of traditional JPEG compression, as shown in Table 2, reveal its limitations in balancing compression efficiency and classification performance. For instance, at the lowest quality setting (JPEG_10), the F1 score was 0.543 with a PSNR of 25.12, while

the average BPP was 0.5111. Although increasing the quality to JPEG_100 improved the PSNR to 37.41, the F1 score only marginally increased to 0.5918, and the BPP escalated dramatically to 8.914. This demonstrates that while JPEG compression can preserve image quality at higher settings, it remains inefficient in terms of compression and offers limited improvements in classification accuracy.

In contrast, neural compression models, particularly those leveraging fused representations, exhibited far superior performance. The *bmshj2018_hyperprior* model, as detailed in Table 3, achieved an F1 score of 0.9181 at Quality 2, with a PSNR of 26.5543 and a BPP of 0.6226. This marks a significant improvement over JPEG_10, which had a much lower F1 score (0.543) with a comparable PSNR of 25.12 and a slightly lower BPP of 0.5111. Notably, fused neural compression models consistently outperformed JPEG across all metrics, striking an effective balance between compression efficiency and classification accuracy.

At Quality 5, the *bmshj2018_hyperprior* model continued to excel, with the fused representation achieving an F1 score of 0.9111, a PSNR of 30.8682, and a BPP of 0.4056. In comparison, JPEG_40, which has a similar PSNR of 28.93, produced a much lower F1 score of 0.5721 and a significantly higher BPP of 1.1732. Additionally, the RDAI metric, which reflects both compression efficiency and classification accuracy, further emphasizes the advantage of neural compression: the fused *bmshj2018_hyperprior* model at Quality 5 achieved an RDAI of 0.79, while JPEG_40 only scored 0.6450. At the highest quality setting (Quality 8), the model *bmshj2018_hyperprior* achieved an F1 score of 0.9493 with a PSNR of 34.9266 and a BPP of 0.5308 for the fused representation. In contrast, JPEG_70 recorded an F1 score of 0.6095, a PSNR of 30.73, and a much higher BPP of 1.8098. Even JPEG_100, which reached a PSNR of 37.41, only achieved an F1 score of 0.5918, with an inefficient BPP of 8.914. Similarly, the *bmshj2018_factorized* model demonstrated the value of fused representations over both singular modalities and JPEG compression. The *mbt2018_mean* model further underscored the superiority of fused neural compression. At Quality 2, the fused representation achieved an F1 score of 0.9064, a BPP of 0.6235, and a PSNR of 26.6673. In comparison, JPEG_40, which had a similar PSNR of 28.93, produced a significantly lower F1 score of 0.5721 with a higher BPP of 1.1732. At Quality 8, the *mbt2018_mean* model's fused data yielded an F1 score of 0.9335, a PSNR of 35.0763, and a BPP of 0.5244. In contrast, JPEG_70 recorded a lower F1 score of 0.6095 with a PSNR of 30.73 and a much higher BPP of 1.8098. The *cheng2020_attn* model, which incorporates attention mechanisms, further amplified the advantages of fused representations. At Quality 2, the fused modality

achieved an F1 score of 0.9014, a PSNR of 26.8801, and a BPP of 0.3686, outperforming all JPEG settings, including JPEG_100, which only managed an F1 score of 0.5918. At Quality 6, the fused representation reached an F1 score of 0.9468 with a PSNR of 32.2457 and a BPP of 0.2976.

Overall, across all neural compression models and quality levels, fused representations consistently demonstrated superior classification accuracy, compression efficiency, and image quality when compared to singular modalities and traditional JPEG compression. These results underscore the effectiveness of integrating multimodal data (SAR and optical) in remote sensing applications.

Method	**Classifier**	**Average BPP ↓**	**PSNR ↑**	**F1 ↑**	**RDAI ↑**
JPEG_10	MLP	0.5111	25.12	0.543	0.6362
JPEG_40	MLP	1.1732	28.93	0.5721	0.6450
JPEG_70	MLP	1.8098	30.73	0.6095	0.6462
JPEG_100	MLP	8.914	37.41	0.5918	0.4409

TABLE 4.2: Results for JPEG Quality Levels using MLP classifier.

4.4.1.2 Classification Performance: Fused vs Singular Modalities

The neural compression models using fused representations of Sentinel-1 SAR and Sentinel-2 optical data consistently outperform singular modalities (either SAR or optical alone) and JPEG compression across all quality metrics, especially in terms of classification accuracy. The fused representations lead to significantly higher F1 scores across all compression models and quality levels, with negligible impact on reconstruction quality, as highlighted in Tables 2 and 3.

For example, as shown in Table 3, the *bmshj2018 hyperprior* model at Quality 2 achieves an F1 score of 0.9181 for fused data, compared to 0.435 for optical-only data and 0.2725 for SAR-only data. This stark contrast illustrates how the fused latent representation effectively captures complementary information from both SAR and optical sources, enabling the model to extract more discriminative features for classification. As the compression quality improves, this trend persists: at Quality 8, the F1 score for fused data reaches 0.9493, while optical and SAR modalities achieve F1 scores of 0.4913 and 0.2884, respectively.

In addition to higher classification accuracy, the fused representations maintain strong image reconstruction quality. Despite combining two distinct data modalities, the neural

Model	Type	Quality	Accuracy	F1	BPP	PSNR	RDAI
bmshj2018_hyperprior	Image	2	0.435	0.3972	0.651	26.5543	0.59
bmshj2018_hyperprior	SAR	2	0.2725	0.145	0.5943	25.9357	0.51
bmshj2018_hyperprior	Fused	2	0.9184	0.9181	0.6226	26.5543	0.77
bmshj2018_hyperprior	Image	5	0.7594	0.7409	0.4472	30.8682	0.74
bmshj2018_hyperprior	SAR	5	0.3381	0.2382	0.364	31.7197	0.58
bmshj2018_hyperprior	Fused	5	0.9137	0.9111	0.4056	30.8682	0.79
bmshj2018_hyperprior	Image	8	0.4913	0.3919	0.5155	34.9266	0.64
bmshj2018_hyperprior	SAR	8	0.2884	0.1775	0.5462	39.4463	0.59
bmshj2018_hyperprior	Fused	8	0.9497	0.9493	0.5308	34.9266	0.83
bmshj2018_factorized	Image	2	0.4656	0.449	0.449	25.2253	0.61
bmshj2018_factorized	SAR	2	0.3291	0.2485	0.6642	23.4472	0.52
bmshj2018_factorized	Fused	2	0.7053	0.69	0.6743	25.2253	0.68
bmshj2018_factorized	Image	5	0.5531	0.504	0.5428	29.4571	0.65
bmshj2018_factorized	SAR	5	0.2762	0.1604	0.4932	29.4293	0.53
bmshj2018_factorized	Fused	5	0.6263	0.6222	0.518	29.4571	0.69
bmshj2018_factorized	Image	8	0.6344	0.6049	0.6824	34.15	0.70
bmshj2018_factorized	SAR	8	0.2456	0.0969	0.647	37.74	0.55
bmshj2018_factorized	Fused	8	0.5887	0.5813	0.6647	34.154	0.69
mbt2018	Image	2	0.4938	0.4779	0.6525	26.7201	0.62
mbt2018	SAR	2	0.3059	0.1911	0.5934	26.1436	0.52
mbt2018	Fused	2	0.7753	0.7693	0.623	26.7201	0.72
mbt2018	Image	5	0.5512	0.5024	0.9112	31.4213	0.64
mbt2018	SAR	5	0.3741	0.2641	0.8053	32.6866	0.58
mbt2018	Fused	5	0.94	0.9404	0.8582	31.4213	0.79
mbt2018	Image	8	0.3581	0.2923	0.4987	35.044	0.61
mbt2018	SAR	8	0.3881	0.283	0.4797	40.1978	0.63
mbt2018	Fused	8	0.9181	0.9178	0.4892	35.044	0.82
mbt2018_mean	Image	2	0.4356	0.4064	0.6528	26.6673	0.59
mbt2018_mean	SAR	2	0.3078	0.2551	0.8942	26.1977	0.53
mbt2018_mean	Fused	2	0.9066	0.9064	0.6235	26.6673	0.76
mbt2018_mean	Image	5	0.6849	0.6061	0.9349	31.3317	0.68
mbt2018_mean	SAR	5	0.42	0.3757	0.8392	32.5241	0.61
mbt2018_mean	Fused	5	0.9287	0.9278	0.887	31.3317	0.79
mbt2018_mean	Image	8	0.5697	0.4946	0.5277	35.1228	0.68
mbt2018_mean	SAR	8	0.4172	0.2852	0.5232	40.2193	0.63
mbt2018_mean	Fused	8	0.9337	0.9335	0.5244	35.0763	0.82
cheng2020_anchor	Image	2	0.5991	0.5693	0.397	26.9826	0.66
cheng2020_anchor	SAR	2	0.4084	0.2818	0.3264	26.5346	0.56
cheng2020_anchor	Fused	2	0.9169	0.9168	0.3617	26.9826	0.78
cheng2020_anchor	Image	4	0.7625	0.7564	0.5288	29.9655	0.73
cheng2020_anchor	SAR	4	0.4553	0.3121	0.4289	30.5567	0.59
cheng2020_anchor	Fused	4	0.9306	0.9302	0.4789	29.9655	0.79
cheng2020_anchor	Image	6	0.6903	0.6024	0.3572	32.3602	0.70
cheng2020_anchor	SAR	6	0.2931	0.1921	0.2382	34.7597	0.58
cheng2020_anchor	Fused	6	0.9153	0.9158	0.2977	32.3602	0.81
cheng2020_attn	Image	2	0.4431	0.373	0.4014	26.8801	0.59
cheng2020_attn	SAR	2	0.3916	0.2697	0.3359	26.5917	0.56
cheng2020_attn	Fused	2	0.9013	0.9014	0.3686	26.8801	0.77
cheng2020_attn	Image	4	0.7797	0.7754	0.5215	29.9436	0.74
cheng2020_attn	SAR	4	0.4781	0.3443	0.4293	30.4895	0.60
cheng2020_attn	Fused	4	0.9372	0.9367	0.4754	29.9436	0.80
cheng2020_attn	Image	6	0.6759	0.6006	0.3562	32.2457	0.70
cheng2020_attn	SAR	6	0.3209	0.2075	0.2416	34.6206	0.59
cheng2020_attn	Fused	6	0.9466	0.9468	0.2976	32.2457	0.82

TABLE 4.3: Classification Performance of Neural Compression Models: This table compares neural compression models across quality levels and data modalities (Sentinel-2 optical, Sentinel-1 SAR, and fused). Metrics include accuracy, F1 score (classification), BPP (compression efficiency), PSNR (image quality), and RDAI. Models (e.g., bmshj2018_hyperprior, mbt2018, cheng2020_anchor) are tested on individual modalities and their fusion. Fused representations consistently outperform in classification (F1) while maintaining competitive BPP and PSNR, highlighting the benefits of fusing Sentinel-1 and Sentinel-2 data for remote sensing.

compression models achieve comparable or even better PSNR and SSIM values compared to singular modalities, indicating that the fusion process preserves important structural and spectral details without degradation. This shows that the additional complexity introduced by fusion does not negatively affect the model's ability to accurately reconstruct the data.

These findings underscore the value of multimodal data fusion in remote sensing applications. SAR data captures critical structural details and performs well in poor visibility conditions, while optical data provides essential spectral information for more nuanced interpretation. When combined into a unified latent space, the neural models effectively leverage the strengths of both modalities, resulting in a robust improvement in classification performance without sacrificing the integrity of the original data during reconstruction. This fusion strategy has significant implications for enhancing the accuracy and efficiency of remote sensing tasks.

4.4.1.3 Reconstruction Quality: Fused Representations vs Singular Modalities

One of the primary concerns with fusing disparate modalities into a single latent representation is the potential impact on reconstruction quality. However, our results show that neural compression models maintain excellent reconstruction performance even when operating on fused data. The fused representations, despite containing a richer blend of information from both SAR and optical sources, do not suffer from significant degradation in terms of reconstruction quality metrics such as PSNR and SSIM. Moreover, this ability to preserve quality across modalities highlights the robustness of the neural compression framework in effectively balancing the complexity introduced by multi-modal fusion without sacrificing reconstruction fidelity.

For instance, in the *bmshj2018_hyperprior* model at Quality 8, the fused representation achieves a PSNR of 34.9266, which is virtually identical to the PSNR values for the optical-only and SAR-only reconstructions, which are 34.9266 and 39.4463, respectively. This indicates that despite the additional complexity introduced by combining two different data types, the model is able to effectively reconstruct the image without a loss in quality. Even at lower quality settings, the fused representation holds its ground: at Quality 2, the PSNR for fused data is 26.5543, compared to 26.5543 for optical-only and 25.9357 for SAR-only. These results underscore that reconstructing from the fused representation does not compromise the integrity of the original data, ensuring high-quality

outputs. This trend is consistent in all neural compression models as shown in Table 4.3.

4.4.1.4 Compression Efficiency and Rate-Distortion Trade-offs

In addition to the enhanced classification performance, fused representations also achieve superior compression efficiency. Across all models and quality levels, the fused modality demonstrates competitive bits-per-pixel (BPP) rates, often lower than the singular modalities for a similar level of classification accuracy and PSNR. For example, in the *bmshj2018_hyperprior* model at Quality 5, the fused representation achieves a BPP of 0.4056 with an F1 score of 0.9111 and a PSNR of 30.8682, whereas SAR-only data achieves a lower F1 score of 0.2382 but requires a BPP of 0.364 to do so. The fused representation, therefore, provides a better rate-distortion trade-off, ensuring efficient compression while maintaining high classification accuracy and image quality.

The RDAI metric, which encapsulates both rate distortion and classification performance, further highlights the efficiency of fused neural compression models. At Quality 5, the fused *bmshj2018_hyperprior* model achieves an RDAI score of 0.79, compared to 0.58 for SAR-only data. Similarly, in the *mbt2018_mean* model at Quality 2, the fused representation achieves an RDAI of 0.76, while SAR-only data scores just 0.53. While the highest RDAI is achieved by both the *cheng_2020_attn* and *mbt2018* models when fusing the representations at the highest compression quality setting surpassing the best JPEG results by 30%.

4.4.2 Quality Metrics Comparison

In this section, we compare the performance of the proposed neural compression models with traditional data fusion techniques, including Principal Component Analysis (PCA), Discrete Wavelet Transform (DWT), and Spatial Averaging (SA). The comparison is based on various quality metrics, both with and without reference images, as presented in Tables 4.4 and 4.5.

4.4.2.1 Quality Metrics with Reference Image

Table 4.4 highlights the quality metrics when the reference image is the original Sentinel-2 optical image. These metrics include Relative Bias, Relative Variance, Relative Standard Deviation, Correlation Coefficient, Universal Image Quality Index (UIQI), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Spectral Angle Mapper (SAM).

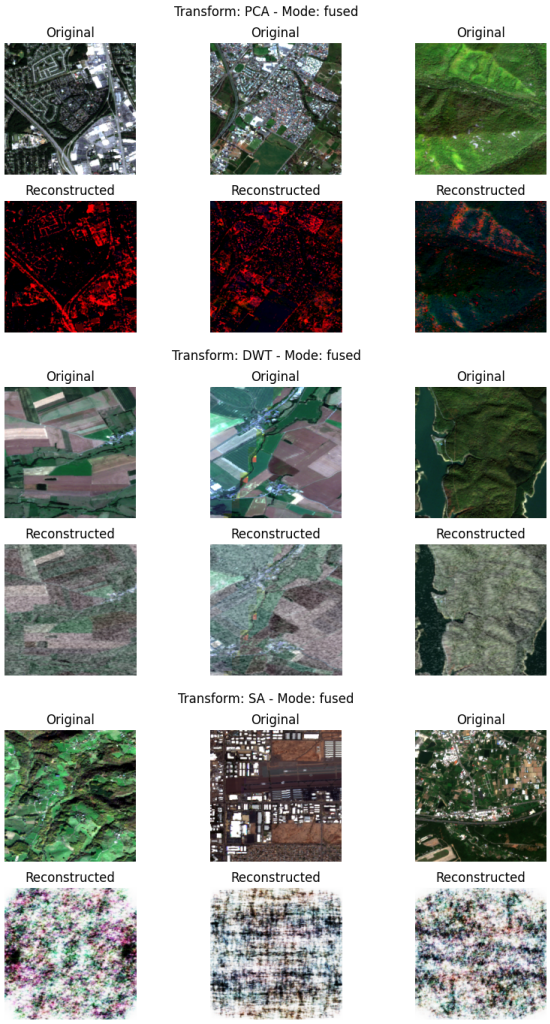
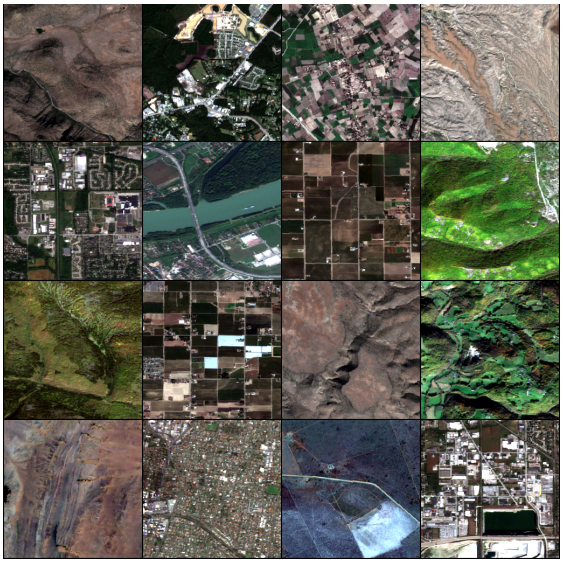


FIGURE 4.3: Fused representation comparison between original and reconstructed satellite images using different fusion techniques. Each row represents the original image (top) and its reconstructed counterpart (bottom) using three different fusion methods: Principal Component Analysis (PCA), Discrete Wavelet Transform (DWT), and Spectral Analysis via Fast Fourier Transform (SA-FFT). These transformations fuse data from different modalities, showing the effects of each method on the visual characteristics of the reconstructed images.

(A) Original Images



(B) Cheng2020 attn Fused Reconstruction

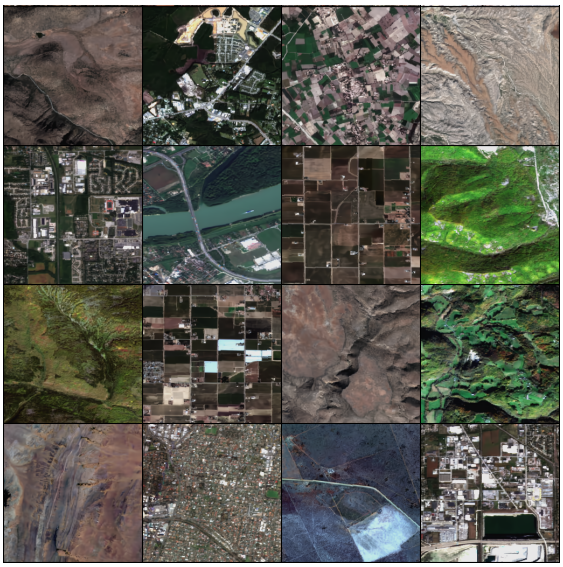


FIGURE 4.4: Comparison of satellite images. The top is the original optical satellite image, while the bottom image (*Cheng Attention 6 Fused*) is the result of fusing SAR and optical data in the latent space using the Cheng2020 VAE at quality level 6. The fused representation closely resembles the original optical image, demonstrating effective reconstruction through data fusion.

The neural compression models consistently outperform PCA, DWT, and SA in all key quality metrics. For example, at Quality 2, the fused *bmshj2018_hyperprior* model achieves a low relative bias and a high relative variance, indicating a close match to the reference image as shown in Table 4.4. In contrast, PCA and SA exhibit higher relative bias and variance, suggesting that these traditional methods introduce more distortion during the compression and fusion process.

With respect to Correlation Coefficient and UIQI the neural models also demonstrate superior performance. The fused neural compression models achieve high correlation coefficients, even at lower quality settings, reflecting strong preservation of structural features in the compressed data. On the other hand, PCA and SA struggle in this area, with correlation coefficients near zero or negative, reflecting poor structural similarity with the original data.

On the same line fused models consistently achieve high SSIM values and PSNR values that increase with quality level. In contrast, traditional methods such as PCA and SA result in significantly lower SSIM and PSNR values, indicating poorer image reconstruction quality and higher loss of critical information. The Spectral Angle Mapper (SAM) values provide additional insights into spectral fidelity. Lower SAM values indicate better spectral reconstruction and the fused neural models perform remarkably well. In comparison, PCA and SA yield much higher SAM values signaling greater spectral distortion as shown in Table 4.4.

4.4.2.2 Quality Metrics without Reference Image

Table 4.5 present quality metrics that do not rely on a reference image, instead focusing on information content and statistical properties such as Spectral Frequency, Standard Deviation, Entropy, and BPP. This provides a broader perspective on the efficiency and effectiveness of the compression models in terms of retaining information content.

The results show that the fused neural compression models maintain similar spectral frequencies and standard deviations to the individual modalities, suggesting that the fusion process does not degrade the variability or frequency content of the data. For instance, at Quality 2, the fused *bmshj2018_hyperprior* model achieves a spectral frequency of 12.3165 and a standard deviation of 0.2124, which are comparable to the image-only modality. This suggests that the fusion process efficiently integrates information from both SAR and optical data without introducing significant distortions.

Model	Type	Quality	Relative Bias	Relative Variance	Relative Standard Deviation	Correlation Coefficient	Universal Image Quality Index	SSIM	PSNR	Spectral Angle Mapper
bmshj2018 hyperprior	Image	2	0.05023	0.9568	0.9781	0.9778	0.9776	0.9693	26.89	0.1301
bmshj2018 hyperprior	SAR	2	2.2061	0.9236	0.8554	-0.094	-0.0867	-0.0749	8.9413	0.551
bmshj2018 hyperprior	Fused	2	0.0502	0.9568	0.9782	0.9779	0.9776	0.9695	26.89	0.1301
bmshj2018 hyperprior	Image	5	0.0233	0.9834	0.9917	0.9914	0.9914	0.9883	30.9976	0.0828
bmshj2018 hyperprior	SAR	5	2.2079	0.8928	0.9436	-0.0942	-0.0868	-0.0732	8.877	0.5569
bmshj2018 hyperprior	Fused	5	0.0233	0.9834	0.9917	0.9914	0.9914	0.9882	30.9958	0.0828
bmshj2018 hyperprior	Image	8	0.0099	0.994	0.997	0.9966	0.9966	0.9954	35.0437	0.0527
bmshj2018 hyperprior	SAR	8	2.2075	0.9048	0.9499	-0.093	-0.0857	-0.0741	8.8618	0.5569
bmshj2018 hyperprior	Fused	8	0.0099	0.994	0.997	0.9966	0.9966	0.9954	35.0448	0.0527
bmshj2018 factorized	Image	2	0.0657	0.9327	0.9658	0.9673	0.9667	0.9549	25.2336	0.1506
bmshj2018 factorized	SAR	2	2.2044	0.8068	0.8968	-0.0957	-0.0879	-0.074	9.0301	0.5429
bmshj2018 factorized	Fused	2	0.0657	0.9328	0.9658	0.9673	0.9667	0.955	25.2367	0.1506
bmshj2018 factorized	Image	5	0.0291	0.9749	0.9873	0.9878	0.9877	0.9832	29.4689	0.0955
bmshj2018 factorized	SAR	5	2.2062	0.8794	0.9364	-0.0928	-0.0856	-0.0737	8.9055	0.5546
bmshj2018 factorized	Fused	5	0.0291	0.9749	0.9874	0.9878	0.9877	0.9834	29.469	0.0955
bmshj2018 factorized	Image	8	0.011	0.9921	0.9961	0.9959	0.9959	0.9944	34.23	0.0572
bmshj2018 factorized	SAR	8	2.206	0.9039	0.9494	-0.0936	-0.0863	-0.0756	8.8651	0.5585
bmshj2018 factorized	Fused	8	0.011	0.9921	0.9961	0.9959	0.9959	0.9944	34.2322	0.0572
mbt2018	Image	2	0.0493	0.9581	0.9788	0.9792	0.9789	0.9712	27.156	0.127
mbt2018	SAR	2	2.2061	0.857	0.9244	-0.0947	-0.0871	-0.0747	8.9398	0.5511
mbt2018	Fused	2	0.0493	0.9581	0.9788	0.9792	0.9789	0.9711	27.1558	0.127
mbt2018	Image	5	0.0214	0.9839	0.9919	0.9927	0.9926	0.9899	31.666	0.0783
mbt2018	SAR	5	2.2045	0.8956	0.9451	-0.0935	-0.0861	-0.075	8.8806	0.5575
mbt2018	Fused	5	0.0214	0.9839	0.9919	0.9927	0.9926	0.9898	31.6685	0.0782
mbt2018	Image	8	0.0096	0.994	0.997	0.9967	0.9967	0.9954	35.0855	0.0526
mbt2018	SAR	8	2.207	0.9077	0.9513	-0.0925	-0.0857	-0.0748	8.8601	0.5588
mbt2018	Fused	8	0.0096	0.994	0.997	0.9967	0.9967	0.9954	35.085	0.0526
mbt2018 mean	Image	2	0.0503	0.9561	0.9778	0.9783	0.9781	0.9701	26.9901	0.1291
mbt2018	SAR	2	2.2056	0.8537	0.9226	-0.0954	-0.0878	-0.0745	8.9442	0.5507
mbt2018 mean	Fused	2	0.0503	0.9562	0.9778	0.9784	0.9781	0.97	26.9879	0.1291
mbt2018 mean	Image	5	0.0206	0.9842	0.9921	0.9925	0.9925	0.9896	31.5712	0.079
mbt2018 mean	SAR	5	2.203	0.8965	0.9454	-0.0921	-0.085	-0.0741	8.8856	0.5576
mbt2018 mean	Fused	5	0.0206	0.9842	0.9921	0.9925	0.9925	0.9897	31.5724	0.079
mbt2018 mean	Image	8	0.0096	0.9939	0.9969	0.9967	0.9967	0.9955	35.1463	0.0524
mbt2018 mean	SAR	8	2.2246	0.9081	0.9516	-0.0944	-0.0868	-0.0733	8.8491	0.5592
mbt2018 mean	Fused	8	0.0096	0.9939	0.9969	0.9967	0.9967	0.9955	35.1456	0.0524
cheng2020 anchor	Image	2	0.0479	0.9607	0.9802	0.98	0.9798	0.9721	27.337	0.1248
cheng2020 anchor	SAR	2	2.2104	0.8566	0.9242	-0.0942	-0.0866	-0.0735	8.9346	0.5508
cheng2020 anchor	Fused	2	0.0479	0.9607	0.9802	0.98	0.9798	0.9723	27.3354	0.1248
cheng2020 anchor	Image	4	0.0291	0.9828	0.9914	0.9897	0.9897	0.9857	30.199	0.0917
cheng2020 anchor	SAR	4	2.2104	0.8863	0.9402	-0.0941	-0.0866	-0.075	8.8835	0.5555
cheng2020 anchor	Fused	4	0.0291	0.9828	0.9913	0.9897	0.9897	0.9857	30.1996	0.0917
cheng2020 anchor	Image	6	0.0164	0.9899	0.9949	0.9938	0.9938	0.9914	32.377	0.0711
cheng2020 anchor	SAR	6	2.2072	0.8978	0.9463	-0.0918	-0.0848	-0.073	8.874	0.5575
cheng2020 anchor	Fused	6	0.0164	0.9899	0.9949	0.9938	0.9938	0.9914	32.3788	0.0711
cheng2020 attn	Image	2	0.0484	0.9584	0.979	0.9795	0.9793	0.9715	27.231	0.1264
cheng2020 attn	SAR	2	2.2061	0.854	0.9229	-0.0942	-0.0867	-0.0739	8.9412	0.5509
cheng2020 attn	Fused	2	0.0484	0.9584	0.979	0.9795	0.9793	0.9715	27.2311	0.1264
cheng2020 attn	Image	4	0.0236	0.9796	0.9898	0.9896	0.9896	0.9856	30.1397	0.0919
cheng2020 attn	SAR	4	2.2013	0.8879	0.9408	-0.0936	-0.0862	-0.072	8.9016	0.5563
cheng2020 attn	Fused	4	0.0236	0.9797	0.9898	0.9897	0.9896	0.9856	30.169	0.0919
cheng2020 attn	Image	6	0.0155	0.9897	0.9948	0.9937	0.9937	0.9913	32.33	0.0715
cheng2020 attn	SAR	6	2.2013	0.8879	0.9408	-0.0936	-0.0862	-0.072	8.90	0.5563
cheng2020 attn	Fused	6	0.0155	0.9896	0.9948	0.9937	0.9937	0.9913	32.32	0.0715
PCA	Image		0	0.9913	0.9956	0.5607	-0.001	-0.0176	8.0261	1.4527
PCA	SAR		0	0.9497	0.9735	0.2956	-0.0005	-0.005	7.4763	1.509
PCA	Fused		0	1.3018	1.1204	0.3356	-0.0006	-0.0026	7.2987	1.5034
DWT	Image		0	1	1	0.5607	1	1	133.196	0.0002
DWT	SAR		0	0.9546	0.9761	0.4557	0.4649	-0.005	71.0285	0.2795
DWT	Fused		0	1.2153	1.0894	0.5355	0.4521	-0.0026	49.0349	0.2998
SA	Image		0	1.4952	1.2215	-0.0088	-0.0075	0.0027	7.269	0.579
SA	SAR		0	1.4809	1.2155	-0.006	-0.0051	0.0042	7.2068	0.5798
SA	Fused		0	1.43	1.194	-0.0062	-0.0052	0.0044	7.1488	0.5723

TABLE 4.4: Quality Metrics with Reference Image: Performance comparison of neural compression and other models across different quality levels and data modalities (Sentinel-2 optical, Sentinel-1 SAR, and fused). The table presents results for various quality metrics with the reference image being the original optical image. Each model is evaluated on individual modalities (Image and SAR) and their fusion (Fused). PCA, DWT and SA transforms are fully inverted to collect metric data before and after fusion on a modality basis.

Entropy, which measures the richness of information in the compressed data, remains high for the fused neural models. At Quality 2, the fused *bmshj2018_hyperprior* model retains an entropy value of 22.35, close to that of the individual modalities, indicating minimal information loss. By comparison, traditional methods such as PCA and SA demonstrate lower entropy values, indicating the potential loss of important data during transformation.

Compression efficiency is another key strength of the neural compression models. The fused representations consistently achieve lower BPP values compared to PCA, DWT, and SA, demonstrating more efficient compression. For example, the fused *bmshj2018_hyperprior* model at Quality 2 has a BPP of 0.5276, whereas PCA, DWT, and SA maintain fixed high BPP values of 96, reflecting their less efficient compression techniques.

4.4.2.3 Comparison with Traditional Methods

The results clearly demonstrate that neural compression models significantly outperform traditional methods such as PCA, DWT, and SA across all evaluated metrics. Structurally, the neural models retain far more information from the original image, with high SSIM and PSNR values confirming superior reconstruction quality. In contrast, PCA and SA suffer from poor structural and spectral fidelity, as evidenced by lower SSIM and PSNR values and higher SAM values.

Neural compression models also excel in compression efficiency. The lower BPP values achieved by neural models, coupled with high entropy, demonstrate that they provide more efficient data representation. Traditional methods, lacking the advanced encoding capabilities of neural networks, result in much higher BPP values, reflecting less efficient compression and greater storage or transmission requirements.

Furthermore, the information content in the neural compression models is better preserved, as reflected in the high entropy values. This indicates that neural compression retains more of the essential details and variability present in the original data. In comparison, PCA and SA exhibit lower entropy, which points to the potential loss of important data during the transformation and fusion process. The ability to maintain higher entropy suggests that neural compression models are more adept at capturing the full complexity of the data, making them important in applications where data richness and detail are critical. This preservation of information is crucial for downstream tasks, such as classification and anomaly detection, where small details can significantly impact performance.

Model	Type	Quality	Spectral Frequency	Standard D	Entropy	BPP
bmshj2018_hyperprior	Image	2	12.3	0.2124	22.35	0.5833
bmshj2018_hyperprior	SAR	2	16.1214	0.2004	22.159	0.5323
bmshj2018_hyperprior	Fused	2	12.3165	0.2124	22.35	0.5276
bmshj2018_hyperprior	Image	5	14.5822	0.2154	22.33	0.3936
bmshj2018_hyperprior	SAR	5	18.6838	0.2047	22.177	0.3239
bmshj2018_hyperprior	Fused	5	14.5816	0.2154	22.3336	0.323
bmshj2018_hyperprior	Image	8	15.8567	0.2166	22.3596	0.4685
bmshj2018_hyperprior	SAR	8	19.9865	0.206	22.2158	0.4927
bmshj2018_hyperprior	Fused	8	15.8265	0.2164	22.3592	0.481
bmshj2018_factorized	Image	2	10.7259	0.2098	22.3425	0.6767
bmshj2018_factorized	SAR	2	12.9859	0.1942	22.1659	0.6604
bmshj2018_factorized	Fused	2	10.7259	0.2098	22.3425	0.66
bmshj2018_factorized	Image	5	13.6945	0.2144	22.3437	0.5218
bmshj2018_factorized	SAR	5	17.4632	0.2029	22.168	0.4801
bmshj2018_factorized	Fused	5	13.6944	0.2145	22.3437	0.4764
bmshj2018_factorized	Image	8	15.6675	0.2162	22.3355	0.6334
bmshj2018_factorized	SAR	8	19.8376	0.2058	22.1842	0.5976
bmshj2018_factorized	Fused	8	15.6678	0.2164	22.3357	0.5888
mbt2018	Image	2	12.5095	0.2126	22.3393	0.5814
mbt2018	SAR	2	16.305	0.2004	22.1881	0.5236
mbt2018	Fused	2	12.5102	0.2126	22.3395	0.5194
mbt2018	Image	5	14.9917	0.2154	22.3633	0.8077
mbt2018	SAR	5	19.1415	0.205	22.2093	0.7138
mbt2018	Fused	5	14.9918	0.2154	22.3634	0.7098
mbt2018	Image	8	15.8645	0.2165	22.3606	0.4542
mbt2018	SAR	8	20.0404	0.2061	22.2085	0.4321
mbt2018	Fused	8	15.8643	0.2165	22.3605	0.4286
mbt2018_mean	Image	2	12.3655	0.2124	22.3475	0.5842
mbt2018_mean	SAR	2	16.1873	0.2001	22.1897	0.5316
mbt2018_mean	Fused	2	12.3661	0.2125	22.3476	0.5264
mbt2018_mean	Image	5	14.9739	0.2155	22.3494	0.8267
mbt2018_mean	SAR	5	19.1709	0.205	22.2025	0.7443
mbt2018_mean	Fused	5	14.9742	0.2155	22.3495	0.7472
mbt2018_mean	Image	8	15.8444	0.2164	22.3524	0.4742
mbt2018_mean	SAR	8	19.9684	0.206	22.1999	0.4713
mbt2018_mean	Fused	8	15.8444	0.2161	22.3519	0.4702
cheng2020_anchor	Image	2	12.5933	0.213	22.3401	0.3549
cheng2020_anchor	SAR	2	15.7871	0.2003	22.1624	0.2912
cheng2020_anchor	Fused	2	12.5932	0.213	22.3401	0.2954
cheng2020_anchor	Image	4	14.38	0.2154	22.3433	0.4716
cheng2020_anchor	SAR	4	18.0338	0.204	22.174	0.3795
cheng2020_anchor	Fused	4	14.3801	0.2152	22.3431	0.366
cheng2020_anchor	Image	6	15.1874	0.2161	22.3429	0.3224
cheng2020_anchor	SAR	6	19.1489	0.2051	22.1718	0.2212
cheng2020_anchor	Fused	6	15.1876	0.2161	22.3428	0.2176
cheng2020_attn	Image	2	12.451	0.2127	22.3488	0.3584
cheng2020_attn	SAR	2	15.5836	0.2003	22.1874	0.2958
cheng2020_attn	Fused	2	12.4501	0.2127	22.3487	0.2942
cheng2020_attn	Image	4	14.4078	0.215	22.3475	0.4642
cheng2020_attn	SAR	4	18.0871	0.204	22.1874	0.3766
cheng2020_attn	Fused	4	14.408	0.2151	22.3475	0.3774
cheng2020_attn	Image	6	15.1626	0.2161	22.3458	0.3227
cheng2020_attn	SAR	6	18.0871	0.204	22.1874	0.3766
cheng2020_attn	Fused	6	15.1625	0.2159	22.3457	0.2302
PCA	Image		12.715	0.2158	20.0907	96
PCA	SAR		12.0723	0.2108	14.6625	96
PCA	Fused		14.2561	0.2426	16.8419	96
DWT	Image		16.65	0.2167	11.4294	96
DWT	SAR		18.4464	0.2112	11.3368	96
DWT	Fused		21.1773	0.2358	11.4065	96
SA	Image		21.4051	0.2639	23.6063	96
SA	SAR		20.6694	0.2626	25.3134	96
SA	Fused		20.9059	0.2579	24.5498	96

TABLE 4.5: Quality Metrics without Reference Image: Performance comparison of neural compression and other models across different quality levels and data modalities (Sentinel-2 optical, Sentinel-1 SAR, and fused). The table presents results for various quality metrics without a reference image focusing on information content. Each model is evaluated on individual modalities (Image and SAR) and their fusion (Fused). PCA, DWT and SA transforms are fully inverted to collect metric data before and after fusion on a modality basis.

4.4.2.4 Comparison of Latent Representations: UMAP vs. t-SNE Visualizations

In addition to the classification performance metrics, the t-SNE and UMAP visualizations provide further insight into the structure and separability of the latent spaces learned by the neural compression models. These visualizations, presented in Figures 5 through 8, illustrate how the models project the optical, SAR, and fused data into lower-dimensional spaces. The visual representations help us understand how well the neural networks separate the data points corresponding to different modalities (optical, SAR, and fused) and classes, and how effective the fusion process is at combining complementary information from the two modalities.

The UMAP visualizations in Figure 5 represent the latent space distribution for optical, SAR, and fused modalities. In these visualizations, the three distinct clusters correspond to the three modalities—optical, SAR, and fused—rather than class separability within each modality. This reflects how each modality occupies its own region in the latent space, with the fused representation forming its own distinct cluster, separate from the singular optical and SAR modalities.

For example, in the *bmshj2018_hyperprior* model at Quality 2 (Figure 5a), the optical, SAR, and fused modalities are clearly separated. The distinct clustering of the fused representation suggests that it captures features that combine the strengths of both SAR and optical modalities, resulting in a richer and more informative latent space. As the quality of compression increases (Figures 5b and 5c), the separation between these modalities remains clear, indicating that the neural compression models maintain consistent distinctions between the different data sources, even at higher compression qualities.

The ability of UMAP to maintain clear distinctions between the optical, SAR, and fused modalities highlights the model's capacity to encode the unique characteristics of each modality. While the optical and SAR modalities form distinct clusters, the fused representation benefits from integrating complementary information from both modalities, leading to more robust and distinct features that are advantageous for downstream tasks like classification.

The t-SNE visualizations offer a different perspective, showing how the data points within each modality are distributed in the latent space. Figure 6 presents t-SNE projections for the optical latent spaces, Figure 7 shows the SAR latent spaces, and Figure 8 illustrates the fused latent spaces. These visualizations allow us to analyze the internal

structure of each modality's latent space and assess how well the data points within each modality are clustered by class.

In Figure 6, the optical latent spaces across different models and quality levels show moderate clustering by class, though some overlap between class boundaries is present. The SAR latent spaces in Figure 7, however, exhibit more dispersion and less defined clustering, reflecting the greater challenge of using SAR data alone for tasks like classification, where spectral details (lacking in SAR data) are important.

The t-SNE visualizations in Figure 8, which represent the fused latent spaces, show significantly better clustering by class compared to the singular modalities. For example, in the *bmshj2018_hyperprior* model at Quality 2 (Figure 8a), the fused representation results in tighter clusters, reflecting the model's improved ability to differentiate between classes. As the compression quality improves to Quality 8 (Figure 8c), the class clusters become even more distinct, further demonstrating how the fusion of SAR and optical data leads to better separation of classes within the latent space. Similar trends are observed in other models such as *bmshj2018_factorized* and *mbt2018*, where fused latent spaces consistently demonstrate better class separability than the individual optical or SAR modalities.

Both UMAP and t-SNE visualizations provide complementary insights into the learned latent spaces. UMAP is particularly effective at showing the overall separation between the modalities (optical, SAR, and fused), emphasizing how neural compression models maintain distinctions between the data sources while encoding each modality in its own subspace. On the other hand, t-SNE focuses more on class separability within each modality, showing how fused representations lead to more coherent and distinct clusters by class. This suggests that the fusion of modalities improves the model's ability to capture discriminative features, ultimately leading to enhanced classification performance.

The consistency between the UMAP and t-SNE visualizations highlights the strength of neural compression models in handling multimodal data. By effectively separating the optical, SAR, and fused modalities, and improving class separability in the fused latent spaces, these models demonstrate their capacity to integrate complementary information from different sources, resulting in better feature representation and overall classification performance.

4.4.2.5 Analysis of Individual Neural Models

Each neural compression model analyzed exhibits unique strengths in terms of compression efficiency and quality preservation. The *bmshj2018_hyperprior* model excels at balancing compression efficiency with reconstruction quality, achieving high SSIM and PSNR values across different quality levels. The *bmshj2018_factorized* model performs similarly, especially at higher quality settings, where it maintains high levels of structural similarity and information content.

The *mbt2018* and *mbt2018_mean* models offer strong overall performance, particularly in entropy retention, highlighting their capability to preserve essential information during compression. Additionally, the *cheng2020_anchor* and *cheng2020_attn* models show competitive results, achieving lower BPP values while maintaining quality, showcasing their efficiency in data compression without sacrificing too much on the quality front.

The comparative analysis highlights the superiority of neural compression models over traditional data fusion methods such as PCA, DWT, and SA. Neural models not only deliver enhanced classification performance but also provide superior image quality and compression efficiency. The fusion of SAR and optical data in these models leverages the strengths of both modalities without degrading reconstruction quality, making neural compression models the preferred choice for remote sensing data fusion tasks. Traditional methods, while useful, fall short in terms of structural and spectral fidelity, compression efficiency, and overall quality, further underscoring the advantages of neural-based approaches.

4.5 Discussion

The findings from this study highlight the significant advantages of neural compression models, especially when applied to multimodal data fusion involving Synthetic Aperture Radar (SAR) and optical imagery. By utilizing fused representations, these models consistently outperform traditional compression and fusion techniques, such as PCA, DWT, and SA, across a variety of image quality, compression efficiency, and classification accuracy metrics.

A key observation is the ability of neural compression models to capture complementary information from SAR and optical data, leading to enhanced classification performance. SAR data, with its ability to penetrate clouds and provide structural details,

complements the rich spectral information available from optical data. When fused into a single latent representation, neural models can extract more discriminative features, improving classification accuracy without the need for explicit decompression. This approach contrasts with traditional methods, where each modality is processed separately, often leading to suboptimal feature extraction and fusion strategies.

The analysis of image quality metrics, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Spectral Angle Mapper (SAM), further demonstrates the superiority of neural compression models in preserving both structural and spectral fidelity. Despite the compression process, the fused neural models were able to maintain high PSNR and SSIM values, indicating that the fused representations do not compromise reconstruction quality. In fact, these models often outperformed singular modalities in terms of overall image quality, suggesting that the fusion process itself contributes positively to the preservation of key features.

Moreover, the efficiency of neural compression models is evident in the significant reduction in BPP compared to traditional methods, without sacrificing classification accuracy or image quality. Traditional compression methods, such as JPEG, require higher BPP to achieve similar levels of quality, making neural compression models a more efficient solution for remote sensing applications where data transmission and storage are key considerations. The Rate-Distortion Accuracy Index (RDAI) metric also reinforces the ability of neural models to balance compression efficiency with classification performance, achieving superior trade-offs compared to traditional techniques.

Additionally, the use of neural compression models for data fusion opens up new possibilities for remote sensing applications beyond classification. The fusion of SAR and optical data has potential utility in other areas such as object detection, change detection, and anomaly detection, where the combination of structural and spectral information can provide a more holistic understanding of the scene. Future work should explore how neural compression models can be adapted or extended to these tasks, as the findings presented here suggest promising opportunities for improved performance in a wider range of remote sensing applications.

4.6 Limitations

Despite the promising results achieved by the neural compression models in this study, several limitations need to be addressed for a more comprehensive understanding of their capabilities and applicability.

First, the computational complexity of neural compression models, especially during training, is significantly higher compared to traditional methods such as PCA, DWT, and SA. Neural networks require considerable computational resources, especially when processing large-scale datasets and high-resolution satellite images like those from Sentinel-1 and Sentinel-2. Training these models involves optimizing numerous parameters across multiple layers, making it both time- and resource-intensive. This complexity limits the accessibility of these models for applications where computational power is constrained or where real-time processing is required. In contrast, traditional methods like PCA and DWT are more computationally efficient and easier to implement, though they offer less robust performance. Future advancements in hardware acceleration, such as the use of specialized AI chips or cloud-based processing, could mitigate some of these computational challenges.

Another limitation is the sensitivity of neural compression models to hyperparameters and training conditions. The performance of these models can vary significantly depending on the chosen architecture, learning rates, batch sizes, and quality settings. Small variations in these parameters can lead to noticeable differences in classification accuracy and image quality, making it difficult to generalize results across different datasets or sensor modalities without careful tuning. This lack of consistency underscores the need for more standardized training and evaluation protocols that can ensure reliable and reproducible results. Moreover, neural models can sometimes exhibit overfitting, particularly when the dataset is not sufficiently large or diverse, which can reduce their ability to generalize to unseen data.

Additionally, while the fusion of SAR and optical data has demonstrated significant benefits in terms of classification accuracy and image quality, multimodal fusion may not always be suitable for all remote sensing applications. Certain environmental or meteorological conditions may favor one modality over the other, making the fusion process less effective or even redundant. For instance, in regions with limited cloud cover, optical data may already provide sufficient information, and the inclusion of SAR data might not yield substantial improvements. Moreover, fusing multiple modalities could potentially introduce irrelevant or conflicting information, leading to decreased performance in some cases. Thus, further research is needed to better understand the specific conditions under which multimodal fusion provides the greatest benefit and when single-modality models might be more appropriate.

A key limitation is also the need for large datasets to train neural compression models effectively. Neural networks, particularly in the context of multimodal data fusion, rely

heavily on extensive, high-quality datasets to capture the nuances of the data and extract meaningful patterns. In cases where data is scarce or the available dataset is too small, the models may fail to generalize well, and the fused representations might artificially inflate performance compared to singular modalities. This limitation underscores the importance of large-scale, well-labeled datasets in achieving the full potential of neural compression models, especially in scenarios where data collection is expensive or difficult.

Furthermore, although the reconstruction quality of fused representations was shown to be on par with or even superior to singular modalities, the impact of fusion on other downstream tasks, such as object detection, time-series analysis, or change detection—has not been fully explored. The evaluation in this study focused primarily on image quality metrics and classification performance. However, remote sensing applications often require more specialized analyses, and it is unclear whether the benefits of fusion seen here will translate to those tasks. For example, in time-series analysis, maintaining temporal coherence across multiple observations is critical, and fusion might introduce inconsistencies that could affect performance. Similarly, object detection and change detection could be impacted by how well the fused representations retain spatial and spectral features over time.

4.7 Conclusion

In this paper, we presented a comprehensive evaluation of neural compression models using fused representations of Sentinel-1 SAR and Sentinel-2 optical data. The results demonstrate that neural compression models significantly outperform traditional data fusion techniques such as PCA, DWT, and SA across a range of quality metrics. The fused representations consistently yielded higher classification accuracy, better compression efficiency, and superior image quality compared to singular modalities.

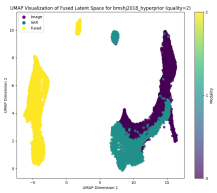
One of the key findings is that the fusion of SAR and optical data enhances both structural and spectral fidelity, without negatively impacting reconstruction quality. The complementary information provided by the two modalities allows neural models to capture more discriminative features, leading to substantial improvements in both compression and classification performance. Furthermore, the fused models maintained low bits-per-pixel (BPP) rates while preserving essential image details, making them highly efficient in terms of data storage and transmission.

However, the study also highlighted several limitations of neural compression models, particularly in terms of computational demands and sensitivity to hyperparameters.

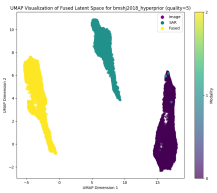
While these models offer substantial improvements over traditional methods, their complexity may limit their use in applications with real-time constraints or limited computational resources.

Overall, neural compression models represent a significant step forward in multimodal data fusion for remote sensing applications. By leveraging the strengths of both SAR and optical data, these models offer a powerful solution for improving remote sensing data processing quality and efficiency. Future research should address the computational limitations, explore the impact of fusion on a wider range of downstream tasks, and refine the models for broader practical applications.

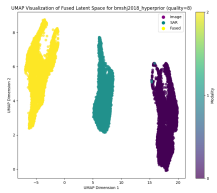
4.8 Appendix: Latent Space Visualizations



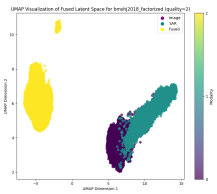
(A) bmshj2018_hyperprior 2



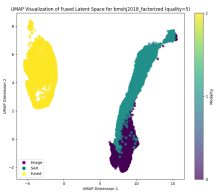
(B) bmshj2018_hyperprior 5



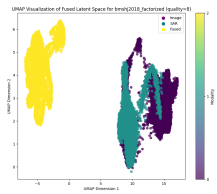
(C) bmshj2018_hyperprior 8



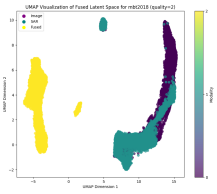
(D) bmshj2018_factorized 2



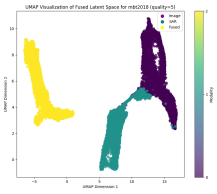
(E) bmshj2018_factorized 4



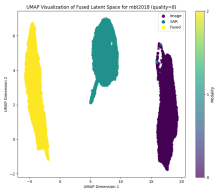
(F) bmshj2018_factorized 8



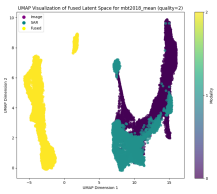
(G) mbt2018 2



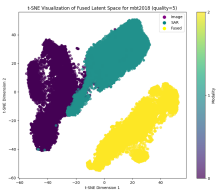
(H) mbt2018 5



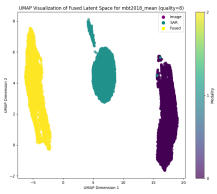
(I) mbt2018 8



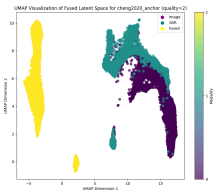
(J) mbt2018_mean 2



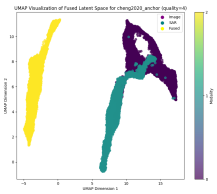
(K) mbt2018_mean 5



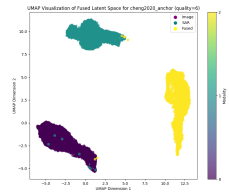
(L) mbt2018_mean 8



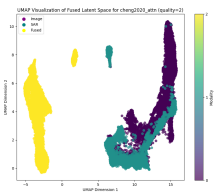
(M) cheng2020_anchor 2



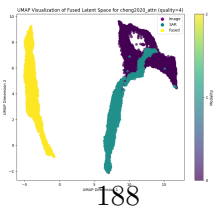
(N) cheng2020_anchor 4



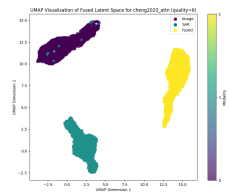
(O) cheng2020_anchor 6



(P) cheng2020_attn 2

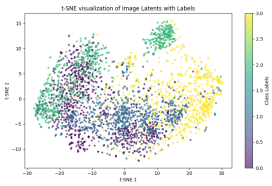


(Q) cheng2020_attn 4

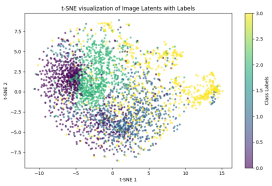


(R) cheng2020_attn 6

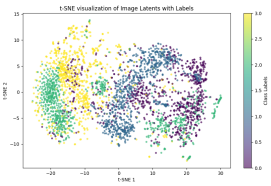
FIGURE 4.5: UMAP visualizations of models constructed optical, SAR and Fused latent spaces, with labels.



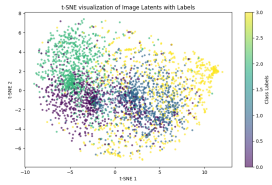
(A) bmshj2018_hyperprior 2



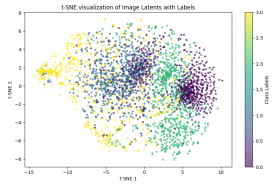
(B) bmshj2018_hyperprior 5



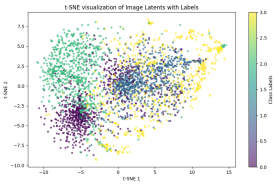
(C) bmshj2018_hyperprior 8



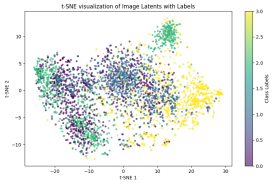
(D) bmshj2018_factorized 2



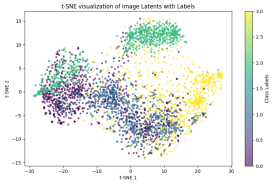
(E) bmshj2018_factorized 4



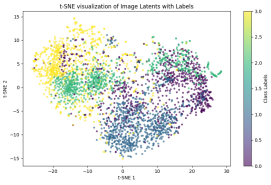
(F) bmshj2018_factorized 8



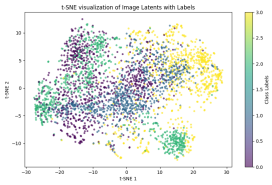
(G) mbt2018 2



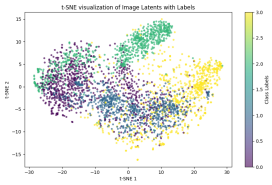
(H) mbt2018 5



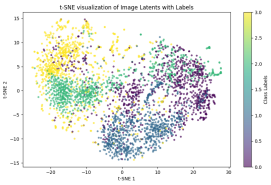
(I) mbt2018 8



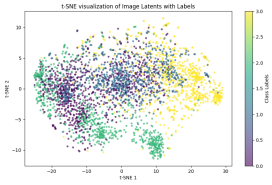
(J) mbt2018_mean 2



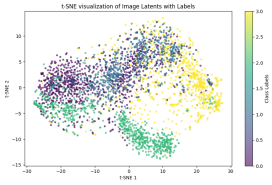
(K) mbt2018_mean 5



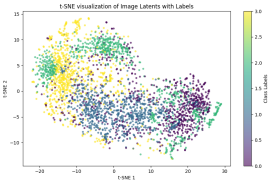
(L) mbt2018_mean 8



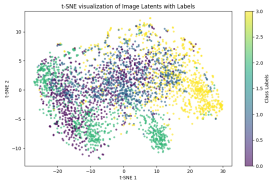
(M) cheng2020_anchor 2



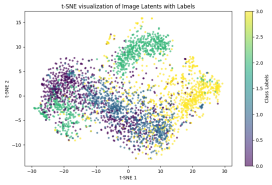
(N) cheng2020_anchor 4



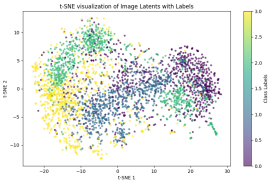
(O) cheng2020_anchor 6



(P) cheng2020_attn 2

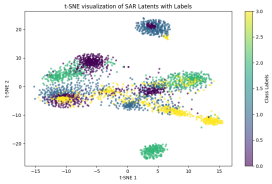


(Q) cheng2020_attn 4

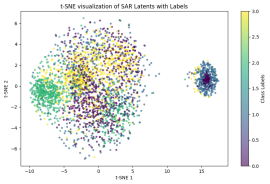


(R) cheng2020_attn 6

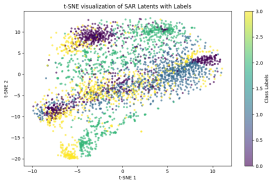
FIGURE 4.6: t-SNE visualizations of models constructed optical latent spaces, with labels.



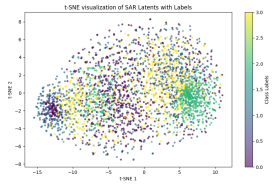
(A) bmshj2018_hyperprior 2



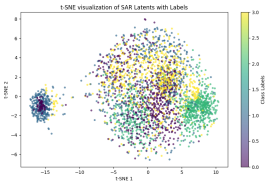
(B) bmshj2018_hyperprior 5



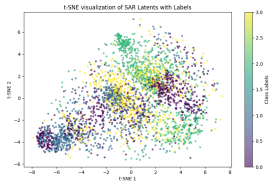
(C) bmshj2018_hyperprior 8



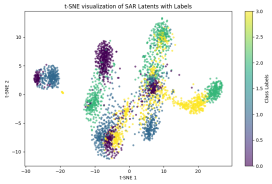
(D) bmshj2018_factorized 2



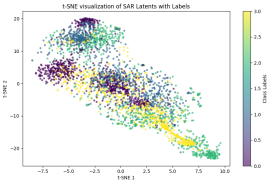
(E) bmshj2018_factorized 4



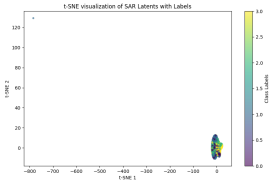
(F) bmshj2018_factorized 8



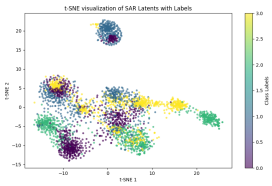
(G) mbt2018 2



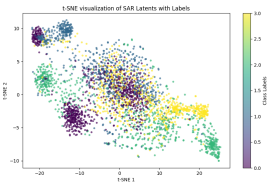
(H) mbt2018 5



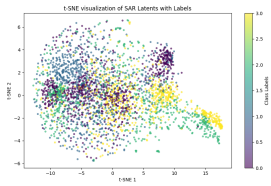
(I) mbt2018 8



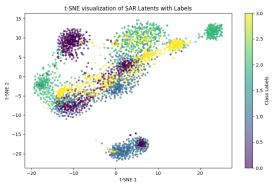
(J) mbt2018_mean 2



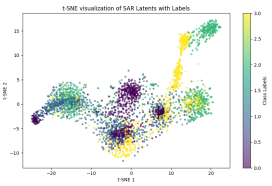
(K) mbt2018_mean 5



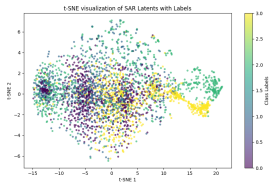
(L) mbt2018_mean 8



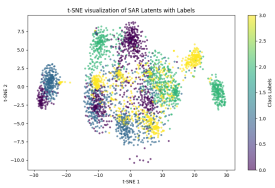
(M) cheng2020_anchor 2



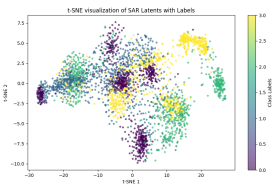
(N) cheng2020_anchor 4



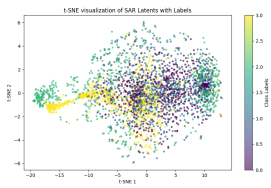
(O) cheng2020_anchor 6



(P) cheng2020_attn 2

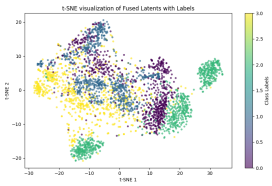


(Q) cheng2020_attn 4

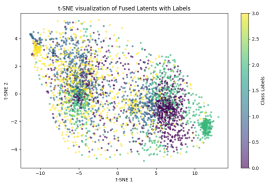


(R) cheng2020_attn 6

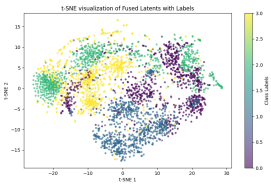
FIGURE 4.7: t-SNE visualizations of models constructed SAR latent spaces, with labels.



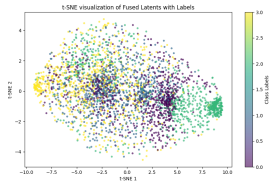
(A) bmshj2018_hyperprior 2



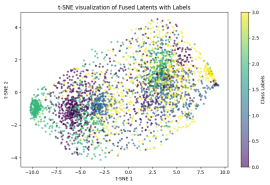
(B) bmshj2018_hyperprior 5



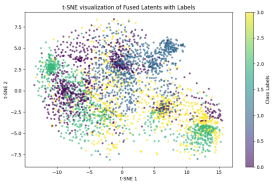
(C) bmshj2018_hyperprior 8



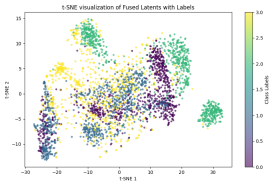
(D) bmshj2018_factorized 2



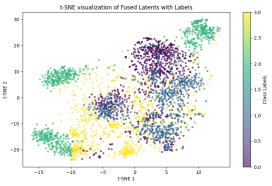
(E) bmshj2018_factorized 4



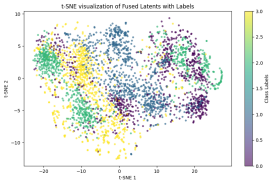
(F) bmshj2018_factorized 8



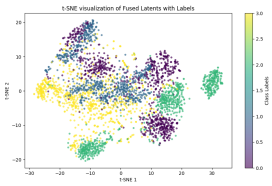
(G) mbt2018 2



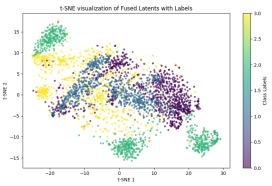
(H) mbt2018 5



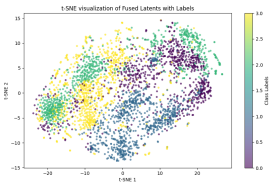
(I) mbt2018 8



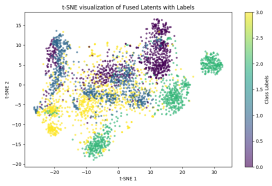
(J) mbt2018_mean 2



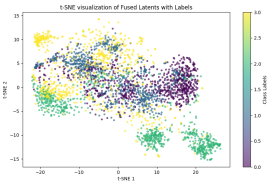
(K) mbt2018_mean 5



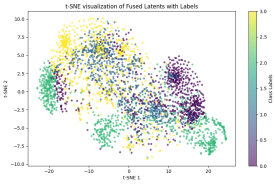
(L) mbt2018_mean 8



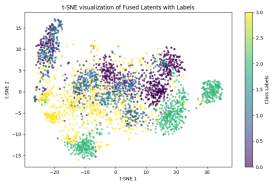
(M) cheng2020_anchor 2



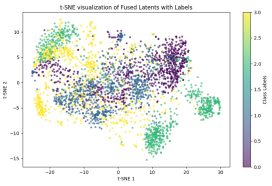
(N) cheng2020_anchor 4



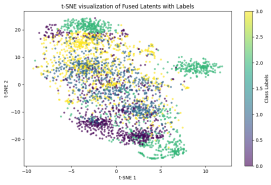
(O) cheng2020_anchor 6



(P) cheng2020_attn 2



(Q) cheng2020_attn 4



(R) cheng2020_attn 6

FIGURE 4.8: t-SNE visualizations of models constructed fused latent spaces, with labels.

Chapter 5

Using VAEs for Anomaly Detection

The content of this chapter is a second revision of the manuscript text for publication under the following citation:

Giuliano, A. (2024). Anomaly Detection of Under-Over Current in Magnetorheological Damper Suspension Using Variational Autoencoders. IEEE/ASME Transactions on Mechatronics.

Anomaly Detection of Under-Over Current in Magnetorheological Damper Suspension Using Variational Autoencoders

Alessandro Giuliano
Faculty of Engineering
McMaster University, Hamilton, ON, Canada
Email: giuliana@mcmaster.ca

Yuandi Wu
Faculty of Engineering
McMaster University, Hamilton, ON, Canada
S. Andrew Gadsden
Faculty of Engineering
McMaster University
Email: gadsdesa@mcmaster.ca

John Yawney
Adastra Corp.
Email: john.yawney@adastragrp.com

Abstract

Detecting anomalies in complex sensor-based systems, such as Magnetorheological (MR) dampers, is critical for ensuring operational reliability and safety. Traditional anomaly detection methods often struggle with the high-dimensional, noisy nature of MR damper data, particularly when anomalies are induced through variable voltage inputs. In this study, we propose a novel hybrid approach using a Variational Autoencoder (VAE) with an integrated Multilayer Perceptron (MLP) classifier to directly classify anomalies within the VAE's latent space. This end-to-end model jointly optimizes unsupervised representation learning and supervised anomaly classification, effectively balancing reconstruction, regularization, and classification objectives. We evaluate the model on

MR damper data collected from position and force sensors, with anomalies induced by over- and under-voltage scenarios. Experimental results demonstrate that our integrated VAE-MLP framework outperforms traditional methods, such as standalone PCA, Isolation Forest, and conventional Autoencoders, in both classification accuracy and robustness to noise. Additionally, we show that the inclusion of an MLP during training enhances the interpretability of the VAE's latent space, clustering anomalies distinctly from normal operational data. This approach not only achieves superior anomaly detection performance but also provides a promising framework for hybrid representation and classification in high-dimensional sensor data applications.

Keywords: *Variational Autoencoder, Anomaly Detection, Magnetorheological Dampers, Machine Learning, Latent Space Classification*

5.1 Introduction

Anomaly detection plays a pivotal role in ensuring the reliability and safety of critical engineering systems. The ability to detect deviations from expected behavior is essential in applications ranging from industrial automation to aerospace, where system failures can lead to catastrophic consequences. In recent years, there has been a growing interest in the use of machine learning techniques to improve anomaly detection capabilities, especially in systems where traditional methods struggle due to high-dimensional and complex data structures. One such system is the Magnetorheological (MR) damper, which is widely used in semi-active control applications, such as vibration control [313][314], automotive suspensions [315] [316], and structural damping [317, 318, 319] and more [320, 321]. Despite its effectiveness, the complexity of MR damper behavior under different operational conditions poses significant challenges for anomaly detection.

Magnetorheological dampers are devices that use a magnetorheological fluid (a suspension of micrometer-sized magnetic particles in a carrier liquid) that can change its rheological properties in response to an applied magnetic field [314]. This ability to vary the fluid viscosity in real time allows MR dampers to adapt their damping characteristics dynamically, making them ideal for applications requiring precise and responsive control over vibration. However, the non-linear and complex dynamics of MR dampers, coupled with the noise-prone data generated by position and force sensors, make it difficult to detect anomalies using conventional threshold-based methods. Anomalies in MR dampers can occur due to various factors, including voltage irregularities, mechanical wear, or degradation in the magnetorheological fluid properties [322]. Identifying these

anomalies is crucial for predictive maintenance and for preventing potential failures that could compromise system performance or safety.

Traditional anomaly detection methods, such as statistical process control, rule-based systems, and threshold monitoring, have been employed in monitoring MR dampers [323]. However, these approaches often rely on handcrafted features and require extensive domain knowledge to establish effective detection criteria. Furthermore, these methods can struggle with distinguishing between normal variations in damper behavior and actual anomalies, especially when anomalies are subtle or develop gradually over time. As a result, more sophisticated data-driven techniques are being explored, particularly in the realm of machine learning. Unsupervised learning models, such as Principal Component Analysis (PCA), Isolation Forests, and Autoencoders, have shown promise in detecting anomalies in high-dimensional sensor data by learning patterns of normal behavior and flagging deviations. However, these models still face limitations in capturing the intricate and multi-modal characteristics of MR damper data, where noise and dynamic changes add further complexity.

Variational Autoencoders (VAEs) have emerged as a powerful tool for anomaly detection in high-dimensional data. Unlike traditional autoencoders, which aim solely to reconstruct input data, VAEs impose a probabilistic structure on the latent space, allowing for a more compact and meaningful representation of data variability. This characteristic is particularly advantageous in anomaly detection, as the VAE's latent space can effectively capture the underlying distribution of normal data while flagging outliers that do not conform to this distribution. By training a VAE on sensor data from MR dampers under normal operating conditions, it is possible to learn a latent representation that encapsulates the typical dynamics of the system. When presented with anomalous data, the VAE should either produce a high reconstruction error or a latent representation that significantly deviates from the normal distribution, indicating an anomaly.

In this study, we propose a novel approach that integrates a Multilayer Perceptron (MLP) classifier directly within the VAE framework to enhance the model's anomaly detection capabilities. The integration of the MLP allows us to perform anomaly classification within the VAE's latent space, creating a hybrid model that benefits from both unsupervised representation learning and supervised classification. This end-to-end training approach, which simultaneously optimizes the VAE's reconstruction and regularization objectives alongside the MLP's classification objective, enables the model to learn a latent space tailored specifically for anomaly detection. By incorporating the

classification task during training, the latent space becomes more discriminative, clustering normal and anomalous samples distinctly and thereby facilitating more accurate anomaly detection.

Our approach is evaluated on sensor data collected from an MR damper system, where anomalies are induced through controlled over- and under-voltage scenarios. The MR damper data consists of time-series measurements from position and force sensors, capturing the damper's response to changes in input voltage. This dataset presents several challenges, including temporal dependencies, and significant noise, making it an ideal test case for our proposed method. We compare the performance of the VAE-MLP hybrid model against several baseline models, including traditional PCA, Isolation Forests, and standalone Autoencoders. Evaluation metrics include accuracy, F1 score, and area under the precision-recall curve (AU-PRC), as well as qualitative assessments of the latent space structure to examine the model's interpretability.

The contributions of this paper are threefold. First, we introduce an integrated VAE-MLP model that jointly optimizes unsupervised and supervised learning objectives, providing a robust framework for anomaly detection in complex systems like MR dampers. Second, we demonstrate that the inclusion of a classifier during training enhances the discriminative power of the latent space, improving anomaly detection performance and interpretability. Finally, we conduct extensive experiments on MR damper data, showing that our approach outperforms traditional anomaly detection methods in terms of both detection accuracy and robustness to noisy sensor data.

In summary, this study aims to advance the field of anomaly detection by presenting a novel VAE-based approach tailored to the challenges of high-dimensional and noisy sensor data in MR dampers. The proposed VAE-MLP hybrid model achieves superior detection performance and provides insights into the underlying data distribution, making it a valuable tool for predictive maintenance and fault diagnosis in engineering systems. The results of this study demonstrate the potential of hybrid unsupervised-supervised models for anomaly detection and open avenues for future research in applying methods similar to those of other sensor-driven applications.

5.2 Related Work

Chong et al. [322] present a first approach employing a nonlinear multiclass support vector machine (NMSVM) as the core classifier in a comprehensive SHM system for buildings equipped with MR dampers. This framework, integrating discrete wavelet

transforms and autoregressive (AR) models to extract damage-sensitive features, aims to improve detection precision under varying damage scenarios. Notably, this study represents the only research to date applying machine learning specifically to SHM in MR-damped buildings, advancing a robust solution capable of effectively distinguishing between multiple damage levels despite the challenges posed by ambient noise and structural complexity.

5.2.1 Anomaly Detection with VAEs

Anomaly detection is an essential task across diverse fields, including cybersecurity, industrial monitoring, and healthcare, due to its potential to identify critical events or irregularities in data. Conventional anomaly detection approaches rely heavily on statistical models, clustering methods, and prediction-based frameworks [324]. However, these traditional techniques often struggle with complex, high-dimensional data due to the lack of robust feature learning mechanisms.

Variational Autoencoders (VAEs) have emerged as a popular tool for anomaly detection because of their ability to learn low-dimensional latent representations that capture underlying data structures. Through probabilistic encoding, VAEs can model normal patterns, making it possible to detect anomalies as deviations from these patterns [325]. VAEs are particularly well-suited for applications in noisy environments, such as web monitoring systems and industrial sensor networks, where they are employed to reconstruct normal patterns while isolating anomalies through reconstruction error.

Recent advances in VAE-based anomaly detection have explored various hybrid architectures. For example, Lin et al. [326] proposed a VAE-LSTM model tailored for time series anomaly detection, where the VAE module learns local temporal features and the LSTM module captures long-term dependencies, effectively identifying both short-term and sustained anomalies. This hybrid approach improves upon the conventional VAE by addressing its limited capacity for long-term temporal modeling, a limitation frequently encountered in purely reconstruction-based models.

Other studies incorporate clustering methods to enhance the interpretability of VAE latent spaces. Zhu et al. [327] introduced a VAE-SOM (Self-Organizing Map) hybrid model for monitoring flexible sensors in wearable health devices. The SOM module clusters the VAE's latent outputs, translating continuous latent features into discrete states, which are subsequently analyzed for temporal dependencies using Markov chains. This method demonstrates that clustering within the latent space can improve anomaly

interpretability and enable the model to adapt to complex temporal patterns in real-time sensor data.

Further innovations in VAE architectures include attention-based mechanisms and graph neural networks (GNNs) to handle multivariate dependencies in time series data. Shi et al. [328] developed a GCN-VAE model for multivariate time series anomaly detection, where a GCN captures interdependencies between time-series variables, and an attention-based VAE reconstructs data while emphasizing the most informative features. This approach outperforms conventional VAEs in high-dimensional sensor networks by leveraging correlations across variables, thus providing a more context-aware anomaly detection framework.

Our work builds on these foundational VAE architectures by proposing a conditioned VAE-MLP model that introduces a classifier within the VAE framework. This integration explicitly conditions the latent space, fostering greater separation between normal and anomalous patterns. Compared to the aforementioned VAE-based models, the conditioned VAE-MLP optimizes the latent representation for anomaly separability, enhancing classification performance and interpretability in anomaly-prone environments such as sensor networks. Additionally, we demonstrate that conditioning the VAE's latent space with a classifier enhances robustness to noise, a persistent challenge in high-dimensional anomaly detection [329].

5.3 Methodology

5.3.1 Experimental Setup

Magnetorheological (MR) dampers represent a class of devices employed in various engineering applications, such as automotive suspension systems, seismic protection, and industrial machinery, where they facilitate adaptive control of vibration and energy dissipation. These dampers function based on the unique properties of magnetorheological fluids, which exhibit a significant change in viscosity in response to an applied magnetic field, thereby allowing for dynamic adjustments to damping characteristics. The performance of MR dampers is characterized by the nonlinear relationship as a result of the innate hysteretic behaviour. Figure 5.1 presents a labelled cutaway of the various components core to the function of a typical MR damper. Its operation centers on converting mechanical energy into frictional losses by exploiting the rheological characteristics of the MR fluid housed within the damper. Under an applied external force, MR fluid is

displaced between chambers via piston orifices. An electrical current induces a magnetic field, aligning the ferrous particles in the MR fluid, thereby modifying its viscosity and consequently altering the damping characteristics [316]. By varying the applied voltage to the coils, precise control over the damping characteristics in real-time may be achieved, thus aligning performance with the demands of particular applications. In light of the significance of controlled damping in engineering systems, the implementation of condition monitoring for MR dampers becomes necessary. Continuous assessment of the damper's performance and operational state enables the maintenance of functionality within desired parameters, contributing to enhanced reliability and potentially extending the lifespan of the equipment.

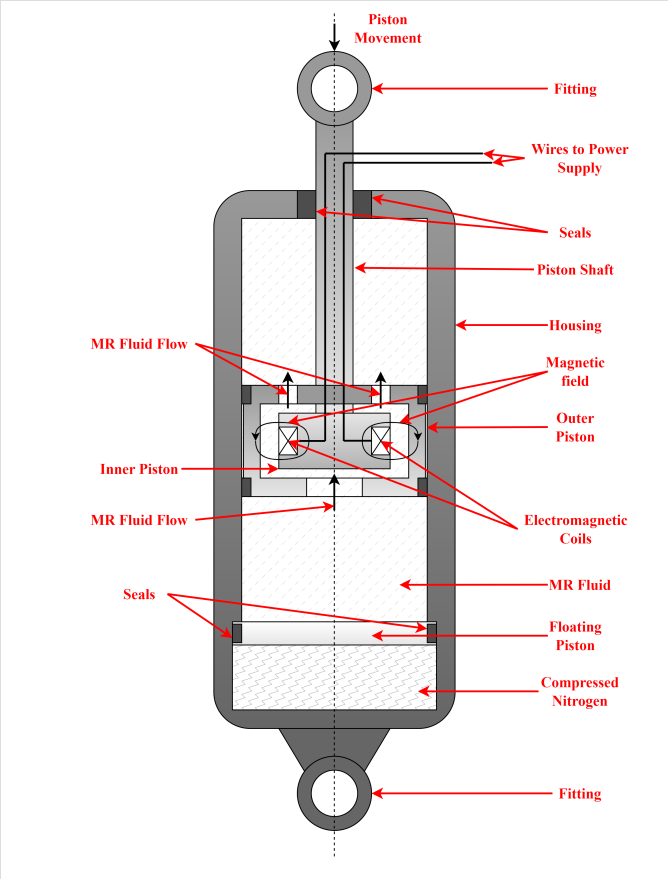


FIGURE 5.1: A labelled depiction of an MR damper.

The Lord RD 8041-1 MR damper employed in this study is classified as a monotube shock absorber, pressurized with nitrogen gas at 300 psi to ensure piston extension under no-load conditions. The response time to changes in the magnetic field for the Lord RD 8041-1 damper is approximately 15 ms [330]. The inclusion of an accumulator

compensates for volumetric changes resulting from piston displacement [331, 332]. The performance of MR dampers is sensitive to ambient temperature fluctuations, with coil resistance directly affecting the generated magnetic field. At 22°C, the coil resistance is typically 5 Ω, increasing to approximately 7 Ω at 71°C [330]. This resistance variation influences the magnetic field strength and particle alignment efficiency. For this study, the damper operates at room temperature (22°C).

A summary of its operational parameters is provided in Table 5.1, with further electrical details available in Table 5.2.

Properties	Value
Stroke length	74 $[mm]$
Extended Length	248 $[mm]$
Body Diameter	42.1 $[mm]$ max
Shaft Diameter	10 $[mm]$
Tensile Strength	8896 $[N]$ max
Peak to Peak Damper Forces 5 cm/sec at 1 A	> 2447 $[N]$
Peak to Peak Damper Forces 5 cm/sec at 1 A	< 667 $[N]$
Operating Temperature	71 $[°C]$ max

Table 5.1: Typical properties of the LORD RD-8041-1 MR damper.

Properties	Value
Input Current: Continuous for 30s	1 $[A]$ max
Input Current: Intermittent	2 $[A]$ max
Input Voltage	12 $[V]$
Resistance at ambient temperature	5 [Ω]
Resistance at maximum operating Temperature (71 $[°C]$)	7 [Ω] max

Table 5.2: Electrical properties of the LORD RD-8041-1 MR damper.

The experimental setup integrates the previously introduced MR damper with a linear actuation system, force sensing, and control mechanisms to facilitate the validation of anomaly detection methods. Figure 5.1 depicts the experimental setup and its various components.

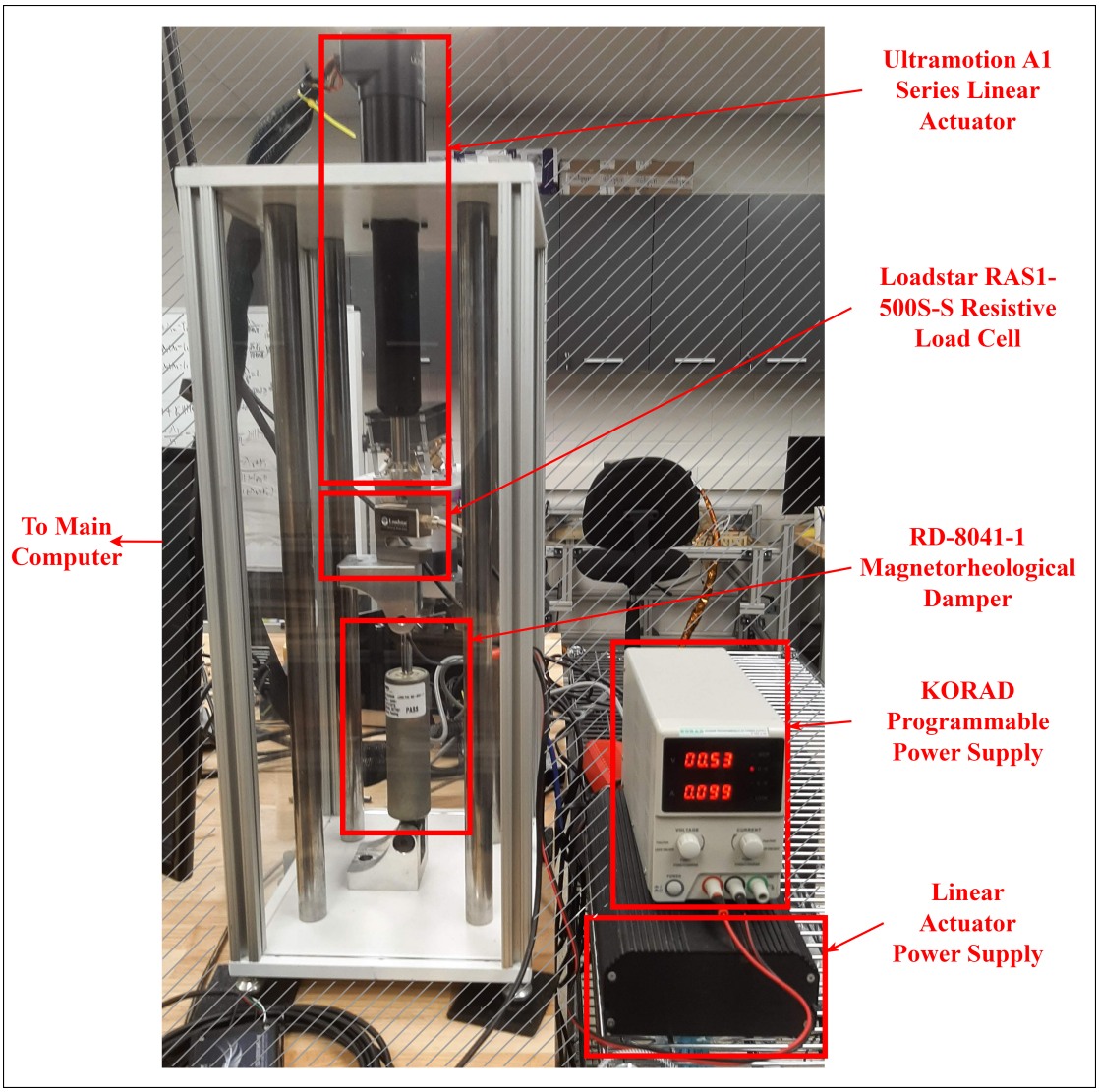


FIGURE 5.2: The experimental setup utilized herein, features the MR damper and necessary components for actuation and sensing.

Control of MR damper damping properties are controlled through the KORAD programmable power supply, allowing for the exploration of varied damping characteristics through excitation adjustments. This power supply's selection was based on its digital control features, including encoder-controlled interfaces and USB control capabilities [333].

The actuation system employs an Ultramotion linear servo, featuring a rod-style actuator paired with a configurable brushless DC motor controller. It provides position feedback with a resolution of 3.1 micrometers and includes a self-locking acme screw mechanism to prevent back driving [334].

In terms of mechanical properties, the actuator exhibits a dynamic continuous load capacity of 756 N and a peak dynamic load capacity of 1512 N, operating with a power rating of 180 W. It achieves a maximum speed of 356 mm/s with a stroke length of 76.2 mm [334]. Control is achieved through RS-422 serial communication, with onboard sensors monitoring position, torque, temperature, and humidity. Position tracking is provided by a multi-turn magnetic encoder, delivering a resolution of 1024 counts per revolution [334].

The force sensor integrated into the setup is the RAS1-500S-S resistive S-Beam load cell, with a capacity of 226.8 kgf (2224.11 N), constructed from tool steel and offering an accuracy of $\pm 0.02\%$ [335]. Calibration is traceable to the National Institute of Standards and Technology (NIST). Data collection from the force sensor is handled via the DI-10000UHS-1K USB interface, enabling data streaming at 1000 Hz [335].

5.3.2 Baseline Models for Anomaly Detection

The proposed VAE-MLP model is evaluated against several baseline anomaly detection models to establish its effectiveness and robustness. These baseline models include a mix of traditional statistical methods and deep learning techniques, each of which serves as a reference point for assessing how well the VAE-MLP can distinguish between normal and anomalous data in high-dimensional sensor readings. By comparing the VAE-MLP to these varied approaches, we aim to demonstrate its advantages in both anomaly separability and classification accuracy.

The following baseline models were used to provide a comprehensive comparison:

- **Principal Component Analysis (PCA)**: PCA is used to reduce data dimensionality by transforming it into a set of principal components that capture the directions of maximum variance. Anomalies are detected based on their deviation

from the variance structure of normal data within the transformed space. This model serves as a simple, interpretable approach to identifying outliers in high-dimensional data.

- **Autoencoder (AE)**: A standard Autoencoder was employed as a baseline, trained solely to minimize reconstruction error. The AE compresses data into a lower-dimensional representation and reconstructs it, identifying anomalies by their high reconstruction errors, as these data points do not conform to the patterns learned from normal data.
- **Isolation Forest**: Isolation Forest is an ensemble-based unsupervised anomaly detection model that isolates anomalies by creating recursive random partitions in the feature space. Data points requiring fewer partitions are likely to be anomalies, as they are more isolated. This method does not rely on assumptions about data distribution, making it suitable for diverse datasets.
- **Vanilla Variational Autoencoder (VAE)**: A standard VAE, or Vanilla VAE, was implemented as a baseline. This model includes only the encoder and decoder components, without an integrated classifier. It learns a latent representation of the data through reconstruction, with anomalies identified based on either high reconstruction error or outlier positions in the latent space. This unsupervised VAE offers a point of comparison to evaluate the benefits of adding a classification layer to the latent space.
- **Convolutional Neural Network (CNN)**: A CNN was trained on segmented time-series data from the sensors, leveraging its ability to automatically extract features from raw data. This model provides a supervised approach to anomaly detection without relying on latent space regularization, serving as a contrasting benchmark to latent space-based methods like the VAE.

5.3.3 Proposed Model: Variational Autoencoder (VAE) with Integrated MLP Classifier

The core of our approach is a hybrid VAE model with an integrated Multilayer Perceptron (MLP) classifier, which jointly optimizes representation learning and anomaly classification. The VAE-MLP model has a two-part architecture consisting of an encoder-decoder structure for representation learning and an additional MLP classifier to enhance anomaly detection.

The VAE Architecture consists of an encoder that compresses high-dimensional input data into a lower-dimensional latent space by generating both the mean (μ) and log-variance ($\log \sigma^2$) vectors. This latent representation allows for effective data compression and regularization through a Gaussian prior. The decoder reconstructs the original input data from the latent space, with the reconstruction loss guiding the encoder to capture essential features of normal and anomalous data.

The MLP Classifier operates directly on the latent mean vector, μ, produced by the encoder, classifying each segment as either normal or anomalous. The classifier network is a three-layer MLP with GELU activations and a softmax layer, optimized alongside the VAE. By integrating the classifier, the model is encouraged to organize the latent space in a way that separates normal and anomalous data, enhancing classification performance in the latent space.

The training objective for the VAE-MLP model is a composite loss function, which combines the reconstruction loss, Kullback-Leibler (KL) divergence, and classification loss:

$$\text{Total Loss} = \text{Reconstruction Loss} + \beta \cdot \text{KL Divergence} + \alpha \cdot \text{Classification Loss} \quad (5.1)$$

where β and α are hyperparameters that control the relative contributions of the KL divergence and classification loss. The reconstruction loss minimizes the error between the input and reconstructed output, the KL divergence regularizes the latent space, and the classification loss enables accurate detection of anomalies.

5.3.4 Training and Evaluation Protocol

To train the VAE-MLP model, we performed hyperparameter tuning, adjusting the latent dimension, learning rate, and the weights of the KL and classification loss components (β and α) to achieve optimal performance. The model was trained using the Adam optimizer over 1500 epochs, monitoring reconstruction, KL, and classification losses throughout.

The time-series data was preprocessed into fixed-length segments of 500 samples to ensure consistency in input dimensions for the model. Each segment comprised ‘Force‘ and ‘Position‘ time-series measurements, which were concatenated into a unified feature array for every sample. If a segment exceeded the target length, it was truncated to the

first 500 samples, while segments shorter than 500 samples were padded with zeros at the end to match the required input length. This segmentation process ensured that all inputs to the model maintained the same dimensionality, making them suitable for batch training. Furthermore, each segment was labeled based on the associated condition of the data, with the assumption that labels remained consistent across each segment.

After segmentation, the data was converted into PyTorch tensors for efficient processing within the deep learning pipeline. Each segment, originally a two-dimensional array of shape (500×2), was flattened into a one-dimensional array of shape (1000) to simplify input handling. This transformation preserved all feature information while aligning with the input requirements of the model.

Normalization was applied to the data to standardize the feature scales and improve the stability of the training process. Normalization was performed by subtracting the mean and dividing it by the standard deviation. A small epsilon value (10^{-8}) was added to the standard deviation to prevent division by zero and ensure numerical stability. Normalization not only brought the features to a common scale but also mitigated the effects of varying magnitudes in 'Force' and 'Position' measurements, which could otherwise hinder model performance.

This preprocessing pipeline, comprising segmentation, tensor conversion, and normalization, was critical in preparing the data for robust and consistent training. It ensured that the input data was both uniform in shape and scaled appropriately, allowing the VAE-MLP to focus on learning meaningful latent representations rather than compensating for inconsistencies in the raw input.

For evaluation, we measured both classification accuracy and F1 score on the test set to assess the model's ability to distinguish between normal and anomalous data. Additionally, to gain insights into the latent space organization, we visualized the latent space using t-SNE and PCA. These visualizations provided a qualitative understanding of how well the model separated anomalies from normal data within the latent space, offering further validation of its anomaly detection capability.

5.4 Results and Analysis

This section presents the performance of different anomaly detection models, with results divided into baseline methods, the Vanilla VAE, and the proposed conditioned VAE-MLP

model. The performance of each model is evaluated in terms of accuracy, F1 score, and AUC, as shown in Tables I, II, and III and Figures 5.3, 5.4, and 5.5.

5.4.1 Baseline Methods

Table III summarizes the results for traditional baseline methods, including PCA, Autoencoder (AE), Isolation Forest, and CNN. These models are commonly used for anomaly detection but lack the explicit representation learning capabilities of deep neural networks. Among these, PCA achieves the lowest accuracy at 42% and an F1 score of 0.6. Isolation Forest and AE perform moderately better, with accuracies around 61-63%, but still fall short in terms of F1 scores, achieving 0.2 and 0.24, respectively.

TABLE 5.3: Accuracy and F1 Scores for Baseline Methods

Model	**Accuracy**	**F1**
PCA	42%%	60%
AE	63%	24%
Isolation Forest	61%	20%
CNN	63%	36%

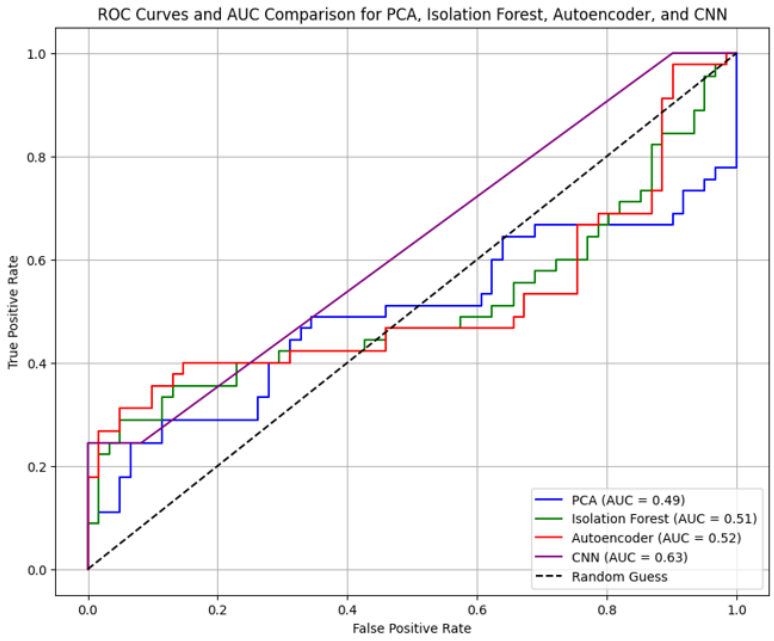


FIGURE 5.3: ROC Curves and AUC Comparison for Traditional Baseline Methods (PCA, Isolation Forest, Autoencoder (AE), CNN). The AUC scores for these methods generally show lower performance, reflecting challenges in detecting anomalies in high-dimensional, noisy data without explicit representation learning.

CNN achieves the best results among the baselines, with an accuracy of 63% and an F1 score of 0.36, showing some capacity for capturing complex patterns in the data. However, the low overall performance of these methods underscores their limitations when applied to high-dimensional, noisy sensor data. The lack of structured latent representation or targeted anomaly classification hinders their ability to effectively separate anomalous from normal data, as reflected in their low ROC-AUC scores in Figure 5.3.

5.4.2 Vanilla VAE

To improve upon these baseline methods, the Vanilla VAE model (Table II) introduces an unsupervised representation learning framework by mapping the data into a latent space through a probabilistic encoder-decoder structure. This model improves performance slightly over traditional baselines, with Gradient Boosting and Random Forest achieving accuracies of 82.08% and 80.19%, respectively, and F1 scores of 76% and 72%. Consequently, while the Vanilla VAE performs moderately well, it struggles to achieve clear separation between normal and anomalous data points in the latent space. The ROC curve in Figure 5.4 reflects this limitation, as the Vanilla VAE's AUC scores, while better than traditional baselines, are still lower than those achieved by the conditioned VAE-MLP.

TABLE 5.4: Accuracy and F1 Scores for Vanilla VAE

Model	**Accuracy**	**F1**
MLP	78.30%	72%
Logistic Regression	53.77%	0%
SVM	67.92%	41%
KNN	68.87%	52%
Naïve Bayes	54.72%	35%
Gradient Boosting	**82.08**%	**76**%
AdaBoost	77.36%	69%
XGBoost	76%	68%
Random Forest (Bagging)	80.19%	72%
Decision Tree (Bagging)	74.53%	65%

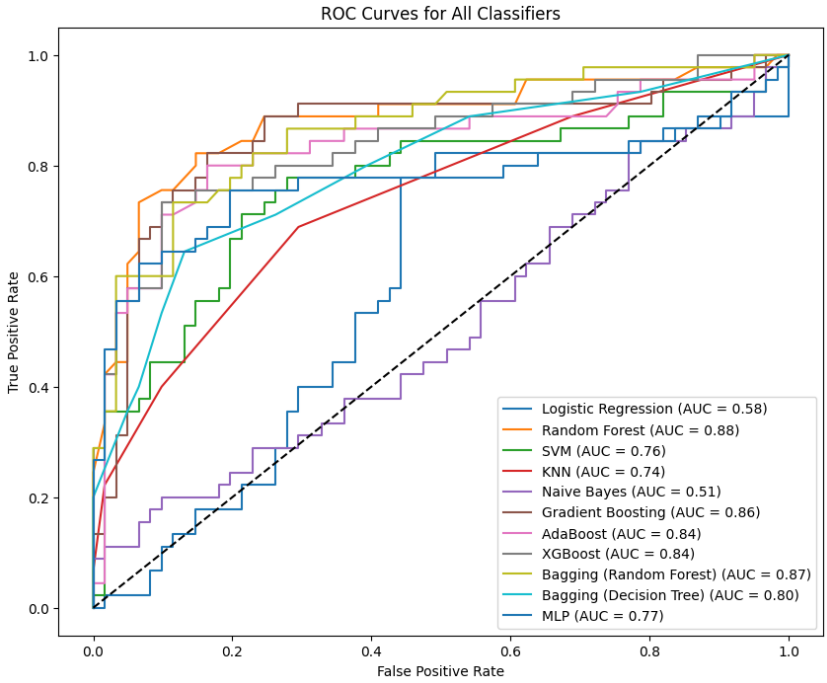


FIGURE 5.4: ROC Curve and AUC for Vanilla VAE. The Vanilla VAE serves as an unsupervised variational model baseline, showing moderate AUC performance. It lacks a dedicated classifier layer to enhance anomaly separability in the latent space, resulting in less distinct classification compared to the conditioned VAE-MLP.

5.4.3 Conditioned VAE-MLP (Proposed Model)

The absence of an integrated classifier limits the model's ability to optimize the latent space specifically for anomaly detection. The proposed conditioned VAE-MLP model further enhances the anomaly detection framework by incorporating an MLP classifier directly within the VAE's latent space. This integration explicitly conditions the latent space to optimize for anomaly separability, leading to significantly improved results. As shown in Table I, the internal MLP classifier achieves an accuracy of 92.45% and an F1 score of 90%, representing the highest performance across all models tested. Even with an external MLP applied to the conditioned VAE's latent space, the model achieves a competitive accuracy of 83.02% and an F1 score of 79%, demonstrating the robustness of the conditioned latent representations. Additionally, tree-based models, such as Random Forest and XGBoost, perform exceptionally well on the conditioned latent space, achieving accuracies of 81.13% and 82.08%, respectively, with high F1 scores, further highlighting the value of the learned representations for anomaly detection. The ROC curves in Figure 5.5 demonstrate the enhanced separability of anomalies, with AUC scores consistently higher than those of the Vanilla VAE and baseline methods. This improvement underscores the importance of conditioning the VAE with a classifier, which shapes the latent space to distinctly cluster normal and anomalous samples, facilitating more effective classification.

Table 5.5: Accuracy and F1 Scores for Conditioned VAE-MLP

Model	**Accuracy**	**F1**
MLP (internal)	**92.45**%	**90**%
MLP (external)	83.02%	79%
Logistic Regression	67.92%	56%
SVM	82.08%	73%
KNN	75.47%	62%
Naïve Bayes	52.83%	43%
Gradient Boosting	80.19%	73%
AdaBoost	63.21%	49%
XGBoost	82.08%	77%
Random Forest (Bagging)	81.13%	75%
Decision Tree (Bagging)	72.64%	62%

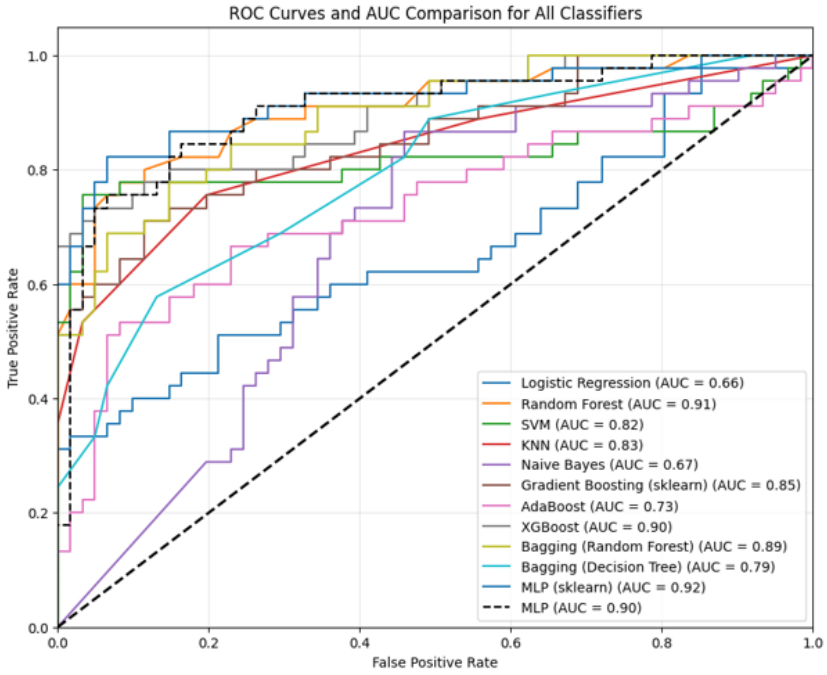


FIGURE 5.5: ROC Curves and AUC Comparison for Conditioned VAE-MLP. ROC curves for various standard classifiers (e.g., Logistic Regression, Random Forest, SVM, KNN) applied to the latent space representations for anomaly classification in the conditioned VAE-MLP. The AUC values indicate the effectiveness of each classifier in distinguishing between normal and anomalous data, showing improved performance due to the integrated MLP layer in the VAE.

5.4.4 Comparative Analysis

Comparing the results across all three groups of models: baseline methods, Vanilla VAE, and conditioned VAE-MLP demonstrates the clear advantage of the proposed approach. Traditional baselines lack the capacity to structure the latent space for anomaly separation, as evidenced by their lower accuracy, F1, and AUC scores. The Vanilla VAE improves upon these baselines through unsupervised representation learning, yet it falls short due to the absence of a classification objective. In contrast, the conditioned VAE-MLP not only learns a meaningful latent representation but also optimizes it for anomaly detection through its integrated MLP classifier.

The conditioned VAE-MLP achieves significantly higher accuracy, F1, and AUC scores, indicating that the explicit conditioning allows for a more discriminative latent space. This structured approach is particularly effective for high-dimensional sensor data, where complex patterns are essential for distinguishing subtle anomalies. In summary, the conditioned VAE-MLP demonstrates superior anomaly detection capabilities, validating the effectiveness of conditioning the VAE with a classifier layer to achieve robust and interpretable results in anomaly detection tasks.

Figures 5.6 and 5.7 illustrate the structure of the latent spaces learned by the conditioned VAE-MLP and the Vanilla VAE models, respectively. Both figures use PCA and t-SNE to project the latent representations of the training and test sets, with anomalies and normal samples color-coded for visual clarity. These projections provide insight into the separability of anomalies from normal data within the latent space.

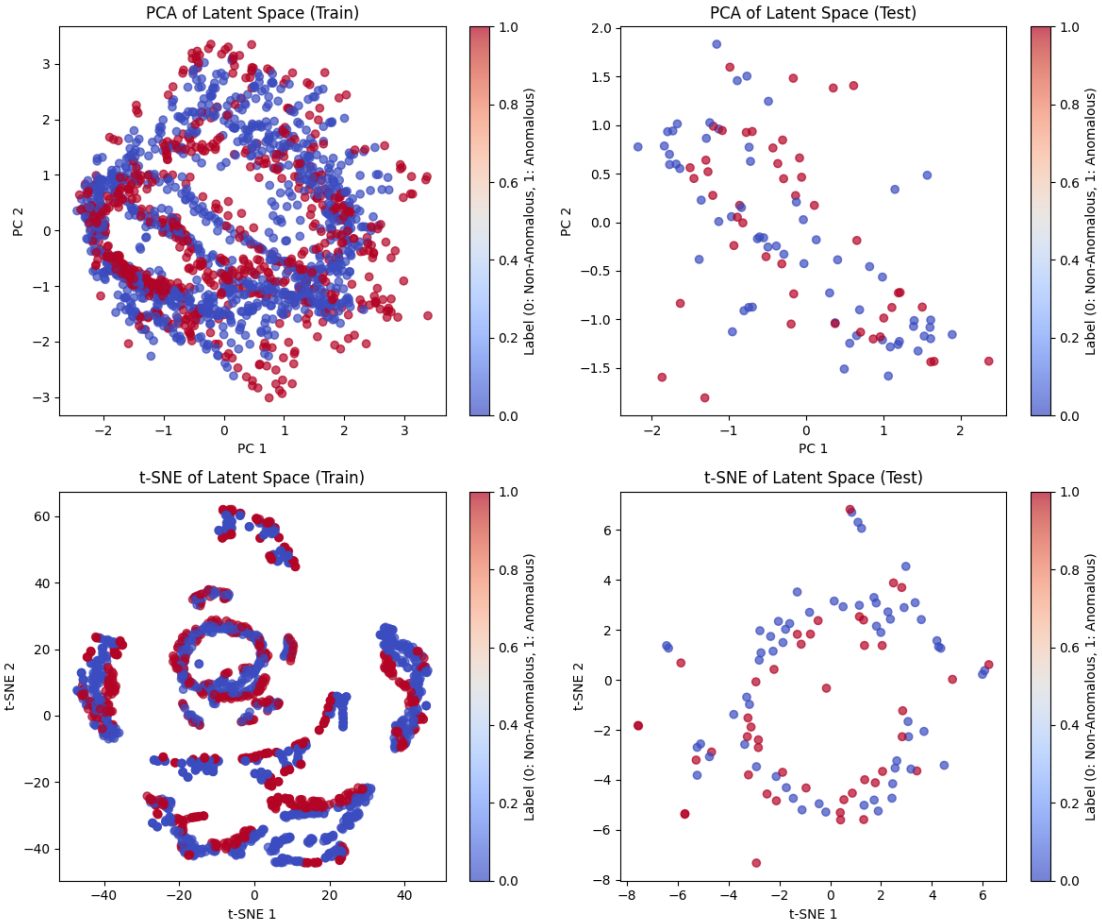


FIGURE 5.6: Latent space representation of the conditioned VAE-MLP model visualized with PCA (top) and t-SNE (bottom) for training and test sets. Anomalous samples are more clearly clustered separately from normal data, especially in the t-SNE projections, demonstrating the enhanced separability achieved through conditioning.

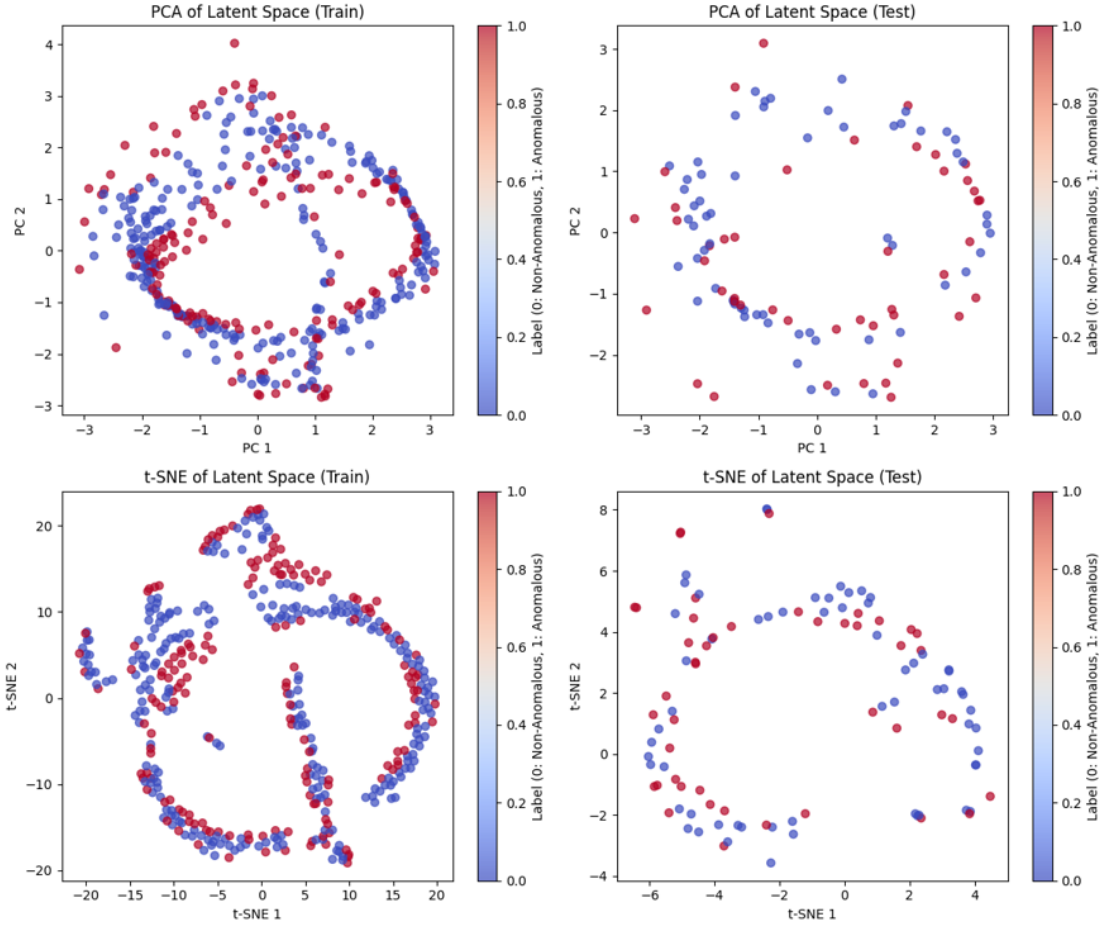


FIGURE 5.7: Latent space representation of the Vanilla VAE model visualized with PCA (top) and t-SNE (bottom) for training and test sets. The clustering of anomalous samples is less distinct compared to the conditioned VAE-MLP, indicating limited anomaly separability in the latent space.

The conditioned VAE-MLP model (Figure 5.6) demonstrates distinct clustering patterns between normal and anomalous data points. In both the PCA and t-SNE projections, the anomalous samples form noticeably separate clusters from the normal samples, especially in the t-SNE plots. This indicates that the conditioned VAE-MLP effectively structures the latent space to enhance separability, likely due to the integrated MLP classifier, which encourages the VAE to organize the latent representations in a way that supports anomaly detection. The training and test sets display consistent clustering patterns, suggesting that the conditioned VAE-MLP generalizes well in distinguishing anomalies from normal data in unseen samples.

In contrast, the Vanilla VAE model (Figure 5.7) exhibits a less distinct separation

between anomalous and normal samples. Although some clustering is observed, particularly in the t-SNE projections, the separation between anomalies and normal data points is less pronounced compared to the conditioned VAE-MLP. This limited separability reflects the absence of an explicit classification objective in the Vanilla VAE, resulting in a latent space that lacks the structured separability seen in the conditioned model. The PCA and t-SNE projections for the Vanilla VAE indicate that, while some latent representations of anomalies are distinguishable, they are often interspersed with normal data, reducing the model's effectiveness in reliably identifying anomalies.

By integrating an MLP classifier directly within the VAE framework, the conditioned VAE-MLP optimizes the latent space for anomaly detection, creating well-defined clusters that facilitate the identification of anomalous samples. This structured separation is essential for high-dimensional, noisy sensor data, where subtle anomalies are challenging to detect. The Vanilla VAE, while effective in compressing data, lacks this enhanced separability, demonstrating the importance of a classification component within the VAE to maximize anomaly detection performance.

5.5 Discussion

The results of this study demonstrate the advantages of integrating a classifier within the VAE architecture for enhanced anomaly detection. The conditioned VAE-MLP model consistently outperforms both the Vanilla VAE and traditional baseline methods across various metrics, including accuracy, F1 score, and AUC. This performance boost can be attributed to the structured latent space produced by the conditioned VAE-MLP, which is explicitly optimized for anomaly separability. By conditioning the VAE's latent space with a classification objective, the model is encouraged to learn representations that are not only compressed but also discriminative, making it highly effective in distinguishing between normal and anomalous data.

The PCA and t-SNE projections of the latent space provide further insight into the model's behavior. For the conditioned VAE-MLP, the latent space clusters anomalies distinctly from normal data, as visualized in both training and test projections. This clustering suggests that the model has learned a stable latent representation that generalizes well to unseen data, a crucial requirement for real-world anomaly detection applications where new, previously unseen anomalies may emerge. In contrast, the Vanilla VAE, while capable of compressing data effectively, lacks this enhanced separability, as

evidenced by the mixed clustering of anomalous and normal samples in its latent space projections.

Moreover, the conditioned VAE-MLP's high performance across tree-based and neural classifiers applied to the latent space suggests that the learned representations are highly versatile and interpretable, making them suitable for downstream tasks beyond simple anomaly detection. The enhanced separability seen in the conditioned VAE-MLP latent space also implies that this approach could be beneficial in scenarios where interpretability is essential, as it allows practitioners to visually inspect clusters and potentially identify the nature of detected anomalies.

Despite these promising results, some limitations should be acknowledged. The conditioning process, while effective, adds computational complexity to the VAE framework, which may not be ideal for environments with limited computational resources. Additionally, while the conditioned VAE-MLP shows strong generalizability, future work could explore techniques to further improve its robustness to highly complex and dynamic anomalies, which may require more sophisticated conditioning mechanisms or the incorporation of domain-specific knowledge.

5.6 Conclusion

In this study, we proposed a conditioned VAE-MLP model for anomaly detection in high-dimensional sensor data. By integrating a classifier within the VAE framework, the model is encouraged to learn a structured latent representation that optimizes separability between normal and anomalous data. The experimental results demonstrate that the conditioned VAE-MLP significantly outperforms both the Vanilla VAE and traditional baseline methods, achieving higher accuracy, F1 scores, and AUC metrics. The visualization of latent space representations via PCA and t-SNE further corroborates the advantages of the conditioned VAE-MLP, highlighting distinct clusters for anomalies and normal samples, particularly in the test set. This clear separation underscores the model's potential for real-world applications where reliable anomaly detection and interpretability are paramount.

The findings of this research provide strong evidence that conditioning the VAE's latent space with a classification objective can enhance anomaly detection capabilities, making the model a valuable tool for applications in high-dimensional, noisy environments. Future research could investigate ways to further reduce computational overhead

and enhance robustness to complex anomaly patterns, potentially through hybrid models that incorporate domain knowledge or adaptive conditioning strategies.

Overall, the conditioned VAE-MLP represents a promising approach for anomaly detection, bridging the gap between unsupervised representation learning and supervised anomaly classification. By fostering a more discriminative latent space, this approach not only improves detection performance but also provides a foundation for interpretability and flexibility in subsequent analysis tasks.

Chapter 6

Conclusion

The integration of IoT and remote sensing systems has created unprecedented challenges in managing, processing, and transmitting heterogeneous data. These challenges are amplified by the diversity of data modalities, such as numerical measurements, optical imagery, and radar data, as well as the need for efficient compression, integration, and anomaly detection. This research leverages existing **compressive Variational Autoencoders (VAEs)**, integrating them into a novel architecture to address these challenges. Through fine-tuning and architectural adjustments, these VAEs are optimized for tasks such as data fusion, classification, and anomaly detection, demonstrating their utility as a unifying framework for heterogeneous data integration.

The study is structured around three core contributions:

- **Leveraging Compressive VAEs for Data Processing and Classification** By employing a range of pre-existing compressive VAEs, this study fine-tuned these models within a tailored architecture to optimize their latent spaces for direct classification and efficient data representation. The approach reduces computational overhead by utilizing latent spaces directly for downstream tasks such as classification, bypassing traditional reconstructive processing. Experimental results illustrate that fine-tuning these models significantly improves classification accuracy and interpretability, while maintaining competitive compression performance metrics such as Bits-Per-Pixel (BPP) and Peak Signal-to-Noise Ratio (PSNR).
- **Data Fusion Using Unified Latent Spaces** Heterogeneous data modalities, such as Synthetic Aperture Radar (SAR) and optical imagery, present unique challenges in integration due to differences in scale, resolution, and semantic content. The proposed architecture fuses disparate data types into unified latent spaces,

enabling robust analytics and classification. Comparative evaluations against traditional fusion methods—such as Principal Component Analysis (PCA), Discrete Wavelet Transform (DWT), and Spectral Analysis—demonstrate that latent-space fusion outperforms these methods in accuracy and data quality. The fused representations also enable more nuanced downstream tasks, such as object detection or environmental monitoring.

- **Anomaly Detection Through Probabilistic Latent Spaces** Detecting anomalies in high-dimensional, noisy datasets is critical for applications such as industrial monitoring, smart cities, and autonomous systems. By fine-tuning compressive VAEs to include anomaly detection capabilities, this study capitalized on the probabilistic structure of latent spaces to identify outliers with high reliability. The proposed architecture integrates a classification layer directly into the latent space, enabling the identification of anomalies with precision even under challenging conditions of incomplete or noisy data.

6.1 Summary of Research

This research makes significant strides in the application of compressive VAEs to heterogeneous data integration:

- **Fine-Tuning for Improved Performance:** By fine-tuning pre-existing compressive VAEs, the proposed architecture achieves a balance between compression efficiency and analytical performance. The approach demonstrates superior results compared to baseline methods, especially in resource-constrained scenarios.
- **Unified Latent Space Representations:** Fusing disparate modalities, such as SAR and optical data, into a single latent representation enables more effective analysis and interpretation, paving the way for advanced applications in remote sensing and IoT.
- **Direct Latent Space Utilization:** The novel utilization of VAE latent spaces for classification and anomaly detection eliminates the need for reconstructive processing, streamlining workflows and reducing computational demands.
- **Enhanced Data Fusion:** The methodology demonstrates that fine-tuned latent-space fusion provides superior performance over traditional fusion techniques, ensuring more reliable and interpretable results for downstream tasks.

6.2 Recommendations for Future Work and Directions

The findings of this research underscore the transformative potential of compressive VAEs in addressing core challenges of IoT and remote sensing systems. By leveraging their inherent flexibility and probabilistic nature, these models provide scalable, interpretable, and resource-efficient solutions for data-intensive applications. The proposed architecture, built upon fine-tuned compressive VAEs, offers a roadmap for future innovations in multi-modal data integration, real-time anomaly detection, and resource-efficient analytics.

This study lays the foundation for future work in:

- Deploying compressive VAEs on edge devices for real-time, low-resource analytics.
- Extending the architecture to incorporate advanced neural components, such as attention mechanisms, to further enhance fusion and classification tasks.
- Expanding anomaly detection frameworks to cover more complex datasets and evolving operational contexts.
- Exploring federated learning to combine compressive VAEs with privacy-preserving distributed systems for collaborative model training.

By focusing on the refinement and application of existing VAE models, this research provides a practical and impactful contribution to the fields of data compression, fusion, and anomaly detection, with broad implications for IoT, autonomous systems, and beyond.

References

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005, doi: 10.1109/JSAC.2004.839380.

[2] W. Hilal, S. A. Gadsden, and J. Yawney, "Cognitive dynamic systems: A review of theory, applications, and recent advances," *Proceedings of the IEEE*, vol. 111, no. 6, pp. 575–622, 2023.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," 2022. [Online]. Available: https://arxiv.org/abs/2108.07258

[4] A. Giuliano, S. A. Gadsden, and J. Yawney, "Optimizing satellite image analysis: Leveraging variational autoencoders latent representations for direct integration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, no. Art no.

5603123, pp. 1–23, 2025.

[5] A. Giuliano, S. A. Gadsden, and J. Yawney, "Cognitive internet of things: A review of frameworks, applications, and recent advances," *IEEE Communication Surveys and Tutorials*, 2024, first round of revisions.

[6] A. Giuliano, S. A. Gadsden, and J. Yawney, "Enhancing data fusion and classification of sentinel-1 and sentinel-2 imagery using neural compression," *Information Fusion*, 2024, under Review.

[7] A. Giuliano, S. A. Gadsden, and J. Yawney, "Anomaly detection of under-over current in magnetorheological damper suspension using variational autoencoders," *IEEE/ASME Transactions on Mechatronics*, 2024, under Review.

[8] A. Giuliano, S. A. Gadsden, and J. Yawney, "Transformer-based transfer learning for battery state of health estimation," *Energy Reports*, 2024, under Review.

[9] A. Giuliano, S. A. Gadsden, W. Hilal, and J. Yawney, "Convolutional variational autoencoders for secure lossy image compression in remote sensing," in *Sensors and Systems for Space Applications XVII*, K. D. Pham and G. Chen, Eds. National Harbor, United States: SPIE, Jun. 2024, p. 18.

[10] N. Alsadi, A. Giuliano, S. A. Gadsden, and J. Yawney, "An adaptive approach to blockchain in smart system applications," in *Big Data V: Learning, Analytics, and Applications*, vol. 12522. SPIE, 2023, pp. 27–32.

[11] A. Giuliano, G. Bone, S. A. Gadsden, and M. AlShabi, "A comparative analysis of control methods applied to horizontal 2 dof robotic arms," in *2023 Advances in Science and Engineering Technology International Conferences (ASET)*. Dubai, United Arab Emirates: IEEE, Feb. 2023, pp. 01–09.

[12] A. Giuliano, W. Hilal, N. Alsadi, J. Yawney, and S. A. Gadsden, "Normalized determinant pooling layer in cnns for multi-label classification," in *Computational Imaging VII*, J. C. Petruccelli and C. Preza, Eds. Orlando, United States: SPIE, Jul. 2023, p. 17.

[13] N. Alsadi, W. Hilal, O. Surucu, A. Giuliano, A. Gadsden, and J. Yawney, "Visual attention for malware classification," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, T. Pham, L. Solomon, and M. E. Hohil, Eds. Orlando, United States: SPIE, Jun. 2022, p. 74.

[14] N. Alsadi, W. Hilal, O. Surucu, A. Giuliano, S. A. Gadsden, and J. Yawney, "An optimized volumetric approach to unsupervised image registration," in *Big Data IV: Learning, Analytics, and Applications*, F. Ahmad, P. P. Markopoulos, and B. Ouyang, Eds. Orlando, United States: SPIE, May 2022, p. 15.

[15] N. Alsadi, W. Hilal, O. Surucu, and et al., "An anomaly detecting blockchain strategy for secure iot networks," in *Disruptive Technologies in Information Sciences VI*, M. Blowers, R. D. Hall, and V. R. Dasari, Eds. Orlando, United States: SPIE, May 2022, p. 18.

[16] N. Alsadi, W. Hilal, O. Surucu, and et al., "Neural network training loss optimization utilizing the sliding innovation filter," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, T. Pham, L. Solomon, and M. E. Hohil, Eds. Orlando, United States: SPIE, Jun. 2022, p. 76.

[17] A. Giuliano, W. Hilal, N. Alsadi, S. A. Gadsden, and J. Yawney, "A review of cognitive dynamic systems and cognitive iot," in *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2022, pp. 1–7, doi: 10.1109/IEMTRONICS55184.2022.9795834.

[18] A. Giuliano, W. Hilal, N. Alsadi, O. Surucu, A. Gadsden, J. Yawney, and Y. Ziada, "Efficient utilization of big data using distributed storage, parallel processing, and blockchain technology," in *Big Data IV: Learning, Analytics, and Applications*, F. Ahmad, P. P. Markopoulos, and B. Ouyang, Eds., vol. 12097, International Society for Optics and Photonics. SPIE, 2022, p. 1209704, doi: 10.1117/12.2618891.

[19] W. Hilal, A. Giuliano, S. A. Gadsden, and J. Yawney, "A review of cognitive dynamic systems and its overarching functions," in *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2022, pp. 1–10, doi: 10.1109/IEMTRONICS55184.2022.9795764.

[20] W. Hilal, C. Wilkinson, N. Alsadi, and et al., "A topic modeling-based approach to executable file malware detection," in *Disruptive Technologies in Information Sciences VI*, M. Blowers, R. D. Hall, and V. R. Dasari, Eds. Orlando, United States: SPIE, May 2022, p. 4.

[21] W. Hilal, C. Wilkinson, A. Giuliano, and et al., "Minority class augmentation using gans to improve the detection of anomalies in critical operations," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*,

T. Pham, L. Solomon, and M. E. Hohil, Eds. Orlando, United States: SPIE, Jun. 2022, p. 69.

[22] O. Surucu, C. Wilkinson, U. Yeprem, and et al., "Prognos: An automatic remaining useful life (rul) prediction model for military systems using machine learning," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, T. Pham, L. Solomon, and M. E. Hohil, Eds. Orlando, United States: SPIE, Jun. 2022, p. 71.

[23] O. Surucu, U. Yeprem, C. Wilkinson, and et al., "A survey on ethereum smart contract vulnerability detection using machine learning," in *Disruptive Technologies in Information Sciences VI*, M. Blowers, R. D. Hall, and V. R. Dasari, Eds. Orlando, United States: SPIE, May 2022, p. 12.

[24] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems security—a survey," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1802–1831, 2017, doi: 10.1109/JIOT.2017.2703172.

[25] M. A. Ahamat, N. F. Zulkefli, N. M. Nur, A. S. M. Rafie, E. N. Roslin, and R. Abidin, "Innovative approach for biomimicry of marine animals for development of engineering devices," in *Advanced Maritime Technologies and Applications*, A. Ismail, W. M. Dahalan, and A. Öchsner, Eds. Cham: Springer International Publishing, 2022, pp. 301–310, doi: 10.1007/978-3-030-89992-9_26.

[26] E. J. Billingsley, M. Ghommem, R. Vasconcellos, and A. Abdelkefi, *Biomimicry and Aerodynamic Performance of Multi-Flapping Wing Drones*. American Institute of Aeronautics and Astronautics, Inc., 2021, doi: 10.2514/6.2021-0227.

[27] H. Chowdhury, R. Islam, M. Hussein, M. Zaid, B. Loganathan, and F. Alam, "Design of an energy efficient car by biomimicry of a boxfish," *Energy Procedia*, vol. 160, pp. 40–44, 2019, doi: 10.1016/j.egypro.2019.02.116.

[28] M. J. Thompson, J. Burnett, D. M. Ixtabalan, D. Tran, A. Batra, A. Rodriguez, and B. Steele, *Experimental Design of a Flapping Wing Micro Air Vehicle through Biomimicry of Bumblebees*. American Institute of Aeronautics and Astronautics, Inc., 2015, doi: 10.2514/6.2015-1454.

[29] E. N. Davison, K. J. Schlesinger, D. S. Bassett, M.-E. Lynall, M. B. Miller, S. T. Grafton, and J. M. Carlson, "Brain network adaptability across task states," *PLoS Comput Biol*, vol. 11, no. 1, p. e1004029, Jan. 2015, doi: 10.1371/JOURNAL.PCBI.1004029.

[30] S. Haykin, "Cognitive dynamic systems," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1910–1911, 2006, doi: 10.1109/JPROC.2006.886014.

[31] S. Haykin, "Cognitive radar: a way of the future," *IEEE Signal Processing Magazine*, vol. 23, no. 1, pp. 30–40, 2006, doi: 10.1109/MSP.2006.1593335.

[32] S. Haykin, "Cognitive dynamic systems," *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, vol. 5, no. 4, pp. 33–43, 2011, doi: 10.4018/jcini.2011100103.

[33] Q. Wu, G. Ding, Y. Xu, S. Feng, Z. Du, J. Wang, and K. Long, "Cognitive internet of things: A new paradigm beyond connection," *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 129–143, 2014, doi: 10.1109/JIOT.2014.2311513.

[34] A. Sheth, "Internet of things to smart iot through semantic, cognitive, and perceptual computing," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 108–112, 2016, doi: 10.1109/MIS.2016.34.

[35] A. Zelenkauskaite, N. Bessis, S. Sotiriadis, and E. Asimakopoulou, "Interconnectedness of complex systems of internet of things through social network analysis for disaster management," in *2012 Fourth International Conference on Intelligent Networking and Collaborative Systems*, 2012, pp. 503–508, doi: 10.1109/iNCoS.2012.25.

[36] J. Powell, A. McCafferty-Leroux, W. Hilal, and S. A. Gadsden, "Smart grids: A comprehensive survey of challenges, industry applications, and future trends," *Energy Reports*, vol. 11, pp. 5760–5785, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484724003299

[37] A. A. Khan, M. H. Rehmani, and A. Rachedi, "Cognitive-radio-based internet of things: Applications, architectures, spectrum related functionalities, and future research directions," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 17–25, 2017, doi: 10.1109/MWC.2017.1600404.

[38] S. Krčo, B. Pokrić, and F. Carrez, "Designing iot architecture(s): A european perspective," in *2014 IEEE World Forum on Internet of Things (WF-IoT)*, 2014, pp. 79–84, doi: 10.1109/WF-IoT.2014.6803124.

[39] I. Florea, R. Rughinis, L. Ruse, and D. Dragomir, "Survey of standardized protocols for the internet of things," in *2017 21st International Conference on Control*

Systems and Computer Science (CSCS), 2017, pp. 190–196, doi: 10.1109/CSCS.2017.33.

[40] H. Brunner, R. Hofmann, M. Schuß, J. Link, M. Hollick, C. A. Boano, and K. Römer, "Leveraging cross-technology broadcast communication to build gateway-free smart homes," in *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2021, pp. 1–9.

[41] J. Jagannath, N. Polosky, A. Jagannath, F. Restuccia, and T. Melodia, "Machine learning for wireless communications in the internet of things: A comprehensive survey," *Ad Hoc Networks*, vol. 93, p. 101913, 2019, doi: 10.1016/j.adhoc.2019.101913.

[42] P. Rawat, K. D. Singh, and J. M. Bonnin, "Cognitive radio for m2m and internet of things: A survey," *Computer Communications*, vol. 94, pp. 1–29, 2016, doi: 10.1016/j.comcom.2016.07.012.

[43] L. Cui, S. Yang, F. Chen, Z. Ming, N. Lu, and J. Qin, "A survey on application of machine learning for internet of things," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 8, pp. 1399–1417, Aug 2018, doi: 10.1007/s13042-018-0834-5.

[44] B. Li, D. He, and Y. Jiang, "Research on the application of artificial intelligence and computational intelligence in the internet of things," *Journal of Physics: Conference Series*, vol. 1915, no. 4, p. 042020, may 2021, doi: 10.1088/1742-6596/1915/4/042020.

[45] J. M. Fuster, "Frontal lobe and cognitive development," *J Neurocytol*, vol. 31, no. 3-5, pp. 373–385, Mar. 2002, doi: 10.1023/a:1024190429920.

[46] S. Haykin, K. Huber, and Z. Chen, "Bayesian sequential state estimation for mimo wireless communications," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 439–454, 2004, doi: 10.1109/JPROC.2003.823143.

[47] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1136–1159, 2013, doi: 10.1109/SURV.2012.100412.00017.

[48] A. Kaur and K. Kumar, "A comprehensive survey on machine learning approaches for dynamic spectrum access in cognitive radio networks," *Journal of Experimental*

& Theoretical Artificial Intelligence, vol. 34, no. 1, pp. 1–40, 2022, doi: 10.1080/0952813X.2020.1818291.

[49] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019, doi: 10.1109/COMST.2019.2916583.

[50] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016, doi: 10.1109/JIOT.2016.2579198.

[51] G. P. Villardi, G. Thadeu Freitas de Abreu, and H. Harada, "Tv white space technology: Interference in portable cognitive emergency network," *IEEE Vehicular Technology Magazine*, vol. 7, no. 2, pp. 47–53, 2012, doi: 10.1109/MVT.2012.2190221.

[52] R. Ferrus, O. Sallent, G. Baldini, and L. Goratti, "Public safety communications: Enhancement through cognitive radio and spectrum sharing principles," *IEEE Vehicular Technology Magazine*, vol. 7, no. 2, pp. 54–61, 2012, doi: 10.1109/MVT.2012.2190180.

[53] G. P. Joshi, S. Y. Nam, and S. W. Kim, "Cognitive radio wireless sensor networks: Applications, challenges and research trends," *Sensors*, vol. 13, no. 9, pp. 11 196–11 228, 2013, doi: 10.3390/s130911196.

[54] R. Doost-Mohammady and K. R. Chowdhury, "Transforming healthcare and medical telemetry through cognitive radio networks," *IEEE Wireless Communications*, vol. 19, no. 4, pp. 67–73, 2012, doi: 10.1109/MWC.2012.6272425.

[55] R. M. Aileni, G. Suciu, V. Suciu, S. Pasca, and R. Strungaru, *Health Monitoring Using Wearable Technologies and Cognitive Radio for IoT*. Cham: Springer International Publishing, 2019, pp. 143–165, doi: 10.1007/978-3-319-91002-4_6.

[56] T. Jabeen, I. Jabeen, H. Ashraf, A. Ullah, N. Z. Jhanjhi, R. M. Ghoniem, and S. K. Ray, "Smart wireless sensor technology for healthcare monitoring system using cognitive radio networks," *Sensors*, vol. 23, no. 13, 2023, doi: 10.3390/s23136104.

[57] G. M. Dias Santana, R. S. d. Cristo, and K. R. Lucas Jaquie Castelo Branco, "Integrating cognitive radio with unmanned aerial vehicles: An overview," *Sensors*, vol. 21, no. 3, 2021, doi: 10.3390/s21030830.

[58] E. Biglieri, "An overview of cognitive radio for satellite communications," in *2012 IEEE First AESS European Conference on Satellite Telecommunications (ESTEL)*, 2012, pp. 1–3, doi: 10.1109/ESTEL.2012.6400078.

[59] W. U. Khan, Z. Ali, E. Lagunas, A. Mahmood, M. Asif, A. Ihsan, S. Chatzinotas, B. Ottersten, and O. A. Dobre, "Rate splitting multiple access for next generation cognitive radio enabled leo satellite networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 11, pp. 8423–8435, 2023, doi: 10.1109/TWC.2023.3263116.

[60] Z. Li, S. Wang, S. Han, W. Meng, and C. Li, "Joint design of beam hopping and multiple access based on cognitive radio for integrated satellite-terrestrial network," *IEEE Network*, vol. 37, no. 1, pp. 36–43, 2023, doi: 10.1109/MNET.005.2200466.

[61] D. N. Molokomme, C. S. Chabalala, and P. N. Bokoro, "A review of cognitive radio smart grid communication infrastructure systems," *Energies*, vol. 13, no. 12, 2020.

[62] M. A. Hajahmed, M. Hawa, L. A. Shamlawi, S. Alnaser, Y. Alsmadi, and D. Abualnadi, "Cognitive radio based backup protection scheme for smart grid applications," *IEEE Access*, vol. 8, pp. 71 866–71 879, 2020, doi: 10.1109/ACCESS.2020.2987762.

[63] M. Fatemi and S. Haykin, "Cognitive control: Theory and application," *IEEE Access*, vol. 2, pp. 698–710, 2014, doi: 10.1109/ACCESS.2014.2332333.

[64] W. B. Powell, *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2007, vol. 703.

[65] H. Wang, F. R. Yu, L. Zhu, T. Tang, and B. Ning, "A cognitive control approach to communication-based train control systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1676–1689, 2015.

[66] M. Fatemi, P. Setoodeh, and S. Haykin, "Observability of stochastic complex networks under the supervision of cognitive dynamic systems," *Journal of Complex Networks*, vol. 5, no. 3, pp. 433–460, Sep. 2016, _eprint: https://academic.oup.com/comnet/article-pdf/5/3/433/17654837/cnw021.pdf. [Online]. Available: https://doi.org/10.1093/comnet/cnw021

[67] S. Wang, X. Yin, P. Li, X. Wang, G. Wang, and J. Hu, "A cognitive control approach for networked control systems with random packet losses," *Advanced Theory and Simulations*, vol. 4, no. 11, p. 2100163, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/adts.202100163

[68] S. Wang, X. Yin, Y. Zhang, P. Li, and H. Wen, "Event-triggered cognitive control for networked control systems subject to dos attacks and time delay," *Arabian Journal for Science and Engineering*, vol. 48, no. 5, pp. 6991–7004, May 2023. [Online]. Available: https://doi.org/10.1007/s13369-022-07068-x

[69] S. Wang, X. Yin, P. Li, Y. Zhang, X. Wang, and S. Tong, "Cognitive control using adaptive rbf neural networks and reinforcement learning for networked control system subject to time-varying delay and packet losses," *Arabian Journal for Science and Engineering*, vol. 46, no. 10, pp. 10 245–10 259, Oct 2021. [Online]. Available: https://doi.org/10.1007/s13369-021-05752-y

[70] S. Haykin, J. M. Fuster, D. Findlay, and S. Feng, "Cognitive risk control for physical systems," *IEEE Access*, vol. 5, pp. 14 664–14 679, 2017, doi: 10.1109/ACCESS.2017.2726439.

[71] P. Lea, *IOT and edge computing for architects: implementing edge and IoT systems from sensors to clouds with communication systems, analytics, and security*, second edition ed., ser. Expert insight. Birmingham: Packt, 2020.

[72] T. Xu, J. B. Wendt, and M. Potkonjak, "Security of IoT systems: Design challenges and opportunities," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. San Jose, CA, USA: IEEE, Nov. 2014, pp. 417–423. [Online]. Available: http://ieeexplore.ieee.org/document/7001385/

[73] F. Li, K.-Y. Lam, X. Li, Z. Sheng, J. Hua, and L. Wang, "Advances and emerging challenges in cognitive internet-of-things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5489–5496, 2020, doi: 10.1109/TII.2019.2953246.

[74] W. Chen, Z. Yin, and T. He, "Enabling Global Cooperation for Heterogeneous Networks via Reliable Concurrent Cross Technology Communications," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9380996/

[75] O. Kaynak, W. He, F. Flammini, and Z. Liu, "Towards symbiotic autonomous systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical*

and Engineering Sciences, vol. 379, no. 2207, p. 20200359, 2021, doi: 10.1098/rsta.2020.0359.

[76] Y. Wang, F. Karray, S. Kwong, K. N. Plataniotis, H. Leung, M. Hou, E. Tunstel, I. J. Rudas, L. Trajkovic, O. Kaynak, J. Kacprzyk, M. Zhou, M. H. Smith, P. Chen, and S. Patel, "On the philosophical, cognitive and mathematical foundations of symbiotic autonomous systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2207, p. 20200362, 2021, doi: 10.1098/rsta.2020.0362.

[77] R. Saracco, K. Grise, and T. Martinez, "The winding path towards symbiotic autonomous systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2207, p. 20200361, 2021, doi: 10.1098/rsta.2020.0361.

[78] R. Harrison, D. Vera, and B. Ahmad, "Towards the realization of dynamically adaptable manufacturing automation systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2207, p. 20200365, 2021, doi: 10.1098/rsta.2020.0365.

[79] M. Hou, Y. Wang, L. Trajkovic, K. N. Plataniotis, S. Kwong, M. Zhou, E. Tunstel, I. J. Rudas, J. Kacprzyk, and H. Leung, "Frontiers of brain-inspired autonomous systems: How does defense r&d drive the innovations?" *IEEE Systems, Man, and Cybernetics Magazine*, vol. 8, no. 2, pp. 8–20, 2022, doi: 10.1109/MSMC.2021.3136983.

[80] A. E. Braten, N. Tamkittikhun, F. A. Kraemer, and D. Ammar, "Towards cognitive device management: a testbed to explore autonomy for constrained iot devices," in *Proceedings of the Seventh International Conference on the Internet of Things*, ser. IoT '17. New York, NY, USA: Association for Computing Machinery, 2017, doi: 10.1145/3131542.3140282.

[81] A. E. Braten and F. A. Kraemer, "Towards cognitive iot: Autonomous prediction model selection for solar-powered nodes," in *2018 IEEE International Congress on Internet of Things (ICIOT)*, 2018, pp. 118–125, doi: 10.1109/ICIOT.2018.00023.

[82] F. A. Kraemer, D. Palma, A. E. Braten, and D. Ammar, "Operationalizing solar energy predictions for sustainable, autonomous iot device management," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 803–11 814, 2020, doi: 10.1109/JIOT.2020.3002330.

[83] A. E. Braten, F. A. Kraemer, and D. Palma, "Autonomous iot device management systems: Structured review and generalized cognitive model," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4275–4290, 2021, doi: 10.1109/JIOT.2020.3035389.

[84] S. Feng, P. Setoodeh, and S. Haykin, "Smart home: Cognitive interactive people-centric internet of things," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 34–39, 2017, doi: 10.1109/MCOM.2017.1600682CM.

[85] A. Serianni, F. De Rango, and P. Raimondo, "Cognitive iot enabled by layered architecture and neural networks in a smart home environment," in *2021 Wireless Days (WD)*, 2021, pp. 1–7, doi: 10.1109/WD52248.2021.9508326.

[86] S. Rinaldi, P. Ferrari, A. Flammini, M. Pasetti, E. Sisinni, L. C. Tagliabue, A. C. Ciribini, F. Martinelli, and S. Mangili, "A cognitive strategy for renovation and maintenance of buildings through iot technology," in *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*, 2020, pp. 1949–1954, doi: 10.1109/IECON43393.2020.9254980.

[87] M. Z. Gunduz and R. Das, "Cyber-security on smart grid: Threats and potential solutions," *Computer Networks*, vol. 169, p. 107094, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128619311235

[88] E. Y. Song, G. J. FitzPatrick, and K. B. Lee, "Smart sensors and standard-based interoperability in smart grids," *IEEE Sensors Journal*, vol. 17, no. 23, pp. 7723–7730, 2017.

[89] M. I. Oozeer and S. Haykin, "Cognitive dynamic system for control and cyber-attack detection in smart grid," *IEEE Access*, vol. 7, pp. 78 320–78 335, 2019.

[90] S. Feng and S. Haykin, "Cognitive dynamic system for future race vehicles in smart cities: A risk control perspective," *IEEE Internet of Things Magazine*, vol. 2, no. 1, pp. 14–20, 2019, doi: 10.1109/IOTM.2019.1900008.

[91] P. Vlacheas, R. Giaffreda, V. Stavroulaki, D. Kelaidonis, V. Foteinos, G. Poulios, P. Demestichas, A. Somov, A. R. Biswas, and K. Moessner, "Enabling smart cities through a cognitive management framework for the internet of things," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 102–111, 2013, doi: 10.1109/MCOM.2013.6525602.

[92] J.-h. Park, M. M. Salim, J. H. Jo, J. C. S. Sicato, S. Rathore, and J. H. Park, "Ciot-net: a scalable cognitive iot based smart city network architecture," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, p. 29, Aug 2019, doi: 10.1186/s13673-019-0190-9.

[93] Y. C. Pranaya, M. N. Himarish, M. N. Baig, and M. R. Ahmed, "Cognitive architecture based smart grids for smart cities," in *2017 3rd International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET)*, 2017, pp. 44–49, doi: 10.1109/PGSRET.2017.8251799.

[94] D. N. Molokomme, C. S. Chabalala, and P. N. Bokoro, "A review of cognitive radio smart grid communication infrastructure systems," *Energies*, vol. 13, no. 12, 2020, doi: 10.3390/en13123245.

[95] M. Ozger, O. Cetinkaya, and O. B. Akan, "Energy harvesting cognitive radio networking for iot-enabled smart grid," *Mobile Networks and Applications*, vol. 23, no. 4, pp. 956–966, Aug 2018, doi: 10.1007/s11036-017-0961-3.

[96] M. S. Farag, R. Ahmed, S. A. Gadsden, S. R. Habibi, and J. Tjong, "A comparative study of li-ion battery models and nonlinear dual estimation strategies," in *2012 IEEE Transportation Electrification Conference and Expo (ITEC)*, 2012, pp. 1–8.

[97] S. Guo and X. Zhao, "Deep reinforcement learning optimal transmission algorithm for cognitive internet of things with rf energy harvesting," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1216–1227, 2022, doi: 10.1109/TCCN.2022.3142727.

[98] R. Perez-Torres, C. Torres-Huitzil, and H. Galeana-Zapien, "An on-device cognitive dynamic systems inspired sensing framework for the iot," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 154–161, 2018, doi: 10.1109/MCOM.2018.1700224.

[99] R. Pérez-Torres, C. Torres-Huitzil, and H. Galeana-Zapién, "A cognitive-inspired event-based control for power-aware human mobility analysis in iot devices," *Sensors*, vol. 19, no. 4, 2019, doi: 10.3390/s19040832.

[100] F. M. Al-Turjman, "Information-centric sensor networks for cognitive iot: an overview," *Annals of Telecommunications*, vol. 72, no. 1, pp. 3–18, Feb 2017, doi: 10.1007/s12243-016-0533-8.

[101] J. Ploennigs, A. Ba, and M. Barry, "Materializing the promises of cognitive iot: How cognitive buildings are shaping the way," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2367–2374, 2018, doi: 10.1109/JIOT.2017.2755376.

[102] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive internet of things: Improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58–64, 2019, doi: 10.1109/MNET.2019.1800344.

[103] F. Foukalas, "Cognitive iot platform for fog computing industrial applications," *Computers & Electrical Engineering*, vol. 87, p. 106770, 2020, doi: 10.1016/j.compeleceng.2020.106770.

[104] E. B. Varghese and S. M. Thampi, "Application of cognitive computing for smart crowd management," *IT Professional*, vol. 22, no. 4, pp. 43–50, 2020, doi: 10.1109/MITP.2020.2985974.

[105] E. B. Varghese and S. M. Thampi, "A cognitive iot smart surveillance framework for crowd behavior analysis," in *2021 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, 2021, pp. 360–362, doi: 10.1109/COMSNETS51098.2021.9352910.

[106] M. Chen, J. Yang, X. Zhu, X. Wang, M. Liu, and J. Song, "Smart home 2.0: Innovative smart home system powered by botanical iot and emotion detection," *Mobile Networks and Applications*, vol. 22, 12 2017, doi: 10.1007/s11036-017-0866-1.

[107] M. A. Garrett, "Big data analytics and cognitive computing – future opportunities for astronomical research," *IOP Conference Series: Materials Science and Engineering*, vol. 67, no. 1, p. 012017, oct 2014, doi: 10.1088/1757-899X/67/1/012017.

[108] M. A. Garrett, "Seti reloaded: Next generation radio telescopes, transients and cognitive computing," *Acta Astronautica*, vol. 113, pp. 8–12, 2015, doi: 10.1016/j.actaastro.2015.03.013.

[109] X. Liu, M. Jia, M. Zhou, B. Wang, and T. S. Durrani, "Integrated cooperative spectrum sensing and access control for cognitive industrial internet of things," *IEEE Internet of Things Journal*, vol. 10, no. 3, pp. 1887–1896, 2023, doi: 10.1109/JIOT.2021.3137408.

[110] S. M. Nagarajan, G. G. Deverajan, P. Chatterjee, W. Alnumay, and U. Ghosh, "Effective task scheduling algorithm with deep learning for internet of health

things (ioht) in sustainable smart cities," *Sustainable Cities and Society*, vol. 71, p. 102945, 2021, doi: 10.1016/j.scs.2021.102945.

[111] M. Tahir, Q. Mamoon Ashraf, and M. Dabbagh, "Towards enabling autonomic computing in iot ecosystem," in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2019, pp. 646–651, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00122.

[112] K. R. Chowdhury and M. Di Felice, "Search: A routing protocol for mobile cognitive radio ad-hoc networks," in *2009 IEEE Sarnoff Symposium*, 2009, pp. 1–6, doi: 10.1109/SARNOF.2009.4850323.

[113] W. Lu, S. Hu, X. Liu, C. He, and Y. Gong, "Incentive mechanism based cooperative spectrum sharing for ofdm cognitive iot network," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 662–672, 2020, doi: 10.1109/TNSE.2019.2917071.

[114] F. Jalali, O. J. Smith, T. Lynar, and F. Suits, "Cognitive iot gateways: Automatic task sharing and switching between cloud and edge/fog computing," in *Proceedings of the SIGCOMM Posters and Demos*, ser. SIGCOMM Posters and Demos '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 121–123, doi: 10.1145/3123878.3132008.

[115] S. Gupta, A. K. Kar, A. Baabdullah, and W. A. Al-Khowaiter, "Big data with cognitive computing: A review for the future," *International Journal of Information Management*, vol. 42, pp. 78–89, 2018, doi: 10.1016/j.ijinfomgt.2018.06.005.

[116] A. A. Cook, G. Mısırlı, and Z. Fan, "Anomaly detection for iot time-series data: A survey," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6481–6494, 2020.

[117] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: A review of anomaly detection techniques and recent advances," *Expert Systems with Applications*, vol. 193, p. 116429, 2022, doi: 10.1016/j.eswa.2021.116429.

[118] M. Bahrami and M. Singhal, *The Role of Cloud Computing Architecture in Big Data*. Springer, 07 2015, vol. 8, pp. 275–295, doi: 10.1007/978-3-319-08254-7_13.

[119] S. Hamdan, M. Ayyash, and S. Almajali, "Edge-computing architectures for internet of things applications: A survey," *Sensors*, vol. 20, no. 22, 2020, doi: 10.3390/s20226441.

[120] H. Cai, B. Xu, L. Jiang, and A. V. Vasilakos, "Iot-based big data storage systems in cloud computing: Perspectives and challenges," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 75–87, 2017, doi: 10.1109/JIOT.2016.2619369.

[121] M. Chen, W. Li, Y. Hao, Y. Qian, and I. Humar, "Edge cognitive computing based smart healthcare system," *Future Generation Computer Systems*, vol. 86, pp. 403–411, 2018, doi: 10.1016/j.future.2018.03.054.

[122] R. A. C. da Silva and N. L. S. d. Fonseca, "Resource allocation mechanism for a fog-cloud infrastructure," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6, doi: 10.1109/ICC.2018.8422237.

[123] K. Muteba, K. Djouani, and T. Olwal, "Deep reinforcement learning based resource allocation for narrowband cognitive radio-iot systems," *Procedia Computer Science*, vol. 175, pp. 315–324, 2020, doi: 10.1016/j.procs.2020.07.046.

[124] S. Tang, Z. Pan, G. Hu, Y. Wu, and Y. Li, "Deep reinforcement learning-based resource allocation for satellite internet of things with diverse qos guarantee," *Sensors*, vol. 22, no. 8, 2022, doi: 10.3390/s22082979.

[125] S. Yoon, J.-H. Cho, D. S. Kim, T. J. Moore, F. Free-Nelson, and H. Lim, "Desolater: Deep reinforcement learning-based resource allocation and moving target defense deployment framework," *IEEE Access*, vol. 9, pp. 70 700–70 714, 2021, doi: 10.1109/ACCESS.2021.3076599.

[126] Y. Sun, X. Zhang, and Y. Zhu, "Mobility and traffic prediction-based resource allocation with edge intelligence in wireless network," in *2021 13th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2021, pp. 1–6, doi: 10.1109/WCSP52459.2021.9613219.

[127] S. Latif, S. Akraam, T. Karamat, M. A. Khan, C. Altrjman, S. Mey, and Y. Nam, "An efficient pareto optimal resource allocation scheme in cognitive radio-based internet of things networks," *Sensors*, vol. 22, no. 2, 2022, doi: 10.3390/s22020451.

[128] M. Babaioff, Y. Mansour, N. Nisan, G. Noti, C. Curino, N. Ganapathy, I. Menache, O. Reingold, M. Tennenholtz, and E. Timnat, "Era: A framework for economic resource allocation for the cloud," in *Proceedings of the 26th International Conference on World Wide Web Companion*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, p. 635–642, doi: 10.1145/3041021.3054186.

[129] C.-H. Hong and B. Varghese, "Resource management in fog/edge computing: a survey on architectures, infrastructure, and algorithms," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–37, 2019.

[130] A. Mijuskovic, A. Chiumento, R. Bemthuis, A. Aldea, and P. Havinga, "Resource management techniques for cloud/fog and edge computing: An evaluation framework and classification," *Sensors*, vol. 21, no. 5, 2021, doi: 10.3390/s21051832.

[131] M. Chen, W. Li, G. Fortino, Y. Hao, L. Hu, and I. Humar, "A dynamic service migration mechanism in edge cognitive computing," *ACM Trans. Internet Technol.*, vol. 19, no. 2, apr 2019, doi: 10.1145/3239565.

[132] M. Chen and Y. Hao, "Task offloading for mobile edge computing in software defined ultra-dense network," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 587–597, 2018, doi: 10.1109/JSAC.2018.2815360.

[133] M. Chen, Y. Hao, M. Qiu, J. Song, D. Wu, and I. Humar, "Mobility-Aware caching and computation offloading in 5G Ultra-Dense cellular networks," *Sensors (Basel)*, vol. 16, no. 7, 2016.

[134] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 44–55, 2018, doi: 10.1109/TVT.2017.2760281.

[135] P. P. Ray, "A review on tinyml: State-of-the-art and prospects," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1595–1623, 2022, doi: 10.1016/j.jksuci.2021.11.019.

[136] Z. Ballard, C. Brown, A. M. Madni, and A. Ozcan, "Machine learning and computation-enabled intelligent sensor design," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 556–565, Jul 2021, doi: 10.1038/s42256-021-00360-9.

[137] H. Lim, S. Kim, K. Lee, C. Han, N. Kim, J. Ryu, C. Park, and J. Lee, "Speech recognition device and speech recognition method based on sound source direction," Aug 2022. [Online]. Available: https://patents.google.com/patent/US10984790B2/en?oq=US-10984790-B2

[138] Z. Liu, C. Li, H. Qiu, and Y. Liu, "Image processing device and method," Apr 2024. [Online]. Available: https://patents.google.com/patent/WO2020050686A1/en

[139] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[140] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," 2017. [Online]. Available: https://arxiv.org/abs/1707.01083

[141] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size," 2016. [Online]. Available: https://arxiv.org/abs/1602.07360

[142] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, 2023.

[143] H. Cai, C. Gan, L. Zhu, and S. Han, "Tinytl: Reduce activations, not trainable parameters for efficient on-device learning," 2021.

[144] T. Veiga, H. A. Asad, F. A. Kraemer, and K. Bach, "Towards containerized, reuse-oriented ai deployment platforms for cognitive iot applications," *Future Generation Computer Systems*, vol. 142, pp. 4–13, 2023, doi: 10.1016/j.future.2022.12.029.

[145] A. Giuliano, S. Andrew Gadsden, and J. Yawney, "Optimizing satellite image analysis: Leveraging variational autoencoders latent representations for direct integration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–23, 2025.

[146] A. Giuliano, S. A. Gadsden, W. Hilal, and J. Yawney, "Convolutional variational autoencoders for secure lossy image compression in remote sensing," in *Sensors and Systems for Space Applications XVII*, G. Chen and K. D. Pham, Eds., vol. 13062, International Society for Optics and Photonics. SPIE, 2024, p. 130620H. [Online]. Available: https://doi.org/10.1117/12.3013451

[147] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. Vincent Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021, doi: 10.1109/COMST.2021.3075439.

[148] T. Sun, D. Li, and B. Wang, "Decentralized federated averaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4289–4301, 2023, doi: 10.1109/TPAMI.2022.3196503.

[149] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained iot devices," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 1–24, 2022, doi: 10.1109/JIOT.2021.3095077.

[150] H. Nasiri, S. Nasehi, and M. Goudarzi, "A survey of distributed stream processing systems for smart city data analytics," in *Proceedings of the International Conference on Smart Cities and Internet of Things*, ser. SCIOT '18. New York, NY, USA: Association for Computing Machinery, 2018, doi: 10.1145/3269961.3282845.

[151] S. Shadroo and A. M. Rahmani, "Systematic survey of big data and data mining in internet of things," *Computer Networks*, vol. 139, pp. 19–47, 2018, doi: 10.1016/j.comnet.2018.04.001.

[152] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *ACM Computing Surveys*, vol. 53, no. 2, p. 1–33, Mar. 2020, doi: 10.1145/3377454.

[153] J. Wang, Y. Yang, T. Wang, R. S. Sherratt, and J. Zhang, "Big data service architecture: A survey," *Journal of Internet Technology*, vol. 21, pp. 393–405, 2020, doi: 10.3966/160792642020032102008.

[154] D. Borthakur, "The hadoop distributed file system: Architecture and design," *Hadoop Project Website*, vol. 11, no. 2007, p. 21, 2007.

[155] H. Jamil, T. Umer, C. Ceken, and F. Al-Turjman, "Decision based model for real-time iot analysis using big data and machine learning," *Wireless Personal Communications*, vol. 121, no. 4, pp. 2947–2959, Dec 2021, doi: 10.1007/s11277-021-08857-7.

[156] E. Ahmed, I. Yaqoob, I. A. T. Hashem, I. Khan, A. I. A. Ahmed, M. Imran, and A. V. Vasilakos, "The role of big data analytics in internet of things," *Computer Networks*, vol. 129, pp. 459–471, 2017, doi: 10.1016/j.comnet.2017.06.013.

[157] I. Baumgart and S. Mies, "S/kademlia: A practicable approach towards secure key-based routing," in *2007 International Conference on Parallel and Distributed Systems*, 2007, pp. 1–8, doi: 10.1109/ICPADS.2007.4447808.

[158] P. Maymounkov and D. Mazières, "Kademlia: A peer-to-peer information system based on the xor metric," in *Peer-to-Peer Systems*, P. Druschel, F. Kaashoek, and A. Rowstron, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 53–65.

[159] J. Benet, "Ipfs - content addressed, versioned, p2p file system," 2014.

[160] F. Ganz, D. Puschmann, P. Barnaghi, and F. Carrez, "A practical evaluation of information processing and abstraction techniques for the internet of things," *IEEE Internet of Things Journal*, vol. 2, no. 4, pp. 340–354, 2015, doi: 10.1109/JIOT.2015.2411227.

[161] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: Literature review and challenges," *International Journal of Distributed Sensor Networks*, vol. 11, no. 8, p. 431047, 2015, doi: 10.1155/2015/431047.

[162] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-aware computing, learning, and big data in internet of things: A survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 1–27, 2018, doi: 10.1109/JIOT.2017.2773600.

[163] A. Goyal and Y. Bengio, "Inductive biases for deep learning of higher-level cognition," 2022.

[164] J. Grafman, *Neuronal plasticity: Building a bridge from the laboratory to the clinic.* Springer Science & Business Media, 2012.

[165] M. Chen, F. Herrera, and K. Hwang, "Cognitive computing: Architecture, technologies and intelligent applications," *IEEE Access*, vol. 6, pp. 19 774–19 783, 2018, doi: 10.1109/ACCESS.2018.2791469.

[166] V. N. Gudivada, S. Pankanti, G. Seetharaman, and Y. Zhang, "Cognitive computing systems: Their potential and the future," *Computer*, vol. 52, no. 5, pp. 13–18, 2019, doi: 10.1109/MC.2019.2904940.

[167] I. Hasan and S. Rizvi, "Review of ai techniques and cognitive computing framework for intelligent decision support," in *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2021, pp. 891–898.

[168] T. Gunasekhar and M. Teja, *Cognitive Computing.* Scrivener Publishing LLC, 05 2021, pp. 189–217, doi: 10.1002/9781119711308.ch7.

[169] A. Sheth, P. Anantharam, and C. Henson, "Semantic, cognitive, and perceptual computing: Paradigms that shape human experience," *Computer*, vol. 49, no. 3, pp. 64–72, 2016, doi: 10.1109/MC.2016.75.

[170] A. G. E. Collins, "Reinforcement learning: bringing together computation and cognition," *Current Opinion in Behavioral Sciences*, vol. 29, pp. 63–68, 2019, doi: 10.1016/j.cobeha.2019.04.011.

[171] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017, doi: 10.1109/MSP.2017.2743240.

[172] A. Plaat, W. Kosters, and M. Preuss, "Deep model-based reinforcement learning for high-dimensional problems, a survey," 2020.

[173] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013.

[174] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan 2016, doi: 10.1038/nature16961.

[175] M. Al-Shabi, K. S. Hatamleh, S. A. Gadsden, B. Soudan, and A. Elnady, "Robust nonlinear control and estimation of a prrr robot system," *Int. J. Robot. Autom.*, vol. 34, no. 6, 2019.

[176] D. Abel, D. Arumugam, L. Lehnert, and M. Littman, "State abstractions for lifelong reinforcement learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 10–19. [Online]. Available: https://proceedings.mlr.press/v80/abel18a.html

[177] D. A. Ferrucci, "Introduction to "this is watson"," *IBM Journal of Research and Development*, vol. 56, no. 3.4, pp. 1:1–1:15, 2012, doi: 10.1147/JRD.2012.2184356.

[178] A. Gliozzo, O. Biran, S. Patwardhan, and K. McKeown, "Semantic technologies in ibm watson," in *Proceedings of the Fourth Workshop on Teaching NLP and CL*, 08 2013, pp. 85–92.

[179] R. E. Hoyt, D. Snider, C. Thompson, and S. Mantravadi, "IBM watson analytics: Automating visualization, descriptive, and predictive statistics," *JMIR Public Health Surveill*, vol. 2, no. 2, p. e157, Oct. 2016.

[180] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, "Building watson: An overview of the deepqa project," *AI Magazine*, vol. 31, no. 3, pp. 59–79, Jul. 2010, doi: 10.1609/aimag.v31i3.2303.

[181] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, "Watson: Beyond jeopardy!" *Artificial Intelligence*, vol. 199-200, pp. 93–105, 2013, doi: 10.1016/j.artint.2012.06.009.

[182] M. N. Ahmed, A. S. Toor, K. O'Neil, and D. Friedland, "Cognitive computing and the future of health care cognitive computing and the future of healthcare: The cognitive power of ibm watson has the potential to transform global personalized medicine," *IEEE Pulse*, vol. 8, no. 3, pp. 4–9, 2017, doi: 10.1109/MPUL.2017.2678098.

[183] Y. Chen, J. Elenee Argentinis, and G. Weber, "Ibm watson: How cognitive computing can be applied to big data challenges in life sciences research," *Clinical Therapeutics*, vol. 38, no. 4, pp. 688–701, 2016, doi: 10.1016/j.clinthera.2015.12.001.

[184] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[185] K. Lu, A. Grover, P. Abbeel, and I. Mordatch, "Pretrained transformers as universal computation engines," 2021.

[186] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[187] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[188] e. a. Tom Brown, "Language models are few-shot learners," 2020.

[189] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd,

A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, J. Yoo, and P. Yu, "GPT-4 Technical

Report," Mar. 2024. [Online]. Available: http://arxiv.org/abs/2303.08774

[190] e. a. Aakanksha Chowdhery, "Palm: Scaling language modeling with pathways," 2022.

[191] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu, "PaLM 2 Technical Report," Sep. 2023, arXiv:2305.10403 [cs]. [Online]. Available: http://arxiv.org/abs/2305.10403

[192] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, arXiv:2302.13971 [cs]. [Online]. Available: http://arxiv.org/abs/2302.13971

[193] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zhang, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, and B. Catanzaro, "Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model," Feb. 2022, arXiv:2201.11990 [cs]. [Online]. Available: http://arxiv.org/abs/2201.11990

[194] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman,

A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovizh, E. Kim, E. B. Natan, F. De Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. Elsahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim, J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. Von Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy, M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina, V. Laippala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, D. E. Taşar, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H. Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang, O. Press, C. Li, D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névéol, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Mirkin, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger,

Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Unldreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. McDuff, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynok, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrimann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sänger, M. Samwald, M. Cullan, M. Weinberg, M. De Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sang-aroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, and T. Wolf, "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," Jun. 2023, arXiv:2211.05100 [cs]. [Online]. Available: http://arxiv.org/abs/2211.05100

[195] DeepSeek-AI, :, X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, H. Gao, K. Gao, W. Gao, R. Ge, K. Guan, D. Guo, J. Guo, G. Hao, Z. Hao, Y. He, W. Hu, P. Huang, E. Li, G. Li, J. Li, Y. Li, Y. K. Li, W. Liang, F. Lin, A. X. Liu, B. Liu, W. Liu, X. Liu, X. Liu, Y. Liu, H. Lu, S. Lu, F. Luo, S. Ma, X. Nie, T. Pei, Y. Piao, J. Qiu, H. Qu, T. Ren, Z. Ren, C. Ruan, Z. Sha, Z. Shao, J. Song, X. Su, J. Sun, Y. Sun, M. Tang, B. Wang, P. Wang, S. Wang, Y. Wang, Y. Wang, T. Wu, Y. Wu, X. Xie, Z. Xie, Z. Xie, Y. Xiong, H. Xu, R. X. Xu, Y. Xu, D. Yang, Y. You, S. Yu, X. Yu, B. Zhang, H. Zhang, L. Zhang, L. Zhang, M. Zhang, M. Zhang, W. Zhang, Y. Zhang, C. Zhao, Y. Zhao, S. Zhou, S. Zhou, Q. Zhu, and Y. Zou, "Deepseek llm: Scaling open-source language models with longtermism," 2024. [Online].

Available: https://arxiv.org/abs/2401.02954

[196] Y. Wu, M. Rabe, W. Li, J. Ba, R. Grosse, and C. Szegedy, "Lime: Learning inductive bias for primitives of mathematical reasoning," 2022.

[197] A. Tamkin, V. Liu, R. Lu, D. Fein, C. Schultz, and N. Goodman, "Dabs: A domain-agnostic benchmark for self-supervised learning," 2023.

[198] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," 2021.

[199] E. Parisotto, H. F. Song, J. W. Rae, R. Pascanu, C. Gulcehre, S. M. Jayakumar, M. Jaderberg, R. L. Kaufman, A. Clark, S. Noury, M. M. Botvinick, N. Heess, and R. Hadsell, "Stabilizing transformers for reinforcement learning," 2019.

[200] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2022.

[201] U. Upadhyay, N. Shah, S. Ravikanti, and M. Medhe, "Transformer based reinforcement learning for games," 2019.

[202] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015, doi: 10.1109/JPROC.2015.2460697.

[203] A. Vakil, J. Liu, P. Zulch, E. Blasch, R. Ewing, and J. Li, "A survey of multimodal sensor fusion for passive rf and eo information integration," *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 7, pp. 44–61, 2021, doi: 10.1109/MAES.2020.3006410.

[204] H. Durrant-Whyte and T. C. Henderson, *Multisensor Data Fusion.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 585–610, doi: 10.1007/978-3-540-30301-5_26,.

[205] K. P. Murphy, *Probabilistic machine learning: an introduction.* MIT press, 2022.

[206] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: theory algorithms and software.* John Wiley & Sons, 2004.

[207] M. Al-Shabi, A. Cataford, and S. A. Gadsden, "Quadrature kalman filters with applications to robotic manipulators," in *2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS)*, 2017, pp. 117–124.

[208] A. Gadsden, S. Habibi, D. Dunne, and T. Kirubarajan, "Nonlinear estimation techniques applied on target tracking problems," *J. Dyn. Syst. Meas. Control*, vol. 134, no. 5, p. 054501, Sep. 2012.

[209] S. Y. Nathan Gaw and M. R. Gahrooei, "Multimodal data fusion for systems improvement: A review," *IISE Transactions*, vol. 54, no. 11, pp. 1098–1116, 2022, doi: 10.1080/24725854.2021.1987593.

[210] T. Hoang, N. Fahier, and W.-C. Fang, "Multi-leads ecg premature ventricular contraction detection using tensor decomposition and convolutional neural network," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4, doi: 10.1109/BIOCAS.2019.8919049.

[211] R. K. Kaliyar, A. Goswami, and P. Narang, "Deepfake: improving fake news detection using tensor decomposition-based deep neural network," *The Journal of Supercomputing*, vol. 77, no. 2, pp. 1015–1037, Feb 2021, doi: 10.1007/s11227-020-03294-y.

[212] P. Zhou, B. Gao, C. Zhao, and T. Chai, "Heterogeneous data-driven measurement method for feo content of sinter based on deep learning and tensor decomposition," *Control Engineering Practice*, vol. 134, p. 105479, 2023, doi: 10.1016/j.conengprac.2023.105479.

[213] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: a review," *Briefings in Bioinformatics*, vol. 23, no. 2, p. bbab569, 01 2022, doi: 10.1093/bib/bbab569.

[214] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *npj Digital Medicine*, vol. 3, no. 1, p. 136, Oct 2020, doi: 10.1038/s41746-020-00341-z.

[215] K. S. F. Azam, O. Ryabchykov, and T. Bocklitz, "A review on data fusion of multidimensional medical and biomedical data," *Molecules*, vol. 27, no. 21, 2022, doi: 10.3390/molecules27217448. [Online]. Available: https://www.mdpi.com/1420-3049/27/21/7448

[216] Y. Zhang, M. Sheng, X. Liu, R. Wang, W. Lin, P. Ren, X. Wang, E. Zhao, and W. Song, "A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration," *Health Information Science and Systems*, vol. 10, no. 1, p. 22, Aug 2022, doi: 10.1007/s13755-022-00183-x.

[217] B. Hunyadi, P. Dupont, W. Van Paesschen, and S. Van Huffel, "Tensor decompositions and data fusion in epileptic electroencephalography and functional magnetic resonance imaging data," *WIREs Data Mining and Knowledge Discovery*, vol. 7, no. 1, p. e1197, 2017, doi: 10.1002/widm.1197.

[218] A. Newton, A. McCafferty-Leroux, S. A. Gadsden, and K. R. Turpie, "Towards a second-generation robotic telescope mount for the air-LUSI instrument," in *Sensors and Systems for Space Applications XVI*, K. D. Pham and G. Chen, Eds. Orlando, United States: SPIE, Jun. 2023, p. 17. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12546/2663887/Towards-a-second-generation-robotic-telescope-mount-for-the-air/10.1117/12.2663887.full

[219] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022, doi: 10.1016/j.jag.2022.102926.

[220] R. H. C. P. G. V. J. C. N. A. B. S. L. B. L. Saux, "Data fusion contest 2022 (dfc2022)," 2022, doi: 10.21227/rjv6-f516.

[221] C. Persello, R. Hänsch, G. Vivone, K. Chen, Z. Yan, D. Tang, H. Huang, M. Schmitt, and X. Sun, "2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 1, pp. 94–97, 2023, doi: 10.1109/MGRS.2023.3240233.

[222] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, 2021, doi: 10.1109/TITS.2019.2962338.

[223] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, 2021, doi: 10.3390/s21062140.

[224] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran, "Deep learning sensor fusion for autonomous vehicle perception and localization: A review," *Sensors*, vol. 20, no. 15, 2020, doi: 10.3390/s20154220.

[225] P. Wang, L. T. Yang, J. Li, J. Chen, and S. Hu, "Data fusion in cyber-physical-social systems: State-of-the-art and perspectives," *Information Fusion*, vol. 51, pp. 42–57, 2019, doi: 10.1016/j.inffus.2018.11.002.

[226] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, p. 478–493, Mar. 2020, doi: 10.1109/jstsp.2020.2987728.

[227] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," 2021.

[228] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," 2024. [Online]. Available: https://arxiv.org/abs/2404.08471

[229] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," 2023.

[230] Q. Butler, Y. Ziada, D. Stephenson, and S. Andrew Gadsden, "Condition monitoring of machine tool feed drives: A review," *J. Manuf. Sci. Eng.*, vol. 144, no. 10, pp. 1–43, Oct. 2022.

[231] Z. Yu, Z. Wang, Y. Li, H. You, R. Gao, X. Zhou, S. R. Bommu, Y. K. Zhao, and Y. C. Lin, "Edge-llm: Enabling efficient large language model adaptation on edge devices via layerwise unified compression and adaptive layer tuning and voting," 2024. [Online]. Available: https://arxiv.org/abs/2406.15758

[232] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," 2024. [Online]. Available: https://arxiv.org/abs/2308.07633

[233] Z. Huang, Q. Min, H. Huang, D. Zhu, Y. Zeng, R. Guo, and X. Zhou, "Ultra-sparse memory network," 2025. [Online]. Available: https://arxiv.org/abs/2411.12364

[234] Y. Li, Z. Li, Z. Han, Q. Zhang, and X. Ma, "Automating cloud deployment for real-time online foundation model inference," *IEEE/ACM Transactions on Networking*, vol. 32, no. 2, pp. 1509–1523, 2024.

[235] Z. Fu, W. Lam, Q. Yu, A. M.-C. So, S. Hu, Z. Liu, and N. Collier, "Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder," 2023. [Online]. Available: https://arxiv.org/abs/2304.04052

[236] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," 2018. [Online]. Available: https://arxiv.org/abs/1801.10198

[237] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, and T. Gui, "The rise and potential of large language model based agents: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2309.07864

[238] T. Wang, J. Fan, and P. Zheng, "An llm-based vision and language cobot navigation approach for human-centric smart manufacturing," *Journal of Manufacturing Systems*, vol. 75, pp. 299–305, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612524000864

[239] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2024, pp. 902–909.

[240] R. Chen, W. Song, W. Zu, Z. Dong, Z. Guo, F. Sun, Z. Tian, and J. Wang, "An llm-driven framework for multiple-vehicle dispatching and navigation in smart city landscapes," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 2147–2153.

[241] J. Qiu, K. Lam, G. Li, A. Acharya, T. Y. Wong, A. Darzi, W. Yuan, and E. J. Topol, "LLM-based agentic systems in medicine and healthcare," *Nat. Mach. Intell.*, vol. 6, no. 12, pp. 1418–1420, Dec. 2024.

[242] B. Li, K. Mellou, B. Zhang, J. Pathuri, and I. Menache, "Large language models for supply chain optimization," 2023. [Online]. Available: https://arxiv.org/abs/2307.03875

[243] T. Xiao and J. Zhu, "Foundations of large language models," 2025. [Online]. Available: https://arxiv.org/abs/2501.09223

[244] I. Khan, X. Zhang, M. Rehman, and R. Ali, "A literature survey and empirical study of meta-learning for classifier selection," *IEEE Access*, vol. 8, pp. 10 262–10 281, 2020, doi: 10.1109/ACCESS.2020.2964726.

[245] Y. Ma, S. Zhao, W. Wang, Y. Li, and I. King, "Multimodality in meta-learning: A comprehensive survey," *Knowledge-Based Systems*, vol. 250, p. 108976, 2022, doi: 10.1016/j.knosys.2022.108976.

[246] H. Peng, "A comprehensive overview and survey of recent advances in meta-learning," 2020.

[247] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2402.06196

[248] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2303.18223

[249] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2308.10792

[250] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-efficient fine-tuning for large models: A comprehensive survey," 2024. [Online]. Available: https://arxiv.org/abs/2403.14608

[251] K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, C. Tang, C. Wang, D. Zhang, E. Yuan, E. Lu, F. Tang, F. Sung, G. Wei, G. Lai, H. Guo, H. Zhu, H. Ding, H. Hu, H. Yang, H. Zhang, H. Yao, H. Zhao, H. Lu, H. Li, H. Yu, H. Gao, H. Zheng, H. Yuan, J. Chen, J. Guo, J. Su, J. Wang, J. Zhao, J. Zhang, J. Liu, J. Yan, J. Wu, L. Shi, L. Ye, L. Yu, M. Dong, N. Zhang, N. Ma, Q. Pan, Q. Gong, S. Liu, S. Ma, S. Wei, S. Cao, S. Huang, T. Jiang, W. Gao, W. Xiong, W. He, W. Huang, W. Wu, W. He, X. Wei, X. Jia, X. Wu, X. Xu, X. Zu, X. Zhou, X. Pan, Y. Charles, Y. Li, Y. Hu, Y. Liu, Y. Chen, Y. Wang, Y. Liu, Y. Qin, Y. Liu, Y. Yang, Y. Bao, Y. Du,

Y. Wu, Y. Wang, Z. Zhou, Z. Wang, Z. Li, Z. Zhu, Z. Zhang, Z. Wang, Z. Yang, Z. Huang, Z. Huang, Z. Xu, and Z. Yang, "Kimi k1.5: Scaling reinforcement learning with llms," 2025. [Online]. Available: https://arxiv.org/abs/2501.12599

[252] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, Jun. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271619301108

[253] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Information Fusion*, vol. 42, pp. 158–173, Jul. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253517305936

[254] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying Corresponding Patches in SAR and Optical Images With a Pseudo-Siamese CNN," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 784–788, May 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8314449/

[255] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When Deep Learning Meets Metric Learning: Remote Sensing Image Scene Classification via Learning Discriminative CNNs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, May 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8252784/

[256] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 3–22, Nov. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271618301096

[257] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models," *Remote Sensing of Environment*, vol. 237, p. 111322, Feb. 2020, arXiv:1807.05713 [cs]. [Online]. Available: http://arxiv.org/abs/1807.05713

[258] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 60–77, Nov. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271618301229

[259] L. Ma, M. Li, X. Ma, L. Cheng, P. Du, and Y. Liu, "A review of supervised object-based land-cover image classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 277–293, Aug. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S092427161630661X

[260] Y. Yang, S. Mandt, and L. Theis, "An Introduction to Neural Data Compression," Aug. 2023, arXiv:2202.06533 [cs, eess, math]. [Online]. Available: http://arxiv.org/abs/2202.06533

[261] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end Optimized Image Compression," Mar. 2017, arXiv:1611.01704 [cs, math]. [Online]. Available: http://arxiv.org/abs/1611.01704

[262] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules," Mar. 2020, arXiv:2001.01568 [eess]. [Online]. Available: http://arxiv.org/abs/2001.01568

[263] V. Alves De Oliveira, M. Chabert, T. Oberlin, C. Poulliat, M. Bruno, C. Latry, M. Carlavan, S. Henrot, F. Falzon, and R. Camarero, "Reduced-Complexity End-to-End Variational Autoencoder for on Board Satellite Image Compression," *Remote Sensing*, vol. 13, no. 3, p. 447, Jan. 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/3/447

[264] C. Li, B. Zhang, D. Hong, J. Zhou, G. Vivone, S. Li, and J. Chanussot, "CasFormer: Cascaded transformers for fusion-aware computational hyperspectral imaging," vol. 108, p. 102408. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253524001866

[265] C. Li, B. Zhang, D. Hong, X. Jia, A. Plaza, and J. Chanussot, "Learning disentangled priors for hyperspectral anomaly detection: A coupling model-driven and data-driven paradigm," pp. 1–14. [Online]. Available: https://ieeexplore.ieee.org/document/10547283/

[266] C. Li, B. Zhang, D. Hong, J. Yao, X. Jia, A. Plaza, and J. Chanussot, "Interpretable networks for hyperspectral anomaly detection: A deep unfolding solution," pp. 1–1. [Online]. Available: https://ieeexplore.ieee.org/document/10613787/

[267] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, and X. X. Zhu, "Cross-city matters: A multimodal remote sensing

benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," vol. 299, p. 113856. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0034425723004078

[268] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, P. Gamba, J. A. Benediktsson, and J. Chanussot, "SpectralGPT: Spectral remote sensing foundation model," vol. 46, no. 8, pp. 5227–5244. [Online]. Available: https://ieeexplore.ieee.org/document/10490262/

[269] A. Robinson and C. Cherry, "Results of a prototype television bandwidth compression scheme," *Proceedings of the IEEE*, vol. 55, no. 3, pp. 356–364, 1967. [Online]. Available: http://ieeexplore.ieee.org/document/1447423/

[270] D. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952. [Online]. Available: http://ieeexplore.ieee.org/document/4051119/

[271] Welch, "A Technique for High-Performance Data Compression," *Computer*, vol. 17, no. 6, pp. 8–19, Jun. 1984. [Online]. Available: http://ieeexplore.ieee.org/document/1659158/

[272] X. Yu, J. Zhao, T. Zhu, Q. Lan, L. Gao, and L. Fan, "Analysis of JPEG2000 Compression Quality of Optical Satellite Images," in *2022 2nd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*. Shenyang, China: IEEE, Feb. 2022, pp. 500–503. [Online]. Available: https://ieeexplore.ieee.org/document/9820991/

[273] S.-e. Qian, M. Bergeron, I. Cunningham, L. Gagnon, and A. Hollinger, "Near lossless data compression onboard a hyperspectral satellite," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 3, pp. 851–866, Jul. 2006. [Online]. Available: http://ieeexplore.ieee.org/document/4014456/

[274] V. Sanchez, F. Auli-Llinas, and J. Serra-Sagrista, "DPCM-Based Edge Prediction for Lossless Screen Content Coding in HEVC," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 4, pp. 497–507, Dec. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7579210/

[275] F. Mentzer, E. Agustsson, J. Ballé, D. Minnen, N. Johnston, and G. Toderici, "Neural Video Compression using GANs for Detail Synthesis and Propagation," Jul. 2022, arXiv:2107.12038 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2107.12038

[276] L. Helminger, A. Djelouah, M. Gross, and C. Schroers, "Lossy Image Compression with Normalizing Flows," Aug. 2020, arXiv:2008.10486 [cs]. [Online]. Available: http://arxiv.org/abs/2008.10486

[277] C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville, "Neural Autoregressive Flows."

[278] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," May 2018, arXiv:1802.01436 [cs, eess, math]. [Online]. Available: http://arxiv.org/abs/1802.01436

[279] Z. Duan, M. Lu, Z. Ma, and F. Zhu, "Lossy Image Compression with Quantized Hierarchical VAEs," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 198–207, arXiv:2208.13056 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2208.13056

[280] J. Liu, H. Sun, and J. Katto, "Learned Image Compression with Mixed Transformer-CNN Architectures," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 14 388–14 397. [Online]. Available: https://ieeexplore.ieee.org/document/10204195/

[281] R. Yang and S. Mandt, "Lossy Image Compression with Conditional Diffusion Models."

[282] P. Bacchus, R. Fraisse, A. Roumy, and C. Guillemot, "Quasi Lossless Satellite Image Compression," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 1532–1535. [Online]. Available: https://ieeexplore.ieee.org/document/9883135/

[283] G. Guerrisi, F. Del Frate, and G. Schiavon, "Convolutional Autoencoder Algorithm for On-Board Image Compression," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 151–154. [Online]. Available: https://ieeexplore.ieee.org/document/9883256/

[284] D. Minnen, J. Ballé, and G. Toderici, "Joint Autoregressive and Hierarchical Priors for Learned Image Compression," Sep. 2018, arXiv:1809.02736 [cs]. [Online]. Available: http://arxiv.org/abs/1809.02736

[285] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK," 2017.

[286] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[287] X. Luo, H. Talebi, F. Yang, M. Elad, and P. Milanfar, "The Rate-Distortion-Accuracy Tradeoff: JPEG Case Study," Aug. 2020, arXiv:2008.00605 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2008.00605

[288] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video Enhancement with Task-Oriented Flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019. [Online]. Available: http://link.springer.com/10.1007/s11263-018-01144-2

[289] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," vol. 12, no. 7, pp. 2217–2226. [Online]. Available: https://ieeexplore.ieee.org/document/8736785/

[290] H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, and L. Zhao, "Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, p. 1594, 2020.

[291] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," vol. 145, pp. 197–209. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271618300042

[292] M. Schmitt, F. Tupin, and X. X. Zhu, "Fusion of sar and optical remote sensing data — challenges and recent trends," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017, pp. 5458–5461.

[293] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for SAR and optical images: A review," *Information Fusion*, vol. 59, pp. 13–29, Jul. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253519305470

[294] A. Singh and K. Gaurav, "Deep learning and data fusion to estimate surface soil moisture from multi-sensor satellite images," *Scientific Reports*, vol. 13, no. 1, p. 2251, Feb. 2023. [Online]. Available: https://doi.org/10.1038/s41598-023-28939-9

[295] M. R. Metwalli, A. H. Nasr, O. S. F. Allah, S. El-Rabaie, and F. E. A. El-Samie, "Satellite image fusion based on principal component analysis and high-pass filtering," *J. Opt. Soc. Am. A*, vol. 27, no. 6, pp. 1385–1394, Jun 2010. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?URI=josaa-27-6-1385

[296] S. Hong, W. Moon, H.-Y. Paik, and G.-H. Choi, "Data fusion of multiple polarimetric sar images using discrete wavelet transform (dwt)," in *IEEE International Geoscience and Remote Sensing Symposium*, vol. 6, 2002, pp. 3323–3325 vol.6.

[297] W. Han, X. Zhang, Y. Wang, L. Wang, X. Huang, J. Li, S. Wang, W. Chen, X. Li, R. Feng, R. Fan, X. Zhang, and Y. Wang, "A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 87–113, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271623001582

[298] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022, doi: 10.1016/j.jag.2022.102926.

[299] R. H. C. P. G. V. J. C. N. A. B. S. L. B. L. Saux, "Data fusion contest 2022 (dfc2022)," 2022, doi: 10.21227/rjv6-f516.

[300] C. Persello, R. Hänsch, G. Vivone, K. Chen, Z. Yan, D. Tang, H. Huang, M. Schmitt, and X. Sun, "2023 ieee grss data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 1, pp. 94–97, 2023, doi: 10.1109/MGRS.2023.3240233.

[301] D. Lahat, T. Adali, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7214350/

[302] Z. Liu, B. Qu, and J. Guo, "Target recognition of SAR images using fused deep feature by multiset canonical correlations analysis," *Optik*, vol. 220, p. 165156, Oct. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S003040262030992X

[303] Y. Tang and J. Chen, "A multi-view SAR target recognition method using feature fusion and joint classification," *Remote Sensing Letters*, vol. 13, no. 6, pp. 631–642, Jun. 2022. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/2150704X.2022.2063038

[304] J. Sandak, A. Sandak, and M. Cocchi, "Multi-sensor data fusion and parallel factor analysis reveals kinetics of wood weathering," *Talanta*, vol. 225, p. 122024, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0039914020313151

[305] M. Wang, D. Hong, Z. Han, J. Li, J. Yao, L. Gao, B. Zhang, and J. Chanussot, "Tensor decompositions for hyperspectral data processing in remote sensing: A comprehensive review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 1, pp. 26–72, 2023.

[306] A. Li, "Deep learning for multimodal data fusion: A survey," *Journal Name*, vol. 1, pp. 1–12, 2022.

[307] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, pp. 829–864, 2020.

[308] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, vol. 105, p. 104042, 2021.

[309] H. Akbari *et al.*, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," in *NeurIPS*, 2021.

[310] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal Learning With Transformers: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, Oct. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10123038/

[311] Y. Shi, N. Siddharth, B. Paige, and P. H. S. Torr, "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models," Nov. 2019, arXiv:1911.03393 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1911.03393

[312] M. Schmitt, L. H. Hughes, and X. X. Zhu, "The sen1-2 dataset for deep learning in sar-optical data fusion," 2018. [Online]. Available: https://arxiv.org/abs/1807.01569

[313] M. Abdul Aziz, S. M. Mohtasim, and R. Ahammed, "State-of-the-art recent developments of large magnetorheological (MR) dampers: a review," *Korea-Australia Rheology Journal*, vol. 34, no. 2, pp. 105–136, May 2022. [Online]. Available: https://link.springer.com/10.1007/s13367-022-00021-2

[314] M. Rahman, Z. C. Ong, S. Julai, M. M. Ferdaus, and R. Ahamed, "A review of advances in magnetorheological dampers: their design optimization and applications," *Journal of Zhejiang University-SCIENCE A*, vol. 18, no. 12, pp. 991–1010, Dec. 2017. [Online]. Available: http://link.springer.com/10.1631/jzus.A1600721

[315] S. Yaghoubi and A. Ghanbarzadeh, "Modeling and optimization of car suspension system in the presence of magnetorheological damper using Simulink-PSO hybrid technique," *Results in Engineering*, vol. 22, p. 102065, Jun. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2590123024003190

[316] G. Yao, F. Yap, G. Chen, W. Li, and S. Yeo, "Mr damper and its application for semi-active control of vehicle suspension system," *Mechatronics*, vol. 12, no. 7, pp. 963–973, 2002.

[317] C. Guo, X. Gong, L. Zong, C. Peng, and S. Xuan, "Twin-tube- and bypass-containing magneto-rheological damper for use in railway vehicles," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 229, no. 1, pp. 48–57, Jan. 2015. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0954409713497199

[318] Y. Kim, R. Langari, and S. Hurlebaus, "Semiactive nonlinear control of a building with a magnetorheological damper system," *Mechanical Systems and Signal Processing*, vol. 23, no. 2, pp. 300–315, Feb. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0888327008001465

[319] H.-S. Kim and J.-W. Kang, "Smart outrigger damper system for response reduction of tall buildings subjected to wind and seismic excitations," *International Journal of Steel Structures*, vol. 17, no. 4, pp. 1263–1272, Dec. 2017. [Online]. Available: http://link.springer.com/10.1007/s13296-017-1201-1

[320] S. Seid, S. Chandramohan, and S. Sujatha, "Optimal design of an MR damper valve for prosthetic knee application," *Journal of Mechanical Science and Technology*, vol. 32, no. 6, pp. 2959–2965, Jun. 2018. [Online]. Available: http://link.springer.com/10.1007/s12206-018-0552-7

[321] S. Javadinasab Hormozabad and A. K. Ghorbani-Tanha, "Semi-active fuzzy control of Lali Cable-Stayed Bridge using MR dampers under seismic excitation," *Frontiers of Structural and Civil Engineering*, vol. 14, no. 3, pp. 706–721, Jun. 2020. [Online]. Available: https://link.springer.com/10.1007/s11709-020-0612-9

[322] J. W. Chong, Y. Kim, and K. H. Chon, "Nonlinear multiclass support vector machine–based health monitoring system for buildings employing magnetorheological dampers," *Journal of Intelligent Material Systems and Structures*, vol. 25, no. 12, pp. 1456–1468, Aug. 2014. [Online]. Available: https://journals.sagepub.com/doi/10.1177/1045389X13507343

[323] H.-B. Yun and S. F. Masri, "Stochastic change detection in uncertain nonlinear systems using reduced-order models: classification," *Smart Materials and Structures*, vol. 18, no. 1, p. 015004, Jan. 2009. [Online]. Available: https://iopscience.iop.org/article/10.1088/0964-1726/18/1/015004

[324] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Jul. 2009. [Online]. Available: https://dl.acm.org/doi/10.1145/1541880.1541882

[325] Z. Wang, C. Pei, M. Ma, X. Wang, Z. Li, D. Pei, S. Rajmohan, D. Zhang, Q. Lin, H. Zhang, J. Li, and G. Xie, "Revisiting VAE for Unsupervised Time Series Anomaly Detection: A Frequency Perspective," in *Proceedings of the ACM Web Conference 2024*. Singapore Singapore: ACM, May 2024, pp. 3096–3105. [Online]. Available: https://dl.acm.org/doi/10.1145/3589334.3645710

[326] S. Lin, R. Clark, R. Birke, S. Schonborn, N. Trigoni, and S. Roberts, "Anomaly Detection for Time Series Using VAE-LSTM Hybrid Model," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, May 2020, pp. 4322–4326. [Online]. Available: https://ieeexplore.ieee.org/document/9053558/

[327] Z. Zhu, P. Su, S. Zhong, J. Huang, S. Ottikkutti, K. N. Tahmasebi, Z. Zou, L. Zheng, and D. Chen, "Using a VAE-SOM architecture for anomaly detection of flexible sensors in limb prosthesis," *Journal of Industrial Information Integration*, vol. 35, p. 100490, Oct. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2452414X23000638

[328] Y. Shi, B. Wang, Y. Yu, X. Tang, C. Huang, and J. Dong, "Robust anomaly detection for multivariate time series through temporal GCNs and attention-based VAE," *Knowledge-Based Systems*, vol. 275, p. 110725, Sep. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950705123004756

[329] Z. Yang, M. Xu, S. Wang, J. Li, Z. Peng, F. Jin, and Y. Yang, "Detection of wind turbine blade abnormalities through a deep learning model integrating VAE and neural ODE," *Ocean Engineering*, vol. 302, p. 117689, Jun. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0029801824010266

[330] L. Corporation, "Lord-MAR — shoplordmr.com," https://www.shoplordmr.com/mr-products/rd-8041-1-mr-damper-long-stroke, [Accessed 10-10-2024].

[331] A. S. Lee, Y. Wu, S. A. Gadsden, and M. AlShabi, "Interacting multiple model estimators for fault detection in a magnetorheological damper," *Sensors*, vol. 24, no. 1, p. 251, 2023.

[332] A. S. Lee, S. A. Gadsden, and M. Al-Shabi, "Application of nonlinear estimation strategies on a magnetorheological suspension system with skyhook control," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2020, pp. 1–6.

[333] L. Korad Technology CO., "KD SERIES_DONGGUAN KORAD TECHNOLOGY CO., LTD._Programmable DC Power Supplies and Electronic Loads — koradtechnology.com," https://www.koradtechnology.com/product/84.html#, [Accessed 07-06-2024].

[334] U. Motion, "Servo cylinder configurator; ultra motion — ultramotion.com," https://www.ultramotion.com/servo-cylinder-configurator/, [Accessed 07-06-2024].

[335] L. Sensors, "RAS1 — loadstarsensors.com," https://www.loadstarsensors.com/ras1.html, [Accessed 07-06-2024].