AN INTERPRETABLE AI-DRIVEN FRAMEWORK FOR MONITORING BEHAVIOR CHANGES IN CARE ENVIRONMENTS

By FATEME AKBARI, MSc

McMaster University

DOCTOR OF PHILOSOPHY IN BUSINESS ADMINISTRATION (2025)

Hamilton, Ontario, Canada (DeGroote School of Business)

TITLE:	An Interpretable AI-Driven Framework for Monitoring Behavior Changes in Care Environments
AUTHOR:	Fateme Akbari MSc (Computer Science), University of Tehran, Tehran, Iran
SUPERVISOR:	Dr. Yufei Yuan Dr. Kamran Sartipi
NUMBER OF PAGES:	xvii, 149

Abstract

The growing prevalence of functional and cognitive impairments among older adults presents significant societal challenges, particularly because these conditions often remain undetected until they progress into more serious health concerns. Traditional clinical assessments, which rely primarily on self-reported data, can be hindered by recall bias and subjectivity, limiting their utility for early detection. To address these gaps, this thesis proposes an interpretable, AI-driven framework that integrates ambient sensor data with machine learning (ML) and large language models (LLMs) to support the identification of behavioral changes in smart home environments. Rather than replacing self-report, this approach aims to complement it, with the ultimate goal of enabling timely clinical intervention and promoting aging-in-place with dignity and autonomy.

Despite recent advancements, current approaches to behavior anomaly detection face critical limitations, including underutilization of temporal dependencies, narrow focus on intra-activity anomalies, reliance on labeled data, poor model generalizability, and lack of interpretability. This research addresses these challenges by proposing a novel, multi-component framework that integrates: (1) inverse reinforcement learning (IRL) models for scalable, label-efficient behavior change detection;

(2) Transformer-based architectures with transfer learning to improve generalizability and mitigate cold-start issues; (3) a synthetic data generation model (BehavGAN) to augment training data diversity; and (4) an LLM-based interpretability layer to translate activities of daily living (ADL) logs and anomaly detections into human-readable, clinician-friendly summaries.

Grounded in Fogg's Behavior Model, the proposed system captures both point and collective anomalies by modeling inter-activity and temporal patterns of ADLs. Experiments on public smart home datasets (CASAS-Twor, CASAS-Aruba, and Kastaren) demonstrate high performance across modules: over 90% recall in behavior change detection with an 11% false positive rate, effective cross-user generalization, and successful near real-time monitoring capabilities. LLM integration further bridges the gap between quantitative sensor data and qualitative clinical insight, while human-in-the-loop (HITL) mechanisms and risk mitigation strategies address challenges related to bias, hallucinations, and ethical oversight.

This thesis contributes a scalable, explainable, and ethically aware solution to preventive geriatric care, demonstrating how AI and generative technologies can be responsibly deployed in eldercare to enhance quality of life, reduce healthcare burden, and empower clinicians with actionable, real-time insights.

Dedications

To my beloved parents, whose endless love, sacrifices, and unwavering support have shaped who I am today. You taught me to dream of a better life, to value education, and to keep learning no matter the challenges. Everything I have achieved is because of the foundation you gave me.

To my dearest Sajjad, my best friend and partner in life, who has always believed in me even when I doubted myself. Your encouragement, patience, and faith in me have been my strength, and this journey would not have been possible without you.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Kamran Sartipi and Dr. Yufei Yuan, for their guidance, encouragement, and all that I have learned from them throughout this journey. My thanks also go to my committee member, Dr. Manaf Zargoush and examiner Dr. Alexander Ryu for their constructive feedback, which greatly improved this work.

A special acknowledgment goes to Dr. Norm Archer, from whom I learned immensely during his time as my supervisor. His example as a role model, dedicated to helping and caring for people, will always inspire me.

Table of Contents

List of Figures

List of Tables

Definitions and Abbreviations

Definitions

Activities of Daily Living

The activities of everyday life. "Basic ADLs include eating, dressing, getting into or out of a bed or chair, taking a bath or shower, and using the toilet". "Instrumental activities of daily living (IADL) are activities related to independent living and include preparing meals, managing money, shopping, doing housework, and using the telephone." In the scope of this research, ADL covers both basic ADL and IADL.

Behavior An ordered sequence of activities represented as events.

Event An instance of ADL performed by the resident at a specific time, date, and location for a period of time (An activity is defined with some attributes such as activity type, duration, and period-of-day).

Smart Home The resident's home that is equipped with ambient sensors

Wearable Sensors

Privacy-preserving non-intrusive vital sign sensors that are worn on the body.

Baseline Period

A 2-3 months training data collected from ADL monitoring in a smart home

Behavior Routine

The usual behavior of the residents in the baseline period

Behavior Change

A considerable deviation from the behavior routine such as missing activities, unusual durations, changes in the sequence or timing of events, repetitive actions that are out of the ordinary, unexpected interruptions, and instances of idle time that are unusual.

Abbreviations

AAL Ambient Assisted Living

AI Artificial Intelligence

ADL Activities of Daily Living

BCD Behavior Change Detection

BERT Bidirectional Encoder Representations from Transformers

FBM Fogg's Behavior Model

GAN	Generative Adversarial Network
HITL	Human-in-the-Loop
IRL	Inverse Reinforcement Learning
LLM	Large Language Model
MDP	Markov Decision Process
ML	Machine Learning
MLM	Masked Language Model
RL	Reinforcement Learning

Chapter 1

Introduction

The growing prevalence of functional [Hajek and König, 2016] and cognitive [Lisko et al., 2021] impairments among older adults poses significant societal challenges. As the number of individuals experiencing these declines increases, so too does the urgency for effective early intervention strategies. While traditional clinical assessments rely heavily on self-reported data, susceptible to issues like recall bias and variability in reliability [Piau et al., 2019], monitoring technologies can offer complementary insights. However, these systems also face limitations, such as difficulty detecting certain symptoms, potential unreliability, and risks of misclassification. Thus, a more comprehensive and effective approach lies in integrating both self-reported information and monitoring data to better identify early signs of deterioration. Concurrently, a substantial proportion of older adults live independently, making continuous, objective monitoring both a logistical and ethical necessity.

Recent advancements in sensor technology have enabled the seamless collection of data on ADLs within the home environment. These sensors offer a promising alternative to subjective assessments, providing continuous, objective data that can capture subtle, gradual changes in daily routines—changes that may indicate the onset of serious health issues such as dementia or other chronic conditions [Nathan et al., 2018]. This objective monitoring is critical, as unaddressed deviations in ADL patterns can precipitate a cascade of adverse health events, including malnutrition, falls, and further cognitive decline, thereby increasing the likelihood of institutionalization [Claes et al., 2015].

In parallel, the integration of artificial intelligence (AI) into healthcare, particularly through deep learning and reinforcement learning, has opened new avenues for interpreting complex sensor data. By leveraging these advanced methodologies, this research proposes a novel framework to detect deviations from established behavioral patterns in older adults. Furthermore, the emerging role of large language models (LLMs) in transforming raw sensor data into coherent, human-readable clinical narratives presents an innovative method for bridging the gap between complex data analysis and actionable healthcare insights. However, while LLMs offer the potential to enhance care delivery by highlighting behavioral anomalies and suggesting early interventions, they also introduce risks such as misinterpretation, bias, and ethical concerns regarding patient privacy and clinical decision-making.

The motivation behind this research is twofold. First, it seeks to address the pressing need for reliable, continuous monitoring of older adults' health by employing state-of-the-art machine learning techniques to detect early signs of behavioral change. Second, it aims to explore the dual potential and challenges of applying LLMs to convert sensor-derived ADL data into insightful patient notes, ensuring that these technologies are integrated with adequate risk mitigation and human oversight. This thesis endeavors to develop a robust, risk-aware framework that not only improves early detection of health-related events but also supports timely and effective clinical intervention, ultimately enhancing the quality of care for older adults while allowing them to age in place with dignity and independence.

The remainder of this chapter is organized as follows: first, I present the problem statement and research questions. Then, I discuss the theoretical foundation underlying the proposed approach. Next, I highlight the key research contributions, followed by an overview of the dissertation's organization.

1.1 Problem Statement and Research Questions

Behavioral changes in older adults serve as key indicators for maintaining independent living, yet traditional qualitative assessments of ADLs often lack the precision and objectivity needed for early detection of potential health declines. This research proposes a quantitative approach using advanced ML techniques to monitor subtle shifts in ADL patterns, thereby enabling earlier identification of health events that could lead to more severe conditions if not promptly addressed.

In parallel, the study evaluates the potential of LLMs to transform quantitative ADL data into informative clinical narratives. It examines both the efficacy and risks of employing LLMs to detect and explain behavioral anomalies, while rigorously addressing the associated practical challenges. Central to this investigation is the development of risk-aware frameworks and transparent evaluation methodologies that integrate ML and LLM technologies with human oversight. This balanced approach aims to enhance care delivery in healthcare settings through responsible technological integration. In this research, I seek to answer the following questions:

- (RQ1) Detection Efficacy: how effectively can machine learning algorithms analyze ADL patterns to detect early behavioral changes in older adults?
- (RQ2) Evaluation Framework: how can the effectiveness and efficiency of the proposed approach be systematically evaluated?
- (RQ3) Generalizability of Models: to what extent can models trained on the ADL data of one resident be generalized or adapted to predict the behavior of other residents?
- (RQ4) LLM Inference: how LLMs can be applied to produce logical inferences aligned with observed behavioral changes?

1.2 Theoretical Support

Fogg's Behavior Model (FBM) postulates that human behavior is influenced by three principal factors: ability, motivation, and trigger [Fogg, 2009]. Unlike other behavioral theories that focus solely on ability and motivation, FBM introduces the concept of a trigger as a necessary catalyst for behavior change. The model classifies triggers into three distinct types:

1. **Spark** Designed to counteract low motivation, a spark leverages intrinsic or extrinsic motivational factors (e.g., highlighting potential risks) to prompt the desired behavior.
2. **Facilitator** When the challenge lies in the individual's ability to perform a task, facilitators provide the necessary support to enable the behavior.

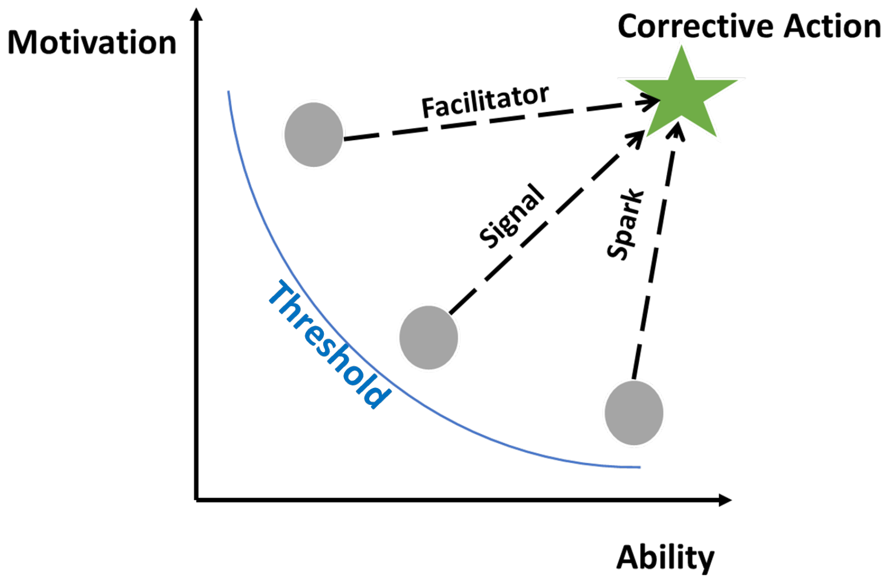


Figure 1.1: A depiction of the trigger's role in Fogg's Behavior Model. The user's motivation and ability should be above a threshold (the blue curve) to make the user respond positively to a trigger

3. **Signal** When both motivation and ability are sufficient, a signal serves as a timely reminder to initiate the behavior.

FBM also emphasizes the importance of the behavior activation threshold; triggers must be administered at the right moment. If a trigger is ill-timed, it may not only fail to produce the intended effect but could also distract or negatively influence the individual's behavior 1.1. Based on FBM, my proposed approach hypothesizes that context-sensitive triggers can effectively support older adults in maintaining their daily routines and accessing needed care. Specifically, when a decline in motivation is detected, an appropriately timed spark may reinvigorate engagement in ADLs. Alternatively, if reduced ability is the underlying issue, interventions in the form of facilitators—such as external support—could be more beneficial. In situations where both motivation and ability are intact, a well-timed signal may suffice to reinforce routine behaviors. Consequently, the proposed system is designed to dynamically evaluate the individual's current state and determine whether to deploy a spark, facilitator, or signal, ensuring that interventions are both appropriate and effective.

1.3 Research Contributions

Inspired by Fogg's behavior model [Fogg, 2009], I design a data-driven solution for the early detection of behavior changes in older adults that are triggered by the onset of health events. I follow a design science research approach in which a new artifact is designed to solve a real-life problem. I use public datasets that are obtained by monitoring the ADLs of older individuals living in smart homes equipped with ambient sensors. The main contributions of this research will be as follows. Each contribution is explicitly linked to the research question it addresses.

- **ML-Based Framework for Behavior Change Detection (addresses Detection Efficacy):** to answer "How effectively can machine learning algorithms analyze ADL patterns to detect early behavioral changes in older adults?", I introduce a novel ML-based framework for behavior change detection in older adults by analyzing sequences of ADLs in order to offer care providers objective indicators about the health status of the older adult living alone at home. This framework employs semi-supervised and supervised learning techniques (e.g., IRL and Transformer-based classifier models) to identify deviations from established routines. By evaluating detection performance (e.g., precision, recall, F1-score) on public smart-home sensor datasets, we demonstrate the framework's capability to flag subtle shifts in behavior that may indicate emerging health issues.

- **Simulation-Based Evaluation with BehavGAN-Augmented Sensor Data (addresses Evaluation Framework):** to answer "How can the effectiveness and efficiency of the proposed approach be systematically evaluated?", I design and conduct simulations on augmented ADL datasets by injecting controlled behavioral deviations, such as increased restroom visits or prolonged inactivity, that correspond to known health events (e.g., urinary tract infections). In addition, I introduce BehavGAN, a novel GAN-based method for synthesizing realistic ADL sequences when real data are scarce. Using these augmented datasets, I quantify detection accuracy (e.g., precision, recall) and false-alarm rates. This contribution defines a reproducible evaluation protocol, covering simulation design and performance metrics for future studies.

- **Transfer Learning for Cold Start Mitigation (addresses Generalizability of Models):** to answer "To what extent can models trained on the ADL data of one resident be generalized or adapted to predict the behavior of other residents?", I apply transfer learning strategies that leverage pre-trained models on existing residents' data to bootstrap detectors for new users. Specifically, I fine-tune BERT Classifier on a small amount of labeled data from a new resident, demonstrating that transferred models achieve superior detection accuracy while requiring fewer individual-specific data. This contribution substantiates the feasibility of cross-resident adaptation in real-world deployments.

- **LLM Interpretation Layer for Clinical Narratives (addresses LLM Inference):** to answer "How LLMs can be applied to produce logical inferences aligned with observed behavioral changes?", I develop a pipeline that converts ML model outputs, such as behavior change scores, and contextual information (e.g., Demographic or medication information along with ADL log) into structured prompts for a large language model (e.g., Gemini, Claude, or GPT-4). The LLM then is tuned to generate evidence-based explanations.

1.4 Thesis Organization

This thesis is organized into six main chapters and supporting appendices. Chapter 1 introduces the research problem, outlines the theoretical foundations, and highlights key contributions. Chapter 2 presents an extensive literature review covering abnormal behavior definitions, ambient assisted living (AAL), behavior anomaly detection, relevant AI and data analytics approaches, and identifies research gaps. Chapter 3 details the proposed methodology, including models for behavior change detection, synthetic data generation using BehavGAN, and an LLM-based clinical note generation framework. Chapter 4 evaluates the proposed system using real-world datasets, reporting experimental results for each system module. Chapter 5 discusses key findings, limitations, and ethical considerations, particularly around trust and privacy in AI-assisted monitoring. Finally, Chapter 6 concludes the thesis by summarizing the main contributions, answering research questions, and outlining future research directions. Appendices provide supplementary background on core AI techniques and technical implementations. This structured approach ensures that theoretical, methodological, and practical dimensions of the research are thoroughly addressed.

Chapter 2

Literature Review

In this section, I provide a review of the related works in behavior change detection. First, I provide a review of "abnormality" definition in the context of activities of daily living. Next, I discuss studies of Ambient Assisted Living (AAL), where sensor networks are utilized to assist residents in their daily living. Then, I provide a summary of studies of behavior anomaly detection in the context of home care. I also review related work that applies different data analytics methods to behavior anomaly detection. Then, I explore the use of LLMs for automating note generation to enhance clinical efficiency and documentation accuracy. Finally, I elaborate on the research gap that is identified through reviewing related work, which is the focus of this research.

2.1 Definition of Abnormal Behaviour

Abnormal behavior in Activities of Daily Living (ADLs) broadly refers to any noticeable deviation from an individual's established patterns of daily routine, across dimensions such as timing, duration, location, order, or statistical likelihood. These deviations can indicate underlying health or cognitive issues, especially in older adults or individuals with chronic conditions. However, since no single, universal definition of abnormality applies across all contexts, the literature presents a variety of interpretations.

The definitions fall under several conceptual categories:

2.1.1 Core Categories of Abnormality

- **Temporal Deviations**: activities performed at atypical times (e.g., preparing dinner at 2 a.m.).
- **Spatial Deviations**: tasks completed in unexpected locations (e.g., brushing teeth in the kitchen).
- **Sequential Deviations**: altered order of events (e.g., taking medication before eating rather than after).
- **Duration Deviations**: time spent on an activity is unusually long or short relative to personal norms.
- **Statistical Deviations**: actions that have low likelihood under learned probabilistic or neural models (e.g., rare event sequences).
- **Contextual Rule Violations**: actions that contradict semantic, behavioral, or sensor-based rules (e.g., motion detected without light activation).
- **Behavioral Omissions or Insertions**: skipping usual steps (e.g., missing meals), or inserting unexpected sub-events (e.g., repeating parts of tasks).

- **Entropy or Disorder Spikes**: increases in entropy over sensor sequences, often indicating environmental disruption (e.g., unplanned visits).
- **Unexpected Inactivity or Idle States**: lingering in a state of inactivity where action is typically expected.

2.1.2 Additional Definitions from the Literature

In addition to the above conceptual types, the following definitions emerged from the broader literature:

- [**Durand and Barlow, 2003**] *Definition:* "actions that are unexpected and often evaluated negatively because they differ from typical or usual behavior."
- [**Mahmoud et al., 2011**] *Definition:* any ADL segment whose feature vector differs significantly (via Euclidean or Mahalanobis distance) from a personalized normal template.
- [**Hoque and Stankovic, 2012**] *Definition:* high-level deviations from expected sequences that violate contextual or causal rules.
- [**Pazhoumand-Dar et al., 2015**] *Definition:* any device usage sequence with low likelihood under an HMM trained on normal usage.
- [**Karakostas et al., 2016**] *Definition:* discrepancies between predicted (expected) and actual behavior based on task models.
- [**Nazerfard, 2018**] *Definition:* violations of learned temporal association rules between activities, including irregular sequencing and unexpected gaps.

- [**Yahaya et al., 2019b**] *Definition:* outliers detected by a one-class SVM trained only on normal ADL data.
- [**Koutli et al., 2019**] *Definition:* sequences with low posterior probability in a hidden semi-Markov model of habit sequences.
- [**Konios et al., 2019**] *Definition:* deviations from a personalized baseline measured by distance functions.
- [**Howedi et al., 2020**] *Definition:* abnormality indicated by spikes in fuzzy entropy over sensor data windows.
- [**Ardebili et al., 2020**] *Definition:* violations of manually defined IF–THEN context-aware sensor rules.
- [**Mustafa et al., 2020**] *Definition:* duration deviations in stages of ADLs beyond group-level statistics.
- [**Yahaya et al., 2019a**] *Definition:* instances where a Normality Score, computed via an ensemble of one-class methods, exceeds a threshold.
- [**Yahaya et al., 2019c**] *Definition:* "any notable departure from a person's typical behavioral pattern," judged by internal and external model consensus.
- [**Arifoglu and Bouchachia, 2019b**] *Definition:* "a departure from expected activity patterns," learned via CNNs trained on both real and synthetic data representing dementia-related behaviors.
- [**Shang et al., 2020**] *Definition:* "a low likelihood of an activity given learned regularity features like time-consistency and frequency."

- **[Fahad and Tahir, 2021]** *Definition:* "missing or extra sub-events in an activity and unusual durations of the activity."
- **[Wang et al., 2023]** *Definition:* deviation from learned patterns in timing, frequency, or order, often related to cognitive impairment.
- **[Tay et al., 2023]** *Definition:*
  - *accidental anomalies:* Detected by abrupt sensor changes (e.g., falls).
  - *non-accidental anomalies:* Violations of sequence/time/location norms (e.g., repetitive or out-of-order actions).

Across a decade of research, the consensus is clear: abnormal behavior in ADLs is any significant deviation, whether in timing, duration, sequence, or context, from a personalized model of normal routine. Techniques vary from probabilistic models to entropy measures and rule-based systems, but all aim to flag deviations indicative of health concerns or safety issues.

Throughout this research, "behavior change" is viewed not just as an "abnormality" but as **any noticeable shift away from a person's normal routine of daily living.** This could involve skipping tasks they usually perform, spending an unusual amount of time on an activity (either much longer or much shorter), changing the order or timing of events, repeating something more than they normally would, experiencing unexpected interruptions, or lingering idle when they typically wouldn't. This thesis adopts a person-centric definition of behavior change. Rather than relying on static rules, we emphasize learning personalized baselines over time and tracking behavioral drift in terms of semantic structure, timing, and frequency. This adaptive approach ensures sensitivity to meaningful anomalies while preserving tolerance for individual variability.

2.2 Ambient Assisted Living

In this subsection, related work on Ambient Assisted Living (AAL) for older adults is reviewed. AAL has proven useful in monitoring Activities of Daily Living because many societies struggle to sustain an aging population due to rising healthcare costs and a shortage of caregivers. Figure 2.1 shows an example home environment that is equipped with different ambient sensors, including motion sensors, door sensors, temperature sensors, and light sensors. Older adults' daily routines and how they conduct ADLs are considered to be reliable signs of healthy aging. Using ambient sensors to monitor older adults' indoor ADLs have several benefits over other monitoring techniques like wearable sensors or video surveillance systems. First, when compared to voice detection or video surveillance systems, ambient sensors present the least privacy risk[Himmel and Ziefle, 2016, Kavitha and Binu, 2019]. Second, since they do not need to be worn constantly and do not interfere with daily activities, ambient sensors offer greater ease of use for older adults[Marschollek et al., 2014, Wilkowska et al., 2022, Kavitha and Binu, 2019]. Third, ambient sensors are suited to be integrated into ADL monitoring systems since they are simple to install and cost-effective for end users [Uddin et al., 2018]. Fourth, compared to video streams, ambient sensor data analysis requires substantially less computational time and power[Kavitha and Binu, 2019]. A variety of aspects of installing high-density environmental sensors in dwellings is investigated by Keohane et al. [Keohane et al., 2018]. Ambient sensors are specifically installed in an environment that supports independent living in order

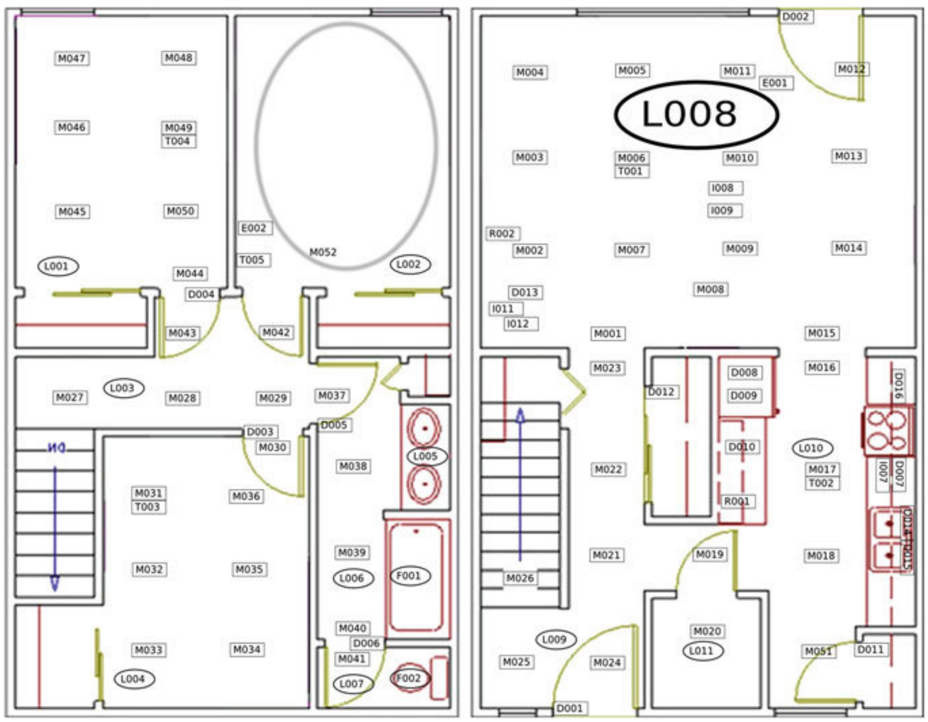


Figure 2.1: Adopted from [Crandall and Cook, 2013]: the floor plan of *Kyoto Testbed* that is equipped with Motion (M), Light (L), Door (D), and Temperature (T) sensors.

to identify ADL behavioral patterns. These patterns are then used to assess the well-being of individuals in their living environments. The study sought to extract from the data the following ADLs: toileting, bathing, sleeping, clothing, mobility (transferring), and cooking (kitchen activity). 48 sensors were distributed throughout an 87-year-old woman's home during a three-month pilot project. The area included a kitchen, bathroom, long hallway, and bedroom. *SmartThings* multi-purpose sensors were attached to items like doors to monitor their opening and closing times as well as their acceleration, direction, and room temperature. *SmartThings* outlets were hooked into wall outlets to control and monitor lights, devices, and small appliances. Also, the *Netatmo Indoor Module* detects pressure, carbon dioxide, sound levels, temperature, humidity, and other variables.

Athanasios et al. [Dasios et al., 2015] detail practical experiences with developing, deploying, and running a wireless sensor network (WSN)-based prototype system for monitoring senior care in residential settings. The monitoring is based on the recording of environmental variables such as temperature, humidity, and light intensity as well as micro-level occurrences that allow inferring daily actions such as moving, sitting, sleeping, and using electrical appliances and plumbing parts. The prototype is composed of low-cost, off-the-shelf components and license-free software, making it an economical solution.

Ambient sensor data are widely used for analyzing behavior patterns. When older adults age in place, there are health events that can be prevented and reduced by analyzing ADLs [Nathan et al., 2018]. Long-term health conditions, such as the risk of falls, have typically been diagnosed through assessments conducted in clinical settings. Smart Homes, however, enable ongoing monitoring of residents' behaviors and vital signs over an extended period. This technology can assist in diagnosing and managing chronic conditions while reducing the burden on healthcare services. Using sensor data, it is possible to develop a behavioral profile for residents and observe changes over time. For example, toileting is essential for survival and a crucial sign of one's health [Cove-Smith and Almond, 2007]. Urinary Tract Infections (UTIs) can be prevented by maintaining a regular washing and toileting schedule, which can also be a sign of urinary incontinence (UI). For those over 70, UTI is a common cause of death that frequently goes undiagnosed [Cove-Smith and Almond, 2007]. Sleep is crucial for cognitive function, particularly for memory consolidation. A deterioration in cognitive function and changes in mood are typically experienced by people who get less than seven hours of sleep each night[Alhola and Polo-Kantola, 2007]. Furthermore, a decrease in mobility is found to be correlated with increased fall risk. Using ambient sensors, it is possible to track sleep duration and patterns, mobility level, patterns of performing ADLs, etc.

Moura et al. [Moura et al., 2022] present an algorithm for learning health changes based on the correlation of context-enriched frequent behavior patterns and cognitive and physical health deterioration. Although the sequence of activities is taken into account in their work, it is only for short-term behavior patterns.

Li et al.[Li et al., 2018] advise examining older adults' movement patterns in their homes using unsupervised learning. They advise utilizing a Bayesian framework with nominal matrix factorization to obtain highly understandable everyday routines.

Saives et al. [Saives et al., 2015], used sequence mining techniques to discover the most frequent patterns in the stream of sensor data that represent the inhabitant's activities. These patterns are subsequently represented using an extended finite automaton, which can be utilized for recognizing activities and generating activity events.

Identifying intra- and inter-activity association patterns from older adults' everyday routines is advised by Lin et al. [Lin et al., 2013]. A data mining technique is used to extract the most frequent sequences of steps within each specific activity (intra-activity pattern) and within a group of everyday activities (inter-activity pattern).

Rashidi et.al [Rashidi and Cook, 2009] introduced CASAS, an adaptable smart-home system that makes use of machine learning to identify trends in occupants' daily routines and create automation rules that mirror those habits. The suggested strategy makes no presumptions regarding the model's underlying properties, such as the activity structure. Nevertheless, finding the patterns of the residents of a smart house is entirely left to the proposed algorithms. To identify recurring patterns, they employ a variation of the Apriori algorithm. They record crucial temporal data, such as event durations and start times, as well as contextual data, such as start-up triggers. In this method, the resident plays a crucial role in determining the automation policy for the environment. The resident can specifically direct the system by offering explicit input, or s/he can let the system learn about and adjust to changes in the pattern of activities on its own.

2.3 Behavior Anomaly Detection in Home Care

Detecting behavioral anomalies is a dynamic area of research that encompasses various elements of Telecare and Telehealth. Recent developments in smart home research for health monitoring demonstrate that these technologies can effectively detect and anticipate health changes in near real-time. In order to show how collaboration between smart homes and care providers may be used to successfully detect and report clinically important health events that can be automatically recognized by smart homes, Fritz et al. [Fritz et al., 2022] present a series of health event examples. In the homes of 25 individuals with various chronic health conditions, ambient sensors were installed. The study examined two cases of congestive heart failure exacerbations, one case of a urinary tract infection, two cases of bowel inflammation flare-ups, and four cases of sleep disturbances among participants. The authors discovered evidence that suggests automated identification of health events may be provided by ongoing sensor-based monitoring of patient behavior in residential settings. Nursing insights into data from smart home sensors are also suggested for use in launching preventive measures and offering prompt care.

Yahaya et al. [Yahaya et al., 2019c], suggested a novel method for assembling a group of novelty detection algorithms. The novelty detection procedure determines whether test data significantly deviates from the data available during training in order to identify new or unknown data. Abnormality in ADLs is defined as any notable departure from a person's typical behavioral pattern. The idea of internal and external consensus serves as the foundation for the suggested Consensus Novelty Detection Ensemble (CNDE) technique. External consensus refers to a voting mechanism among different models in the ensemble. Internal consensus refers to a decision-making process that takes place within each individual model of the ensemble, where multiple components or mechanisms contribute to the final output through an internal voting system. Based on each model's performance and a score termed the "Normality Score," the weight of the model is estimated. The data are classified as abnormal (anomalous) depending on the crossing of a specific threshold, and as normal otherwise, using the computed score.

Dalal et al. outline two key application areas for their rule-based approach to the inference of elders' activity: detecting Independent Activities of Daily Living (IADLs) for the detection of abnormalities in activity data patterns and the passive and covert detection of probable emergency scenarios. Results showed the viability and validity of knowledge-engineered rules, which outperformed mechanically generated rules produced by supervised learning with random forest [Dalal et al., 2005].

Howedi et al. [Howedi et al., 2020] use entropy measurements to do anomaly detection in daily activities. they demonstrated that the suggested method will find anomalies in environments with multiple occupants when there are visitors. The suggested entropy measurements are based on identifying the highest entropy value in routine daily activities, which would be applied as a threshold to identify anomalous ADL behaviors in previously unreported data.

2.4 Data Analytics for Behavior Anomaly Detection

Abnormal behavior can be defined as "actions that are unexpected and often evaluated negatively because they differ from typical or usual behavior" [Durand and Barlow, 2003]. Because the concept of an anomaly is difficult to define precisely and is closely tied to patient behaviors and the types and course of pathologies, artificial intelligence, and more specifically machine learning techniques, have been used to learn to recognize those anomalies. These methods could be categorized into three groups, as follows: [Nie et al., 2015]:

- **Supervised** anomaly detection methods develop classifiers by learning from labeled data that includes both normal and abnormal instances, enabling them to classify new data as either normal or anomalous.
- **Semi-supervised** anomaly detection approaches build a representation of normal behavior based solely on provided examples of typical activity, and then evaluate how likely new data points are to conform to this learned behavior.
- **Unsupervised** anomaly detection methods identify outliers within unlabeled data by assuming that most data points represent normal behavior, allowing deviations to be flagged as anomalies.

From another perspective, anomaly detection methods can be categorized as:

- **Profiling** techniques construct a model that represents normal behavior, enabling the identification of deviations that may indicate anomalies.
- **Discriminating** methods that train a model to discriminate an abnormal activity from a normal activity, assuming that abnormal events have occurred before (only signature anomalies can be detected.) [Sanfeliu and Cortés, 2004].

Scholars have used machine learning methods extensively to analyze ADLs with the goal of providing on-time care and predicting older adults' health conditions. Many studies benefit from the availability of datasets for daily activities, including the use of machine learning methods for predicting/detecting anomalous behavior [Arifoglu and Bouchachia, 2019a, Lotfi et al., 2012a, Riboni et al., 2015a, Suryadevara et al., 2013b].

Fahad et al.[Fahad and Tahir, 2021] propose a method for detecting behavior anomalies by taking into account two types of abnormality: missing or extra sub-events in an activity and unusual durations of the activity. They trained an H2O model to classify events using labeled activities (normal, anomaly). The main problem with such supervised models is that they must be trained using labeled data, which is time-consuming and difficult to generate.

Yahya et al. [Yahaya et al., 2021] suggested an adaptive system pipeline for adjusting to changes in human behavior. The authors propose a forgetting factor feature that enables the model to be adjusted to a person's current habits while forgetting outmoded behavioral patterns. In the forgetting factor, two separate techniques are used to adapt to the dynamic behavior of the person. First, the data dissimilarity technique assesses the similarity of the activity data in order to remove dissimilar data. Second, the data aging technique discards previous behavioral routines depending on the age of the activity data.

Casagrande et al.[Casagrande et al., 2018] have used recurrent neural networks to forecast the future values of the activities for each sensor. When abnormal behavior is anticipated in the near future, the caregiver is informed using the projected values. Investigations into data gathering, classification, and prediction were conducted in actual homes with dementia-affected elderly residents.

In assisted living settings, temporal characteristics of ADLs are taken into consideration to forecast the next activity. Nazerfard [Nazerfard, 2018], presents an association rule mining module that identifies associations among ADLs that are grouped according to the start time and duration of the related ADL. The sequence of the activities is also taken into account.

Karakostas et al. [Karakostas et al., 2016] present an anomaly detection approach in which the predicted user activity is represented by a task model. The predicted and actual behavior are then compared to see if any variance (anomaly) has occurred. The problem with such model-based anomaly detection approaches is that they fail to detect anomalies that have not previously occurred. Ismail et al. [Ismail et al., 2019] propose a context-aware framework for learning and predicting human behavior. Behavior contexts such as weekdays and the time of day are collected from residents' real-life data to improve the accuracy of activity prediction.

Cook et al. [Cook and Schmitter-Edgecombe, 2009] have developed algorithms for automatically learning separate Markov models for each of the five classes of activity (Telephone Use, Hand Washing, Meal Preparation, Eating and Medication Use, and Cleaning). These models are used to both categorize the activities that are carried out in smart homes and to identify errors and inconsistencies in those activities.

Krishna et al. [Krishna et al., 2018], proposed a Long Short-Term Memory(LSTM)-based method for detecting anomalies in daily activity sequences, as well as a comparison of the proposed method with the Hidden Markov Model, which demonstrates comparable results for the LSTM model. Moallem et al. [Moallem et al., 2019] presented an anomaly detection method in smart homes based on deep learning. They used binary sensor data to train a predictor model, which is a recurrent neural network, to predict which sensors will turn on/off and how long the event will last.

Alberdi et al. [Alberdi et al., 2018] considered the viability of detecting the multimodal symptoms that are frequently impaired in Alzheimer's Disease (AD) using covertly obtained activity-aware smart home behavior data. Using longitudinal smart home data collected over an average of more than two years for 29 older adults, the data were automatically assigned to the appropriate activity classes. Time-series statistics with ten behavioral variables were also derived. Every six months, mobility, cognition, and mood were assessed. With the help of these data, regression models were built to forecast test-measured symptoms, and feature selection analysis was carried out. Classification models were created to detect trustworthiness and absolute changes in the scores that indicated symptoms. The findings demonstrate that activity-aware smart home data can be used to forecast symptoms related to mobility, cognition, and sadness.

Arifoglu et al. [Arifoglu and Bouchachia, 2019b] examined the problem of dementia-affected older individuals' activity recognition and inappropriate behavior detection. Given the difficulty in getting real-world data, the research first proposes an approach for creating synthetic data that reflects on some behavioral issues of people with dementia. The second part of the study looked at Convolutional Neural Networks (CNNs), which can be used to predict patterns in activity sequences and identify abnormal behavior associated with dementia. The identification of activities is regarded as a sequence labeling issue, and anomalous behavior is highlighted based on a departure from expected patterns. Additionally, the effectiveness of CNNs is evaluated in comparison to cutting-edge techniques like Conditional Random Fields (CRFs), Hidden Semi-Markov Models, Hidden Markov Models, and Naive Bayes (NB). The outcomes show that CNNs are in a competitive position with the listed state-of-the-art methods.

Shang et al. [Shang et al., 2020] proposed a Feature-based Implicit Irregularity Detection (FIID) method that utilizes unsupervised learning to extract regularity features and estimate the likelihood of implicit irregularities. In their approach, regular daily behaviors are characterized by activities that occur frequently and follow consistent timing patterns. The probability of implicit irregularity in an individual's daily health state is then determined within a multidimensional feature space constructed from these extracted features.

Lago et al. [Lago et al., 2017] introduced contextualized behavior patterns, a long-term behavior model that takes context-related variability into account and then codifies the key ideas relating to activities in Ambient Assisted Living. This study shows that using semantic similarity makes it easier to detect behavioral changes.

2.5 AI for Activities of Daily Living Monitoring

The application of artificial intelligence (AI) in healthcare has garnered significant attention due to its potential to enhance patient monitoring and care delivery. Within this domain, the use of LLMs for processing and interpreting ADLs data represents a novel and rapidly evolving area of research.

Despite advancements in sensor technologies, the interpretation of ADL data remains a challenge due to the complexity and volume of the collected information. Traditional approaches rely on rule-based systems or machine learning models trained on pre-defined features [Timon et al., 2023]. However, these methods often lack the flexibility to generalize across diverse patient populations or adapt to new patterns of behavior. This limitation has motivated the exploration of LLMs as a means to transform raw sensor data into actionable insights.

LLMs have significantly advanced the automation of medical note generation, enhancing clinical efficiency and documentation accuracy. Recent research explores various applications, methodologies, and challenges associated with LLMs in healthcare note generation.

For instance, the HEAL model [Yuan et al., 2024]—a 13B LLaMA2-based LLM—was specifically trained for medical conversations and automated scribing. HEAL outperformed GPT-4 and PMC-LLaMA in PubMedQA with an accuracy of 78.4% and matched GPT-4 in generating high-quality clinical notes, demonstrating the cost-effectiveness of domain-specific LLMs compared to general-purpose models. Similarly, another study [Brake and Schaaf, 2024] compared two approaches for generating clinical notes: one that produced sections independently and another that generated them sequentially. Both methods achieved comparable ROUGE scores and factuality metrics, with LLMs like LLaMA2 exhibiting inter-rater reliability comparable to human annotators for evaluating note consistency.

Beyond accuracy, other research has focused on specific use cases. For example, LLMs have been applied to automate discharge note generation for cardiac patients [Jung et al., 2024], reducing documentation time while maintaining comprehensive and accurate summaries. AscleAI [Han et al., 2024] presented a clinical note management system that automates the creation, organization, and retrieval of notes, streamlining documentation workflows and improving clinician productivity. Additionally, LLMs have been employed to generate synthetic patient-physician dialogues from clinical notes [Das et al., 2024], creating realistic conversational data for training and evaluation purposes.

Further advancements include the use of generative AI to focus on patient-centric note generation [Biswas and Talukdar, 2024], which enhances the quality and relevance of documentation by prioritizing the patient's perspective. Research on generating and de-identifying clinical discharge summaries for the Indian healthcare context [Singh et al., 2024] highlights the potential for LLMs to address privacy-preserving data generation needs. Lastly, frameworks such as MedSyn [Kumichev et al., 2024] leverage LLMs to generate realistic synthetic medical texts, contributing to the development of robust AI systems for healthcare.

Despite these advancements, challenges persist. LLMs can produce hallucinated or incorrect information, potentially compromising patient safety. Data privacy is another critical concern, as LLMs must adhere to strict healthcare data protection regulations. Seamlessly integrating AI-generated notes into clinical workflows also poses challenges, requiring careful consideration to ensure clinician acceptance. Automated evaluation of LLM-generated notes shows promise, with studies suggesting that LLMs can assist in assessing note consistency, though human oversight remains essential [Brake and Schaaf, 2024].

2.6 Research Gap

In this section, I will provide a comparative review of the related works that apply different data-driven methods to detect behavior changes using sensor data. In Table 2.1, important features of these studies are summarized in order to identify the research gap.

The method as well as the **method type** that is used in the work is identified according to Supervised/Unsupervised/Semi-supervised categorization. Related works are also compared in terms of **behavior features** that are considered in formulating behavior. In terms of the analysis **time frame**, some approaches focus on a short-term analysis of behavior while a few works also consider long-term behavior patterns. According to [Lin et al., 2013], the sequence of activities, as well as individual activity, can both exhibit anomalous patterns. Among the studies, some perform abnormality detection within individual activity classes (intra-activity analysis), while some others consider abnormalities that are detected from inter-activity dependencies (**Level of analysis**). Furthermore, I identify the **target subjects** that are studied in each research. Some works focus on patients only, while others consider older adults who live alone. The **type of anomaly** that is considered in the related works is also an important factor when comparing methods. An anomaly can be a point anomaly, a collective anomaly, or a contextual anomaly. As stated by [Erhan et al., 2021], a point anomaly occurs when a single data point deviates from the rest of the dataset. Conversely, a collective anomaly arises when a group of related data points collectively differs from the overall dataset. In this case, while individual points may appear normal, their specific sequence or pattern signifies the anomaly. Finally, a contextual anomaly occurs when a data point is considered anomalous only within a particular context, while in other situations it would be regarded as normal. Additionally, I provide information on the **type of data** utilized in the analyses across various studies.

I hypothesize that analyzing data streams of ADLs from an ambient-assisted environment could help identify the onset of an acute and unusual event in ADLs (e.g., use of the bathroom at night, followed by several hours of immobility) or subtle changes over time (e.g., disorganization in stereotypical habits). By leveraging non-intrusive sensors, my approach can deliver important insights to caregivers, including family members and healthcare providers, enabling effective monitoring of older adults in their own homes and supporting their ability to live independently.

While there are plenty of studies on behavior anomaly detection in older adults, temporal features are not utilized to their full potential. Most of the studies reflect on point anomalies. However, anomalies that can only be identified by considering the sequential features of data are not explored well. Some works are limited to finding abnormalities within activity classes, while there can be abnormalities that can only be detected by a higher-level analysis of activities. Therefore, appropriate behavior granularity needs to be considered. It is also important for the method to present a generalizable solution that can be tuned for different target users in a reasonable time. This feature would allow the method to start learning the behavior patterns from a pre-learned model as opposed to learning from scratch.

To address the above-mentioned issues, I hypothesize that deep learning IRL-based methods (semi-supervised) combined with transformer-based classifier models (supervised) that have been proven effective in analyzing time series data can be effectively applied in analyzing **ADL data streams** for detecting deviations from normal behavior. I propose considering **temporal features** of behavior to detect **collective abnormalities** in **long-term** behavior. This research considers **inter-activity** dependencies in order to understand behavior routines from. I apply state-of-the-art IRL-based methods in order to minimize the need for labeled data. The Transformer-based supervised method will also address the "cold start" issue, in which the algorithm is unable to make any conclusions about residents for whom it has not yet received sufficient training data.

Another significant research gap lies in the translation of ML outputs into actionable insights. Existing studies have predominantly focused on detecting abnormalities using traditional ML methods, without paying sufficient attention to how these outputs are integrated into the decision-making process of care providers—the ultimate users of the system. There is a notable absence of research exploring the presentation of these results in an interpretable and user-friendly format. My research explores the usage of Generative AI capabilities to translate logs and raw ML results into human-understandable notes. This approach aims to support care providers by delivering clear, objective reports that complement self-reported data, rather than replacing it. By automatically generating concise yet comprehensive summaries, the system can help reduce the cognitive burden on care providers, minimizing the need to manually gather and reconcile information from various sources while still acknowledging the value of patient-reported insights.

In summary, this research not only seeks to enhance anomaly detection by incorporating temporal and inter-activity features through advanced ML-based methods, but it also aims to bridge the gap between complex ML outputs and practical clinical decision-making. By integrating Generative AI to produce interpretable summaries, the proposed solution aspires to offer a holistic approach that is both technically robust and user-centric.

Table 2.1: Summary of the related works that apply data-driven methods to detect behavior changes using sensor data

	Research	ML Method	Behavior Features	Time Frame	Level of Analysis	Target Subjects	Anomaly Type	Data Type
Unsupervised	[Banaee et al., 2020]	Temporal Association Rule Mining	Temporal features (Duration, Freq. per day, Freq. per time of day)	Short-term	Inter-activity	Dementia patients	Collective anomaly, Point anomaly	Activity log
	[Suryadevara et al., 2012]	Sequential Pattern Mining	Usage of household appliances and furniture	Short-term	Intra-activity	Older adults	N.A.(Predicts the next sub-sequence of activities)	Activity log
	[Shang et al., 2020]	DBSCAN Clustering	Frequency, Occurrence time	Short-term	Intra-activity.	Older adults	Point anomaly	Activity log
	[Yahaya et al., 2021]	One-class SVM	Start time, Duration, Number of sleep interruptions	Long-term	Intra-activity	Older adults	Contextual anomaly	Sleep log
	[Nazerfard, 2018]	Temporal Association Rule Mining and Clustering	Start time, Duration	Short-term	Inter-activity	Older Adults	Collective anomaly	Activity log
	[Xiao et al., 2024]	Loss-guided Mask Autoencoder	Temporal sensor data and user activity sequences	Short-term	Intra-activity	The public	Point and Contextual anomaly	Multi-sensor time series
Semi-supervised	[Zhu et al., 2015]	DBN (Dynamic Bayesian Network)	Wearable motion sensors (body activity, and hand gesture), location context	Short-term	Inter-activity	Older adults	Point anomaly, collective anomaly, contextual anomaly	Activity log, Location log
	[Virone et al., 2008]	Statistical Predictive Algorithm	Room occupancy period, Activity level	Short-term	Inter-activity	The public	Point anomaly	Movement data
	[Forkan et al., 2015b]	Hidden Markov Model and Fuzzy Rule-based Model	Start time, End time, Type of activity, Location, Vital sign	Short-term	Inter-activity	Older adults	Point anomaly, Collective anomaly	Activity log, Location log, Vital sign data
	[Zekri et al., 2020]	DBSCAN Clustering, Temporal Similarity Score	Start time, Duration, Location	Long-term	Inter-activity	Older adults	Point anomaly	Activity log

	[Cardinaux et al., 2008]	Gaussian Mixture Model	Start time, Duration, Weekend or Weekday, Activity level, Frequency	Short-term	Intra-activity	Older adults	Point anomaly	Motion detection, Space and storage utilization, Appliance use
	Research	ML Method	Behavior Features	Time Frame	Level of Analysis	Target Subjects	Anomaly Type	Data Type
Supervised	[Zhao et al., 2014]	K-Means and Markov Chain Model	Start time and duration of time spent in each room	Short-term	Intra-activity and Inter-activity	Older adults living alone	Point anomaly, Collective anomaly	Spatiotemporal data
	[Arifoglu and Bouchachia, 2019b]	Convolutional Neural Network	The sequence of sensor readings (i.e. sensor firing)	Short-term	Intra-activity	Dementia sufferers	Collective anomaly	Activity log
	[Fahad and Tahir, 2021]	H2O autoencoder	Duration, Number of subevents for each activity class	Short-term	Intra-activity	Older adults	Point anomaly	Activity log
	[Lotfi et al., 2012a]	RNNs	Start time, duration	Short-term	intra-activity	Dementia patients	Point anomaly	Activity log
	[Forkan et al., 2015a]	MapReduce Apriori and Classification	Current and last activity, Room temperature, and Vital signs	Short-term	Intra-activity	Patients	Point anomaly	Activity log, Room temperature log, Vital sign data
	[Alberdi et al., 2018]	Classification	Duration, Time spent per day per activity, Time out of the home, Daily sleep duration and frequency, Total number of activated sensors, Total walking distance per day, Complexity of the daily routine, Number of totals and non-repeated activities per day, Maximum and minimum inactivity times, Similarity with the previous day	Short-term	Inter-activity	AD patients	N.A.(Predicts the multi-modal symptoms of AD	Activity log)

Chapter 3

Proposed Approach

In this chapter, I introduce my proposed approach which integrates three main components, **Smart Monitoring**, **Anomaly Detection**, and **Interpretation** that collaboratively deliver the system's anticipated output. As illustrated in Figure 3.1, the process starts with collecting sensor data from smart homes, which includes extensive logs that are subsequently analyzed to identify ADLs. A variety of ambient sensors installed throughout the home, such as sleep sensors, motion sensors, contact sensors, smart plugs, smart pill boxes, and vital sensors, enable the system to establish a comprehensive and non-intrusive monitoring solution.

Following data collection and ADL detection, the framework employs data augmentation via a synthetic data generation technique [Akbari et al., 2022]. This augmentation not only enlarges the dataset but also diversifies the range of behavioral scenarios, simulating variations that may occur in real-world settings. In parallel, domain-specific anomalies that are identified based on established knowledge of disease symptoms and relevant physiological signals, are deliberately injected into the dataset. These anomalies serve to both challenge the model during training and rigorously assess its detection capabilities during evaluation.

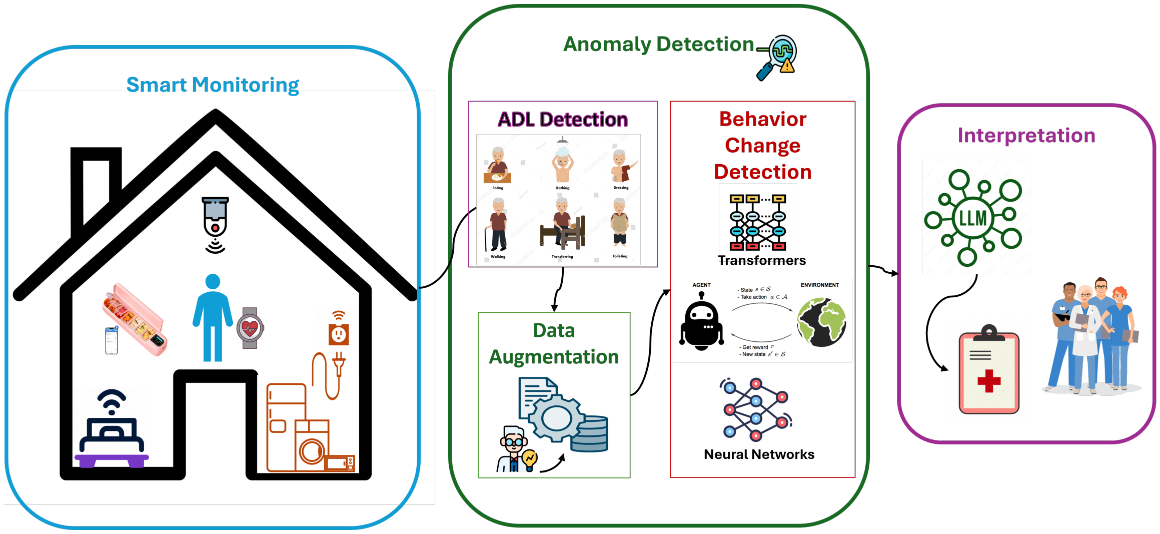


Figure 3.1: The Proposed framework showing the steps in converting raw sensor data into interpretable outputs for the care providers

Subsequent to data preparation, machine learning algorithms are applied to the enriched dataset to identify patterns within the ADL sequences. The primary objective here is to detect deviations from expected behavioral patterns. The system ultimately outputs an indicator that signals whether the observed behavior diverges from normative patterns.

The final stage of the framework is the Interpretation component, which is designed to bridge the gap between the opaque inner workings of the machine learning model and the explicit input-output framework. While the ADL sequences serve as the known inputs and the deviation indicators represent the observable outputs, the transformation process that leads from input to output remains largely hidden. The Interpretation Layer addresses this by extracting and highlighting specific indicators within the behavior sequence that can account for the detected deviation. For instance, an abrupt change in sleep duration or an unexpected shift in meal timings might serve as critical explanatory markers.

To accomplish this, large language models are employed to analyze historical

ADL data alongside the output from the Behavior Change Detection (BCD) module. This analysis culminates in a comprehensive descriptive summary that contextualizes the behavioral anomalies. By elucidating the underlying indicators, the summary provides healthcare professionals and caregivers with clear, evidence-based insights, thereby enhancing their ability to make informed decisions. The workflow of the proposed solution is depicted in Figure 3.2, which will be further discussed in the following sections.

3.1 Modeling and Detecting Changes in ADL

In this section, I introduce two models that I developed to detect behavior changes; an IRL-based Behavior Change Detection(BCD) model and a Transformer-based BCD model. Before introducing the two models, I need to model human indoor behavior for relatively unconstrained environments. This behavior representation will then be utilized in both proposed models as input.

3.1.1 Behavior Representation

Considering behavior as a sequence of discrete tokens (sleeping, eating, watching TV, preparing meals, etc.), two important quantities emerge: i) *content*: activities that constitute a behavior; and ii) *order*: the temporal arrangement of the constituent activities. The idea of tokenizing behavior in this work is similar to the way researchers in Natural Language Processing (NLP) have looked at documents as vectors of their constituent words (see Vector Space Model, VSM [Salton et al., 1975]). Approaches such as VSM capture the content of a sequence in an efficient way. However, they completely ignore its order. Behavior is not fully defined by its activity content alone; rather, by its natural activity orderings. Therefore, a model to capture activity order

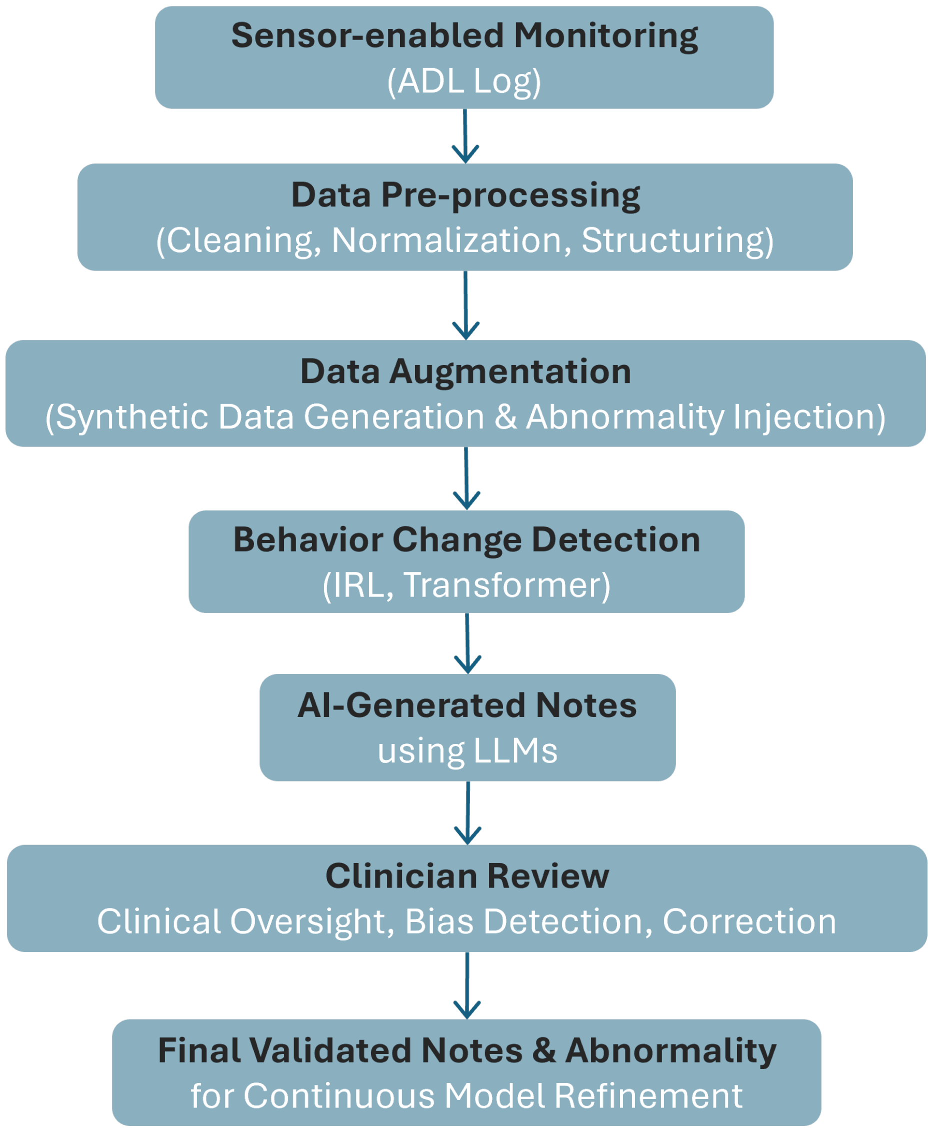


Figure 3.2: The Workflow of the Proposed Solution

in an explicit manner is needed. For this purpose, I consider a sliding window of size W over a behavior sequence to take into account all possible sequences of length T. I consider the start time of ADL as the baseline for the order of tokens in sequences. Therefore, in the case of interleaved ADL, ADL will be put in the sequence according to their start time.

In order to feed the behavior sequence into the Behavior Change Detection (BCD) module, it needs to have a fixed length. However, behavior sequences can be of any length as people perform a different number of ADL each day. To tackle this issue, I define a sliding window (with a shift delta of 1) that allows for sliding over the dynamic-length sequences and capturing ADL dependencies. In this approach, although the length of sequences is fixed to a predefined value (sliding window length), truncating the sequences does not harm the process of capturing ADL dependencies as the dependency between the token at the truncating point and its pre- or post-tokens will be observed in the previous or next sequences, respectively when the window slides over the original sequence. The sliding window size is a parameter of the model that needs to be determined depending on the contextual features of analysis that the generated data will be used for. For example, if data are to be used for learning short patterns, it makes sense to have a small sliding window.

To determine an appropriate value for T, I need to find a small-enough number that, while it limits model complexity, is suitable for covering a representative sequence of the individual's patterns of behavior. In this research, I model human behavior B as an ordered sequence of events:

$$B = e_1, e_2, ..., e_i, ..., e_W \tag{3.1.1}$$

where e_i refers to an event. I define event e_i as a tuple that consists of the activity type a_i, duration d_i, and period-of-day p_i:

$$\begin{aligned} e_i = (a_i, d_i, p_i); where \;\; & a_i \in \{activity \;\; types\} \; and \\ & d_i \in \{activity \;\; duration \;\; range\} \; and \\ & p_i \in \{period\_of\_day \;\; range\} \end{aligned} \tag{3.1.2}$$

An example event is *(Sleeping, Long, Night)*.

Then, I reshape B to a flat tensor B' in order to feed it into the algorithm:

$$\begin{aligned} & B' = y_1, y_2, ..., y_k, ..., y_T; \\ & where \;\; y_k = a_i \; if \;\; k \; \mathbf{rem} \; 3 = 0 \;\; and \\ & \qquad\quad\; y_k = d_i \; if \;\; k \; \mathbf{rem} \; 3 = 1 \;\; and \\ & \qquad\quad\; y_k = p_i \; if \;\; k \; \mathbf{rem} \; 3 = 2 \\ & \qquad\quad\; s.t. \quad i = \lfloor \frac{k+2}{3} \rfloor \end{aligned} \tag{3.1.3}$$

where T is the window size and equals $3 \times W$. It is worth mentioning that activity type and period-of-day are categorical data that need to be encoded in integers so they can be fed into the BCD module. For activity duration, I also discretize the values so the model deals with categorical values. This is helpful to simplify the model by decreasing the state space. As the range of duration in different activity types varies, I first normalize the duration for each activity type, separately. Then, an equal-width discretization method is applied to turn the duration values into categorized values (e.g., extra short, short, medium, long, extra long).

3.1.2 Behavior Change Detection Models

In this section, I discuss the two proposed models for detecting behavior changes. The IRL-based model employs a semi-supervised approach, while the Transformer-based classifier model utilizes a supervised approach.

A: IRL-based BCD Model

In this section, I present my proposed approach for detecting behavior change using Inverse Reinforcement Learning [Ng et al., 2000]. I input recent ADLs into the model to understand the behavior patterns and intentions. I developed a model to detect ADL behavior changes using IRL.

As shown in Figure 3.3, the proposed method works in three layers: input, process, and output. Sensor data logged over 2-3 months (baseline period) are processed in the offline IRL module to learn the weights of the feature vector and the reward function $R(s, a)$. Then, the online IRL module receives the real-time behavior sequence of the resident and calculates its associated reward. Finally, the fusion center compares the calculated reward with a predefined threshold, which represents the average reward for normal sequences, to determine the normality of the real-time behavior.

Problem Formulation I represent the Behavior Change Detection problem as a Markov Decision Process. I define the MDP elements as follows:

- ***State*** $s_t \in STATES$: a sliding window of size W that represents a sequence of the W latest ADL events that the individual has performed at time t: $e_{t-W}, ..., e_t$;
- ***Action*** $a \in Actions$: the next ADL event e_{t+1};
- ***Transition*** $T(s_t, a)$: after taking action a in state s_t, the agent transitions to

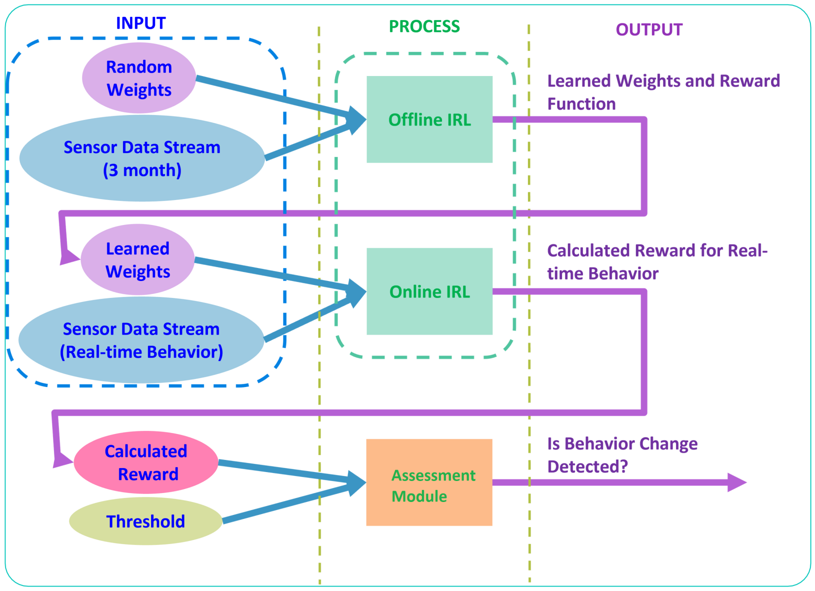


Figure 3.3: The IRL-based Behavior Change Detection Model

state s' that equals $e_{t-W+1}, ..., e_t, e_{t+1}$, which slides the behavior window one token forward.

I propose an IRL algorithm that estimates the reward function $R(s, a)$ from observations. In this model, observations are the ADL trajectories that the individual performs. I use a discount factor to consider the expected future rewards in the long-term reward calculation.

I hypothesize that learning the reward function will enable us to understand unusual ADL sequences. The threshold-based fusion center evaluates the real-time reward and determines the normality of the behavior by comparing the associated reward of the real-time behavior sequence with a predefined threshold R_{th}. In the following blocks, the offline IRL module as well as the online IRL and fusion module are presented.

In algorithm 1, the action space and the observation space are defined based on the number of activity classes (an activity class is a combination of ADL type, timing and duration) and the number of previous activities, respectively. The reward network R

Algorithm 1: Offline IRL

Require: Expert demonstrations $\tau_e = (s_1, a_1, s_2, a_2, ...)$, ADL window size W, Episode length ep_l, Hidden size $hidden_s$, Learning rate lr, Number of epochs num_{epochs}

Ensure: Reward function R

Define the reward network R using a neural network with input size W, hidden size $hidden_s$, and output size equal to the number of activity classes

Define the optimizer (Adam) and the loss function (CrossEntropy) for the reward network

Define a custom Gym environment based on the MDP with parameters (S, A, T, R, γ) , where S is the state space, A is the action space, T is the transition function, and γ is the discount factor.

Train the reward network R using the state-action pairs in τ_e and the optimizer and loss function for a specified number of epochs

return R

is defined using a neural network with input size W, hidden size $hidden_s$, and output size equal to the number of activity classes. The optimizer and the loss function are also defined. The episode length is defined as ep_l. The log data is converted to state-action pairs, and the reward network R is trained using these pairs and the optimizer and loss function for a specified number of training epochs. The trained reward function R is returned as the output of the algorithm.

Algorithm 2: Online IRL and Fusion Module

Data: Real-time ADL Sequence $Input_{seq}$, Reward threshold R_{th}, Reward function R

Result: 0 (No Potential Behavior Change is Detected), 1 (Potential Behavior Change is Detected)

```
Pass the Input_seq to the reward network and get R(Input_seq);
if R[actual action] ≤ R_th then
    return 1;
else
    return 0;
```

Algorithm 5 includes an online IRL module that receives a trained reward function

R, as well as a real-time sequence of ADL and a predefined threshold R_{th} to determine the normality of the behavior sequence. The reward function outputs a reward value for each activity class. In the fusion center, the reward value of the current activity is compared to R_{th} to determine whether the activity conforms to the typical behavior pattern.

B: Transformer-based BCD Model

In this section, I introduce a model that leverages BERT Language models [Devlin et al., 2019] to process ADL sequences, comparable to text corpus.

As illustrated in Figure 3.4, The ADL sequences are fed into the BERT model, which tokenizes them into constituent tokens. After that, tokenized input is embedded. When transforming tokens into tensors of numbers in the embedding layer, the position, context, and token are all taken into account. As a result, a brief nap in the middle of the day is coded differently than a short nap at night, because the position of the token is taken into account when embedding.

The embedded sequences pass through the encoding layer with attention heads to encode the entire sequence in the next layer. I use the BERT-Base-uncased architecture for my implementations, which has 12 attention heads. Finally, a standard classifier, such as a logistic regression classifier, receives the encoded sequence and produces the sequence label, which specifies whether the ADL sequence is a potential anomaly or not.

A major benefit of using BERT models in sequence analysis is their ability to leverage transfer learning. BERT models are initially pre-trained on a large, general-domain text corpus. Afterward, this pre-trained model can be fine-tuned on domain-specific data to perform specialized tasks effectively. Existing research suggests that transfer learning is effective in BERT models. However, a recent study questions the effectiveness of transfer learning when it comes to using pre-trained BERT models for

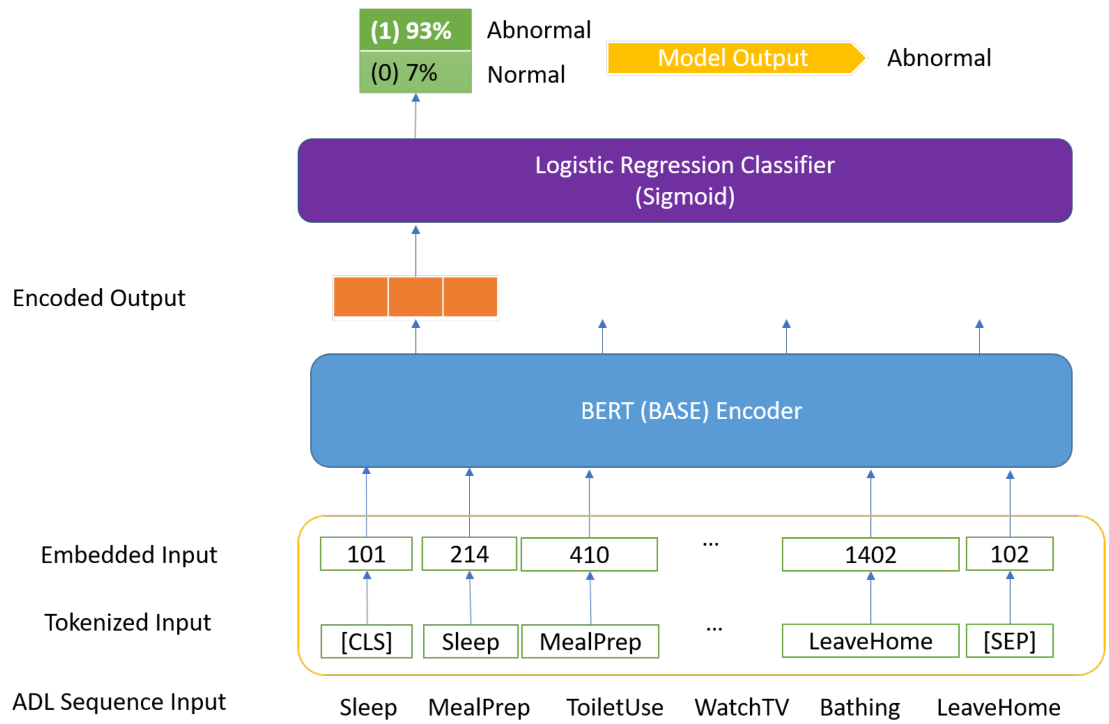


Figure 3.4: The Architecture of the BERT Model for Behavior Change Detection

domains with a high percentage of exclusive vocabulary such as biomedical domain [Gu et al., 2021].

Transfer learning, on the other hand, can be useful for training the BERT model using datasets obtained from multiple residents' ADLs and then fine-tuning the BERT for a specific resident in hand in a shorter amount of time. As a result, for a new resident, the BERT model does not need to be trained from scratch. This feature of the model considerably improves the model's generalizability. I will run the model in two different settings and compare the outcomes, keeping these two characteristics of transfer learning in mind: (1) training the BERT from scratch and fine-tuning on each resident's data. (2) using the pre-trained BERT models and fine-tuning on each resident's ADL sequence data.

3.2 Synthetic Data Generation

Access to datasets detailing individuals' daily behaviors can greatly support a wide range of research, particularly in applying machine learning techniques to predict and identify anomalous behaviors[Li et al., 2020, Suryadevara et al., 2013a, Arifoglu and Bouchachia, 2019c, Kristiansen et al., 2021, Riboni et al., 2015b, Lotfi et al., 2012b, Moutacalli et al., 2015], predicting health conditions [Fritz et al., 2022] or predicting clinical health scores [Cook and Schmitter-Edgecombe, 2021]; and development of reminder and recommender systems in healthcare support and the supervision of long-term behavior [Han et al., 2012, Zhao et al., 2021, Chaminda et al., 2012]. Furthermore, the efficiency and effectiveness of deep learning methods depend on the quality and quantity of training data. Due to the following reasons, existing datasets of real data do not meet the requirements of research in this area: i) *Data scale*: machine learning model training often demands extensive datasets; ii) *Data privay*: health monitoring raises privacy issues for individuals whose activities are tracked [Jourdan et al., 2020]; and iii) *Labeled data*: supervised learning requires labeled datasets, but the process of labeling is labor-intensive and time-consuming.

That being said, synthetic data generation methods have been considered extensively for simulation studies. Generating synthetic datasets is widely used in different domains of study such as computer vision and natural language processing to address the issue of data scarcity. Apart from model-based data generators and simulators [Synnott et al., 2015, Ma and Sartipi, 2015], Generative Adversarial Networks (GANs) have gained significant attention in recent years for their ability to generate realistic images, text, electronic health records (EHRs), and even music using limited real-world data [Radford et al., 2015, Ledig et al., 2017, Yang et al., 2017, Yu et al., 2017, Baowaly et al., 2019]. However, GANs have rarely been used for generating realistic data related to human behavior. Such a dataset could be very beneficial for health monitoring, given the sensitivity of this type of data.

Identical Sample Generation Issue

Generative adversarial networks (GANs) can be difficult to train when it comes to generating sequences consisting of tokens from a limited token space. The issue lies in the fact that the GAN model is trained to generate samples similar to the real data. Thus it is probable that the model will repeat itself and generate records that are identical to the real data. An intuition behind why identical sample generation issue occurs is that the discriminator's output is the only information that is provided to the generator. Therefore, if the discriminator identifies that a generated sample is very similar to the real data, it passes high rewards to the generator and the generator continues to generate from that pattern repeatedly. This issue becomes more severe when it comes to generating data from a limited token space, including behavior sequence generation. In behavior sequences, token space is limited to the activities that an individual can realistically do, which is likely to have less variety than would be seen, for example, in language space or image space.

To address this issue, I introduce a combined reward method that incorporates the BLEU score in the reinforcement mechanism. According to SeqGAN, $R_{D_\phi}^{G_\theta}$ is an action-value function of a sequence, that calculates the expected accumulative reward starting from state s, taking action a, and following policy G_θ. The discriminator's reward is calculated both for complete and partial sequences as follows:

$$R_{D_\phi}^{G_\theta}(a = y_t, s = B'_{1:t-1}) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} D_\phi({B'_{1:t}}^n), {B'_{1:t}}^n \in MC^{G_\beta}(B'_{1:t}; N) & for\, t < T \\ D_\phi(B'_{1:t}) & for\, t = T \end{cases} \tag{3.2.1}$$

where $D_\phi(B'_{1:T})$ is the discriminator's output for a complete sequence (when $t = T$), $B'_{1:T}$, indicates the probability that the sequence is from real sequence data or not. For partial sequences (when $t < T$), N samples of complete sequences (${B'_{1:t}}^n$) that are sequels to the partial sequence will be selected from Monte Carlo tree to be used for estimating the ultimate reward associated with a partial sequence $D_\phi({B'_{1:t}}^n)$. As shown in the above formula, the discriminator reward is calculated after the generation of each token (activity $a = y_t$) with a current state of s.

Since I want to guide the generator in a direction that avoids generating completely identical sequences as the real data and moreover generates a diverse variety of sequences, I need to evaluate it in terms of diversity. The BLEU score gives us a sense of how similar the generated sequence is to the reference set (the real data). Then, I can conclude that samples with a very high BLEU score are likely to trap the model into the issue of identical sample generation. Therefore, I define a new action-value function based on the BLEU score as:

$$R_b^{G_\theta}(a = y_t, s = B'_{1:t-1}) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N} R_b({B'_{1:t}}^n), {B'_{1:t}}^n \in MC^{G_\beta}(B'_{1:t}; N) & for\, t < T \\ R_b(B'_{1:t}) & for\, t = T \end{cases} \tag{3.2.2}$$

where $R_b(B'_{1:T})$ is the BLEU score associated with a complete sequence $B'_{1:T}$ which indicates the similarity of $B'_{1:T}$ to the reference data. Now that the model has a sense of the diversity of the generated sample, I define a combined reward that is a function of both R and R_b:

$$R_{comb} = f(R, R_b) = \begin{cases} max(R) - R & if\ R_b > Threshold \\ R & otherwise \end{cases} \tag{3.2.3}$$

where max(R) equals $max(R^{G_\theta}_{D_\phi}(a = y_t, s = B'_{1:t-1}) : y_t \in \gamma)$ that is the maximum discriminator reward calculated for N generated sequences in every rollout. γ is the vocabulary of candidate tokens. R and R_b also refer to the discriminator's reward and the BLEU reward defined respectively in equations C.0.1 and C.0.2. An overall picture describing the adversarial learning mechanism in my proposed solution is presented in Figure 3.5.

BLEU Score Calculation Computational Cost Issue

As specified by Papineni Papineni et al. [2002], the BLEU score is a modified n-gram measure of the precision of a hypothesis, given a set of references R. 'Modified precision' is the maximum number of occurrences for each n-gram of a hypothesis in the reference set, with an upper bound of the number of occurrences for that n-gram in the hypothesis. The geometric mean is calculated over the precisions for all values of n and multiplied by a brevity penalty, which is 1.0 if the hypothesis sentence is of the same or smaller length than the reference sequence and less than 1.0 otherwise. Thus, a BLEU score of 1.0 means that for all n-grams in the hypothesis, there is at least one sequence in the reference set in which the number of n-gram occurrences is equal to or greater than that of the hypothesis sequence. Its length is also less than or equal to the length of the hypothesis sequence. Assuming that the length

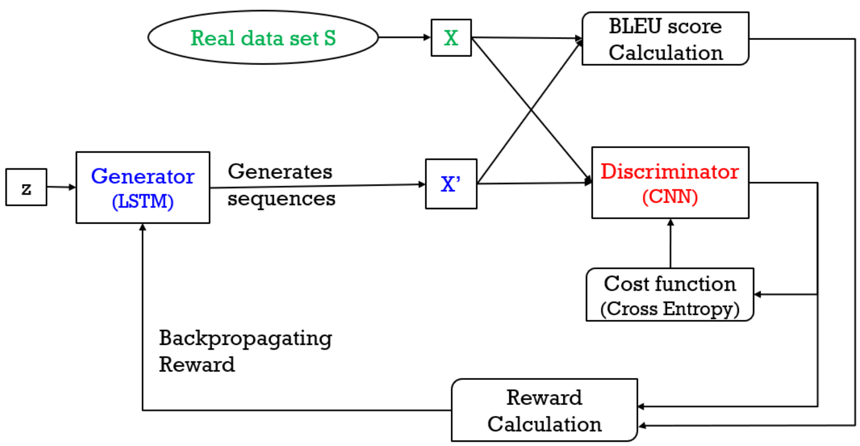


Figure 3.5: Overall architecture of BehavGAN describing the proposed adversarial learning mechanism. The generator network starts generating batches of sequences from a normal distribution. Then, generated sequences with the label ”fake” along with real sequences with the label ”real” are fed into the discriminator network to distinguish real data from fake data. The Discriminator network keeps training based on Cross-Entropy loss. A reward for the generated sequences will be calculated using the BLEU score based on Equation C.0.3. This reward is backpropagated to the generator to guide its learning by reinforcing quality yet diverse sequences.

of generated sequences is fixed to be equal to the length of the reference sequences, brevity penalty would be 1. Usually, n is set to 4 and I denote this metric as BLEU-4 which can measure the similarity between sequences by counting unigrams, bigrams, trigrams and 4-grams. The larger the value of n, the smaller is the BLEU score. In this paper, I apply BLEU-4 as a similarity metric to be calculated in every training loop of the generator.

As demonstrated in Algorithm 3, presented in Appendix C, in every training epoch my algorithm, BehavGAN, calculates the BLEU score for a batch of generated samples. In other words, in each epoch k, the generator generates a sequence which will then be used to calculate the BLEU scores associated with the complete sequence and all the possible partial sequences. The method estimates the BLEU score for partial sequences by applying the rollout mechanism. As explained above, in the rollout mechanism a sample of size N is picked to estimate the BLEU score. Considering the fact that the BLEU score calculation is time-consuming, I need to resolve this issue as this calculation will be performed $N \times T \times K$ times; where N is the rollout number, T is the sequence length and K is the number of epochs.

Therefore, each time I compute the BLEU score for a hypothesis sequence, I compare its n-grams with those of the reference sequences. Since the reference sequences are not changing, I can count the n-grams in the reference sequences only once and utilize that number each time I need to calculate the BLEU score for a new hypothesis, instead of counting the n-grams for every candidate calculation. To optimize the BLEU score calculation I use a hash table data structure. In my implementation, I use the python dictionary data structure where the *key* is the "n-gram" and its associated *value* is the maximum number of occurrences of the corresponding n-gram in a reference sequence. This way, a huge number of calculations are pre-calculated once and the resulting constant values are easily accessible. This implementation makes my algorithm capable of providing feedback based on the BLEU metric in a timely manner.

3.3 LLM-Based Note Generation

The integration of artificial intelligence (AI) into healthcare has catalyzed transformative advancements in patient monitoring and care delivery. Among these, LLMs have emerged as a promising tool for generating human-readable patient notes from sensor-derived data, particularly in the context of ADL monitoring. Smart sensors embedded in the living environment capture continuous streams of data on activities such as eating, sleeping, and mobility patterns, enabling healthcare providers to better understand patients' daily routines and detect potential health issues. Through data analysis, LLMs can convert intricate activity logs into clear, meaningful narratives, identifying anomalies and providing possible explanations to support timely interventions. For instance, an unusual rise in nighttime bathroom visits could indicate a potential urinary tract infection (UTI), enabling early medical evaluation.

However, while LLMs hold significant potential for enhancing care efficiency and accuracy, their deployment in healthcare also introduces critical risks. These models, which rely on probabilistic text generation, can produce misleading or incorrect information, posing challenges to clinical decision-making. Errors such as hallucinations, misinterpretations, or overgeneralizations may lead to misdiagnoses, unnecessary interventions, or delayed responses. Furthermore, ethical concerns including bias in training data, lack of explainability, and potential breaches of privacy, underscore the need for cautious and responsible integration of LLMs into healthcare workflows.

The application of LLMs to ADL monitoring represents a novel approach with dual benefits and challenges. On the one hand, they can bridge the gap between raw sensor data and actionable healthcare insights, offering personalized and proactive care for populations such as older adults or individuals with chronic conditions.

On the other hand, their deployment necessitates robust risk mitigation strategies, including human-in-the-loop oversight to ensure that clinical expertise remains central to decision-making. Over-reliance on AI-generated insights without sufficient human judgment could unintentionally diminish the role of healthcare professionals, exacerbating the risks of automation.

This section explores the efficacy and risks of using LLMs in transforming ADL data into insightful patient notes. Specifically, I investigate the potential of LLMs to enhance care delivery by identifying and explaining behavioral anomalies, while also addressing the ethical, regulatory, and practical challenges of their deployment. My research emphasizes the importance of balancing automation with human judgment through risk-aware frameworks and transparent evaluation methods.

The main goals of the 'Note Generation' component are as follows: 1. **Transforming ADL logs into insightful notes** to generate patient notes from ADL logs collected via smart sensors, providing insights into daily routines and potential health issues. 2. **Investigating risk mitigation strategies**, including human-in-the-loop approaches, to ensure the reliability and accountability of AI-generated notes in clinical decision-making. 3. **Proposing effective methods** for transforming activity log data into prompts for LLMs, leveraging techniques such as Few-shot Prompting and Chain of Thought to improve output quality. 4. **Evaluating the logical inferences of LLMs**, assessing their alignment with actual behavioral anomalies associated with various medical conditions.

In this section, I aim to demonstrate the dual potential and risks of deploying LLMs in healthcare. I advocate for a responsible, risk-aware approach to harnessing the capabilities of generative AI while ensuring accuracy, accountability, and ethical compliance in patient care. Ultimately, my findings contribute to advancing personalized healthcare for vulnerable populations, enabling timely interventions and improved outcomes through a synergistic partnership between AI systems and human expertise.

3.3.1 Background

This subsection provides an overview of fine-tuning approaches for large language models. By fostering a comprehensive comprehension of these concepts, the recognition of the technical and theoretical underpinnings of my proposed approach and its potential applications in the realm of smart home care can be enhanced.

Fine-tuning of LLMs

Language models have made significant advancements in NLP tasks, thanks to the emergence of large-scale language models (LLMs) like GPT-3.5. These models are pretrained on vast amounts of text data and capture extensive linguistic knowledge. However, to adapt LLMs to specific tasks or domains, a process called fine-tuning is necessary.

While pretrained LLMs possess a wealth of general language knowledge, they lack task-specific understanding. Fine-tuning is essential to bridge this gap. By fine-tuning, the pre-trained knowledge can be leveraged and adapted to specific tasks, allowing the model to excel in specialized areas. Fine-tuning is a key component of **transfer learning**. By starting with a pre-trained model and fine-tuning it on a specific task, you can accelerate the training process and leverage the model's general language understanding for the new task. Fine-tuning is beneficial when labeled data for the specific task is limited. By utilizing a pre-trained model's knowledge, fine-tuning can effectively adapt it to the task using a smaller dataset. Fine-tuning on top of a pre-trained model is often more efficient than training from scratch. It saves computational resources and time by skipping the initial training stages and converging faster to a solution.

Fine-tuning is the process of training a pretrained LLM on task-specific or domain-specific data to make it more effective for a particular task. During fine-tuning, the model learns to generalize from the task-specific data while retaining the valuable linguistic knowledge it acquired during pretraining. Fine-tuning helps improve performance, especially when labeled training data for the target task is limited.

There are three primary approaches to fine-tuning LLMs: self-supervised, supervised, and reinforcement learning from human feedback (RLHF) [Ouyang et al., 2022].

- **Self-Supervised Fine-Tuning** [Kalyan et al., 2021] involves training the LLM on a large corpus of unlabeled data for a related auxiliary task. The model learns to predict missing/next words in the input text, effectively creating its own supervision. This self-supervised pretraining provides a strong initialization for fine-tuning, allowing the model to capture general language understanding.

- **Supervised Fine-Tuning** [Devlin et al., 2019] is the most common approach, where the LLM is trained on a labeled dataset specific to the task at hand. The labeled data contains input-output pairs, allowing the model to learn the task-specific mapping. During supervised fine-tuning, the model updates its parameters based on the labeled examples, adapting its knowledge to the specific task requirements.

- **Reinforcement Learning from Human Feedback (RLHF)** [Ouyang et al., 2022] involves training the LLM using reinforcement learning techniques, where human-generated feedback serves as rewards or reinforcement signals. RLHF aligns model outputs with human preferences through feedback mechanisms. This approach is particularly useful in scenarios where obtaining a large labeled dataset is challenging. The model interacts with a human in a dialogue-style setting or receives rewards based on its generated outputs, gradually improving its performance through reinforcement learning.

Each approach has its advantages and considerations. Self-supervised fine-tuning benefits from large amounts of unlabeled data, supervised fine-tuning is effective when labeled data is available, and RLHF is suitable for scenarios where direct supervision is not readily accessible. While fine-tuning offers numerous benefits, it is important to be aware of potential pitfalls. Using a small dataset or excessive epochs can lead to **overfitting**, where the model performs well on the training data but fails to generalize to new data. Insufficient training or an excessively low learning rate can lead to **underfitting**, a situation where the model does not effectively capture the underlying patterns of the task. Fine-tuning poses the risk of **catastrophic forgetting**, where the model loses its broad knowledge acquired during pretraining. This can hinder its performance across various natural language processing tasks.

By understanding these potential pitfalls and employing appropriate strategies, fine-tuning can be effectively utilized to adapt LLMs to specific tasks, domains, and evolving requirements, maximizing their performance and applicability in various NLP scenarios. By fine-tuning, the gap between general language knowledge and task-specific understanding will be closed. The approaches of self-supervised, supervised, and RLHF provide flexibility in adapting LLMs depending on the availability of data and the nature of the task.

3.3.2 Descriptive Note Generation

In the realm of healthcare, understanding the comprehensive well-being of individuals is crucial for providing effective and personalized care. Traditional methods of gathering information, such as in-person assessments or manual data collection, can be time-consuming and resource-intensive. However, recent advancements in NLP and machine learning have introduced a transformative approach: Language Models (LMs) and, more specifically, LLMs. These powerful models have the potential to revolutionize the way healthcare providers obtain insights into the situation of individuals by leveraging data from sensor-monitored activities of daily living.

NLP, the branch of artificial intelligence concerned with the interaction between computers and human language, plays a pivotal role in bridging the gap between technical machine learning models and non- technical healthcare providers. By harnessing the capabilities of NLP, LLMs can effectively communicate complex machine learning concepts and outputs to healthcare professionals in a language they understand. This empowers healthcare providers to harness the potential of LLMs without requiring an in-depth understanding of the underlying technicalities. One promising application of LLMs in healthcare is the generation of comprehensive and human-understandable notes summarizing an individual's behavior patterns based on data collected from sensors monitoring their activities of daily living. ADLs encompass routine activities such as eating, sleeping, mobility, and personal hygiene. By using sensors strategically placed within an individual's living environment, it becomes possible to capture valuable data that can shed light on their overall well-being and daily patterns. However, deriving actionable insights from this wealth of sensor data can be challenging without the aid of advanced analytical tools.

This is where LLMs come into play. By training these models on vast amounts of data, they acquire the ability to understand and analyze patterns within the sensor-monitored ADL data. Furthermore, prompt engineering techniques, such as few-shot prompt engineering, provide an avenue for fine-tuning LLMs and their foundation models using domain-specific data. This process involves providing targeted instructions or examples to the model during training, enabling it to generate comprehensive notes that precisely capture the nuances of an individual's behavior patterns based on the sensor data. These notes can highlight potential abnormalities, symptoms, and underlying reasons behind the observed behavior patterns.

The integration of LLMs and ML-based anomaly detection models can be particularly powerful. While the core ML model detects potential behavior changes through a sequence of ADLs, the LLM model, which is already trained and fine-tuned on domain- and patient-specific data, can generate a comprehensive note on the potential abnormality, pointing to possible patterns, symptoms, and reasons behind the abnormality. This complementary approach provides a holistic understanding of an individual's situation, aiding healthcare providers in monitoring and addressing their well-being effectively. LLMs can be tailored for generating comprehensive notes about the individual's behavior from sequences of ADLs by leveraging prompt engineering techniques and domain-specific training data. Here is an explanation of the process:

- Data Collection: first, data are collected from sensors that monitor the individual's ADLs. These sensors can capture information on various activities, such as eating, sleeping, mobility, personal hygiene, etc. Sensor data is first processed for ADL recognition, producing a sequence of activities that is then fed into the LLM.

- Training the LLM: the pre-trained LLM is further trained using a large corpus of text, which includes domain-specific data related to the individual's behavior patterns and ADLs. This training enables the LLM to learn the statistical patterns, language structures, and domain-specific knowledge necessary to generate meaningful and relevant notes.

- Fine-tuning with Prompt Engineering: to tailor the LLM for generating notes about the individual's behavior, prompt engineering techniques are employed. Prompt engineering involves providing targeted instructions or examples, known as prompts, to the LLM during the fine-tuning process. These prompts guide the model to generate notes specifically focused on summarizing the individual's behavior patterns based on the ADL sequences. Prompt engineering is a low-cost yet effective approach because it lets you "steer" a pre-trained model toward a new task or domain without the need for retraining or modifying the model's internal weights. By simply crafting or refining the input prompt, you can guide the model's behavior without updating any of its millions (or billions) of parameters. This means you can deploy your solution immediately without the heavy computational costs associated with full model fine-tuning. Pretrained LLMs already contain a vast amount of general language and domain-agnostic knowledge. Prompt engineering taps into that existing reservoir, coaxing the model to apply what it knows to a specific problem by framing the input correctly. Since you're not performing any additional training, you don't need large labeled datasets or specialized hardware (like GPUs) to modify the model. This makes prompt engineering especially attractive for rapid prototyping or use in resource-constrained settings.

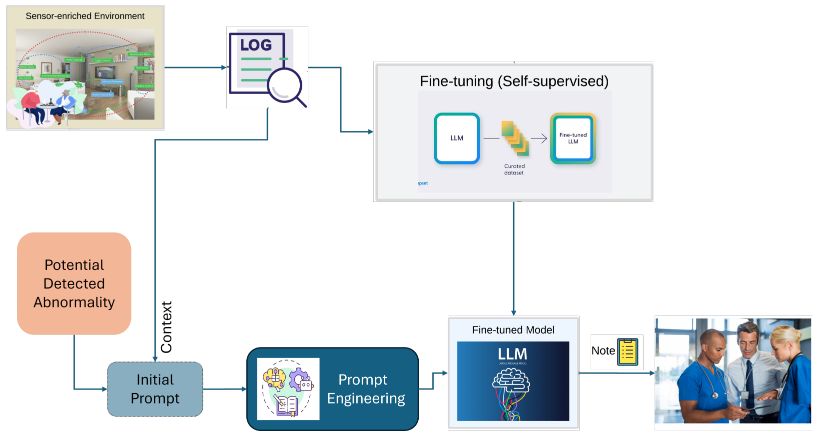
- Domain-Specific Training Data: to enhance the LLM's understanding of the individual's behavior, domain-specific training data is incorporated. These data can include human-generated notes or summaries of disease symptoms. The LLM is fine-tuned using these labeled data to capture the specific reasons behind the changes in the behavior pattern.

- Generating Notes: once the LLM is trained and fine-tuned, it can generate human-understandable notes based on the input sequences of ADLs. The LLM analyzes the patterns in the ADL data and generates coherent and comprehensive notes that highlight potential abnormalities, symptoms, and the underlying reasons behind observed behavior patterns. These notes can aid healthcare providers in monitoring the individual's well-being, identifying issues, and providing personalized care recommendations.

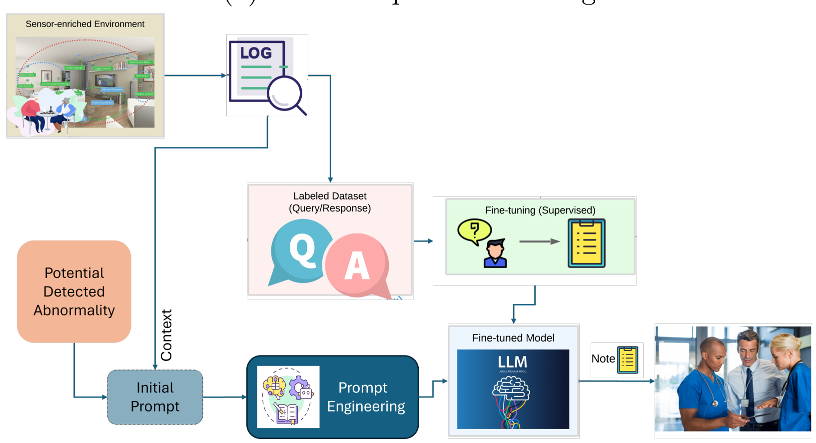
The potential benefits of leveraging LLMs for generating human-understandable notes from ADL sensor data are numerous. Healthcare providers can gain valuable insights into the individual's sleep patterns, mobility, diet, and overall daily routine, enabling them to identify potential issues, monitor changes in health status, and provide personalized care recommendations. Furthermore, the automated generation of comprehensive notes alleviates the burden on healthcare professionals, freeing up their time to focus on critical decision-making and direct patient interaction. In conclusion, the integration of LLMs, NLP, and sensor-monitored ADL data presents an exciting opportunity to enhance the understanding of an individual's behavior patterns for healthcare providers. By combining the power of NLP to communicate machine learning results and the capabilities of LLMs to analyze ADL data, these models can generate detailed, human-understandable notes, enabling healthcare providers to deliver more personalized and effective care. With the aid of prompt engineering techniques, LLMs can be fine-tuned to exhibit domain-specific expertise, further amplifying their value in the healthcare domain.

My proposed approach combines a behavior change detection module with fine-tuned LLMs to generate descriptive notes highlighting abnormal occurrences within the activity log data. To provide meaningful insights and contextual information about the detected behavior changes, I employ LLMs for generating descriptive notes. Figure 3.6 illustrates my proposed architecture for training an LLM for note generation.

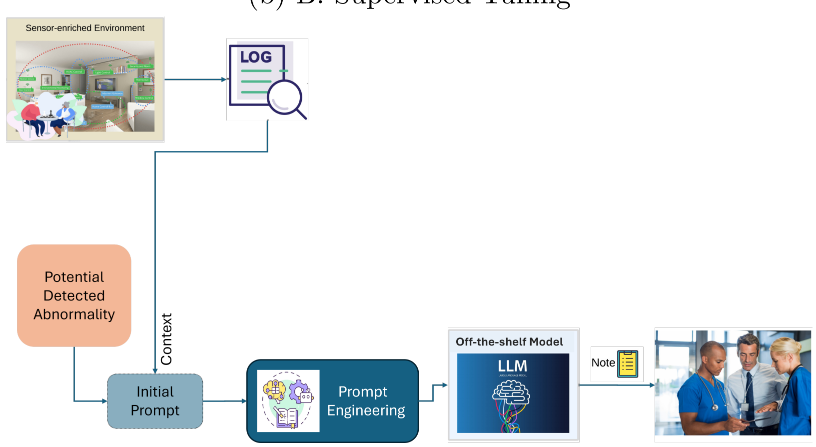
The LLMs can be fine-tuned on the activity log data, which enables them to learn domain-specific language patterns and characteristics. Fine-tuning can involve **self-supervised** training of LLMs on a large corpus of activity log data (Figure 3.6a), **supervised** tuning by further training of LLMs on a labeled Question-Answer (e.g., Prompt-Note) dataset (Figure 3.6b), or utilizing techniques such as **prompt engineering** and few-shot learning to enhance their performance in our specific scenario



(a) A: Self-supervised Tuning



(b) B: Supervised Tuning



(c) C: Prompt Engineering

Figure 3.6: The Proposed Architecture for Note Generation

(Figure 3.6c).

I incorporate prompt engineering techniques, carefully designing informative and context-rich prompts to guide the LLMs in generating notes specifically tailored to behavior change occurrences in the activity log data. Given the behavior change detection module's output, I pass the abnormal days' activity logs to the fine-tuned LLMs. The LLMs utilize the learned knowledge from the fine-tuning process and the context provided by the prompts to generate descriptive notes. These notes highlight the specific locations or instances within the activity logs where behavior change is detected, aiding in the interpretation and understanding of the abnormality patterns.

My proposed approach combines the strengths of behavior change detection modules with fine-tuned LLMs to provide descriptive notes that pinpoint abnormal occurrences in activity log data. By leveraging prompt engineering, the LLMs' ability to generate contextually relevant and informative notes is enhanced. This approach contributes to a comprehensive understanding of abnormal behavior, enabling domain experts to take appropriate actions based on the generated notes and improve anomaly detection and monitoring systems.

3.3.3 Human-in-the-Loop AI Systems

The integration of LLMs into patient monitoring and note generation presents significant opportunities but also introduces risks related to reliability, bias, and ethical concerns. A human-in-the-loop approach, which incorporates human oversight into AI-driven processes, is essential for ensuring the safe and effective deployment of generative AI in healthcare. By leveraging Human-AI collaboration and interaction dynamics, as shown in Figure 3.7 , HITL frameworks can enhance model accountability, improve accuracy, and mitigate potential harms.

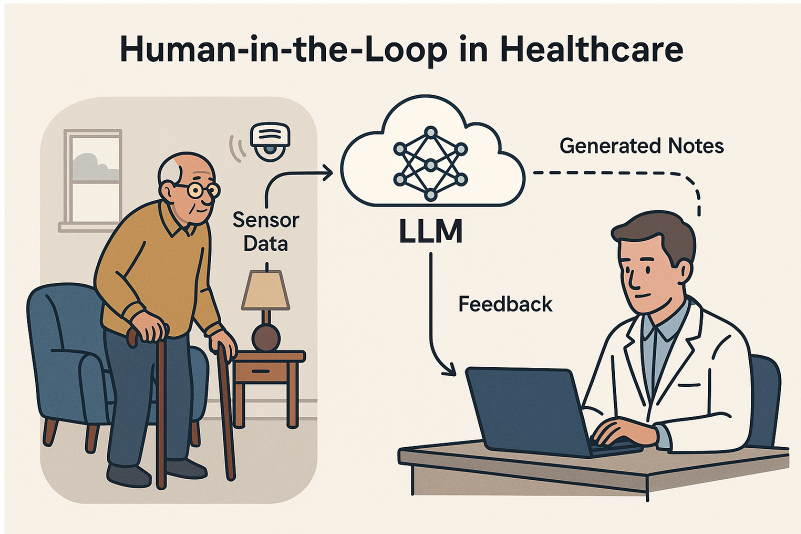


Figure 3.7: The Interaction of AI and Human to Mitigate Risks of Leveraging AI in Healthcare

Ensuring Clinical Oversight and Decision Support

Generative AI models are highly capable of synthesizing large volumes of patient data, but they lack the contextual awareness and clinical reasoning skills that human healthcare professionals possess. A HITL approach ensures that AI-generated patient notes and anomaly detections are reviewed, validated, and corrected by clinicians before they influence medical decisions. This prevents over-reliance on AI outputs, reducing the risk of misdiagnoses, hallucinations, and misleading insights that could compromise patient care.

Bias Detection and Continuous Model Refinement

One of the primary risks of LLMs in healthcare is the potential for biased outputs, particularly if the training data does not adequately represent diverse patient populations. HITL systems enable continuous feedback loops where healthcare providers can flag and correct biased or inaccurate AI-generated content. This feedback can be incorporated into reinforcement learning from human feedback (RLHF) models, improving future iterations of the system and making them more reliable across different demographic and medical contexts.(refer to Deepseek RL approach)

Context-Aware Interpretation of Patient Data

AI models may misinterpret activity log data, generating insights that lack clinical relevance or fail to account for patient-specific factors such as comorbidities, environmental influences, or behavioral patterns. Human experts play a crucial role in contextualizing AI-generated notes by integrating external knowledge and patient history, ensuring that outputs align with real-world medical conditions. By maintaining human interpretability, AI-generated insights can be adjusted to better support individualized patient care.

Reducing Automation Bias and Promoting Trust

Automation bias, the tendency of users to over-trust AI-generated outputs, can lead to blind acceptance of incorrect or incomplete information. HITL systems mitigate this risk by enforcing a verification protocol where human clinicians must approve or modify AI-generated patient notes before they are used in clinical settings. This structured oversight not only improves patient safety but also enhances trust in AI-assisted healthcare, ensuring that the technology is viewed as a decision-support tool rather than a replacement for human expertise.

Ethical and Legal Accountability

AI-generated patient documentation raises concerns about liability, ethical responsibility, and transparency in medical decision-making. Integrating a HITL mechanism ensures that critical clinical decisions are ultimately governed by qualified healthcare professionals, preserving human oversight and accountability in AI-supported healthcare systems [Shahriari and Shahriari, 2017, Guidance, 2021]. This safeguards against legal ambiguities and ensures compliance with regulatory frameworks, such as HIPAA and GDPR, which govern patient data privacy and medical decision-making.

The success of generative AI in patient monitoring and note generation depends on effective Human-AI collaboration. By maintaining a structured human-in-the-loop framework, healthcare professionals can maximize the benefits of AI while minimizing risks related to reliability, bias, and ethical concerns. Future research should focus on optimizing these collaborative dynamics, ensuring that AI systems are designed to complement, rather than replace, human expertise in clinical decision-making.

Chapter 4

Evaluation and Results

In this section, I discuss my evaluation method to show the efficiency of my proposed framework in detecting and reporting behavior changes. I, first, introduce the datasets I used. Then, I discuss the evaluation metrics and evaluation process. Finally, I present the results of my work for each component of the framework separately.

4.1 Datasets

In this section, I introduce three public datasets that I used throughout my experiments.

4.1.1 CASAS-Twor Dataset

To evaluate the proposed Transformer-based model for behavior change detection, I chose the CASAS-Twor2010 dataset [Cook, 2010] which consists of normal daily activities that two residents,R1 and R2, performed in the WSU smart apartment testbed during the 2009-2010 academic year. Some examples from this data set are shown in Table 4.2. In this dataset, thirteen types of indoor activities were included. Bathing, Bed-Toilet-Transition, Eating, Enter-Home, Housekeeping, Leave-Home, Meal-Preparation, Personal-Hygiene, Sleep, Sleeping-Not-in-Bed, Wandering-in-Room, Watch-TV, and Work were recorded using motion sensors, door sensors and temperature sensors. As shown in Table 4.2, start and end times for each activity were recorded, making it possible to calculate the duration of the activity. Also, the time ordering of activities was captured. As there is no overlap in the times of activities performed, I can conclude that concurrent activities were not considered. The CASAS-Twor2010 dataset has 2,804,813 records which have recorded a total 3744 number of activities comprising 1903 activities for resident A and 1841 activities for resident.

As the original data is not labeled, in order to use it for training the supervised change detection model, I inject samples of behavior abnormalities by rearranging ADLs and manipulating activity duration. For example, while in the original ADL sequences eating occurs after meal preparation, I reversed the ADLs' order to inject partially misordered sequences. I also created some abnormalities by randomly shuffling the ADLs. I intentionally make abnormal records frequent (oversampling) in order to avoid imbalanced data problems.

4.1.2 CASAS-Aruba

The CASAS-Aruba dataset [Cook, 2010] consists of activities that a woman performed at home during a period of seven months. A few examples from this dataset are shown in Table 4.2. In this dataset, eleven types of indoor activities were included. Meal preparation, Relaxing, Eating, Working, Sleeping, Washing Dishes, Bed to Toilet, Entering Home, Leaving Home, Housekeeping, and Respiration were recorded using motion sensors, door sensors, and temperature sensors. As shown in Table 4.2, start and end times for each activity were recorded, making it possible to calculate the

Table 4.1: Example data from CASAS-Twor2010 dataset.

Date	**Time**	**SensorID**	**SensorState**	**Activity**
24-8-09	00:15:25	M034	ON	R2_Sleep begin
24-8-09	00:15:27	M047	OFF	
.	.	.	.	
24-8-09	00:16:27	M047	ON	R1_Sleep end
24-8-09	00:16:29	M048	ON	R1_Wandering_in_room begin
.	.	.	.	
24-8-09	00:23:44	M048	OFF	
24-8-09	00:23:52	M048	ON	R1_Wandering_in_room end
24-8-09	00:23:53	M047	ON	R1_Sleep begin
24-8-09	00:23:53	M046	ON	
.	.	.	.	
24-8-09	06:32:46	P001	507	R1_Sleep end
24-8-09	06:32:46	D005	CLOSE	R1_Personal_Hygiene begin
24-8-09	06:32:47	M038	OFF	
.	.	.	.	
24-8-09	06:37:48	M040	OFF	R1_Personal_Hygiene end
24-8-09	06:38:22	P001	579	R1_Bathing begin
24-8-09	06:39:08	T004	20.5	
.	.	.	.	
24-8-09	06:51:00	M040	OFF	
24-8-09	06:51:02	P001	5053	R1_Bathing end
24-8-09	06:51:04	M038	OFF	R1_Personal_Hygiene begin
.	.	.	.	
24-8-09	06:54:37	D005	OPEN	R1_Personal_Hygiene end
24-8-09	07:07:50	M034	OFF	R2_Sleep end
.	.	.	.	
24-8-09	07:07:57	M038	ON	R2_Personal_Hygiene begin
.	.	.	.	
24-8-09	07:08:45	M019	ON	R1_Meal_Preparation end
.	.	.	.	
24-8-09	07:08:48	M024	ON	R1_Leave_Home begin
.	.	.	.	
24-8-09	07:08:58	M024	OFF	R1_Leave_Home end
.	.	.	.	
24-8-09	07:10:43	M037	ON	R2_Personal_Hygiene end

Table 4.2: Example Data from CASAS-Aruba Dataset.

Date	**Time**	**SensorID**	**SensorState**	**Activity**
2010-11-04	00:03:50	M003	ON	Sleeping begin
2010-11-04	00:03:57	M003	OFF	
2010-11-04	00:15:08	T002	21.5	
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2010-11-04	05:40:43	M003	OFF	Sleeping end
2010-11-04	05:40:51	M004	ON	
2010-11-04	05:40:52	M005	OFF	BedToToilet begin
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2010-11-04	05:43:30	M004	OFF	BedToToilet end

Table 4.3: CASAS-Aruba Dataset Statistics

Number of Records	**Number of ADL**
1,719,558	6,477

duration of the activity. Also, the time ordering of activities was captured. Table 4.5 and Table 4.4 present some overall statistics on the Aruba dataset.

Figure 4.1 illustrates the behavior trend of the Aruba resident for two weeks. Looking at the plot, it can be seen that there is a clear pattern in the activities over the course of the month. For example, there are periods where the woman is predominantly sleeping or eating, followed by periods where she is predominantly working or engaging in other activities. Additionally, we can see that there is some variation in the activities from day to day, with some days showing more variety in activities than others. Figure 4.2, It also demonstrates the behavior trend over two days, highlighting the similarity in behavior patterns between the two days.

Table 4.4: ADL Types in CASAS-Aruba Dataset

ADL Type	**Number of Records**
eating	257
enter_home	431
housekeeping	33
leave_home	431
meal_preparation	1606
Bed_to_Toilet	157
Wash_Dishes	65
sleeping	401
work	171
Relax	2910
Resperate	6

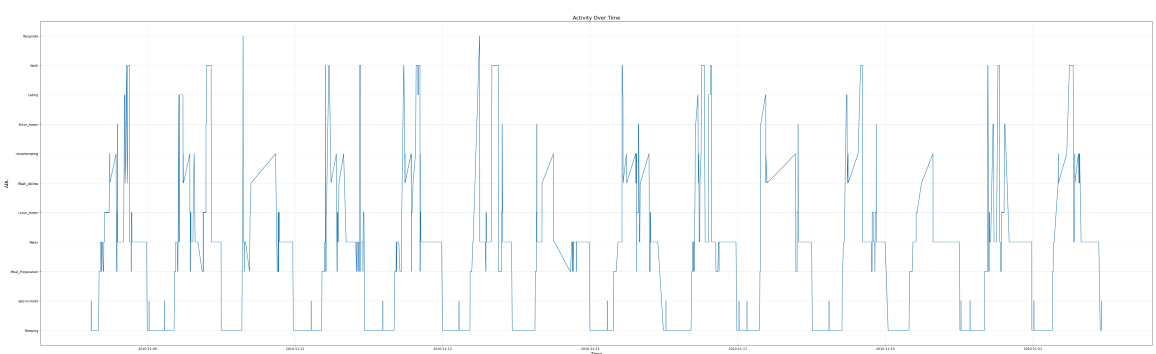


Figure 4.1: The Behavior Trend of the Aruba Resident over Two Weeks

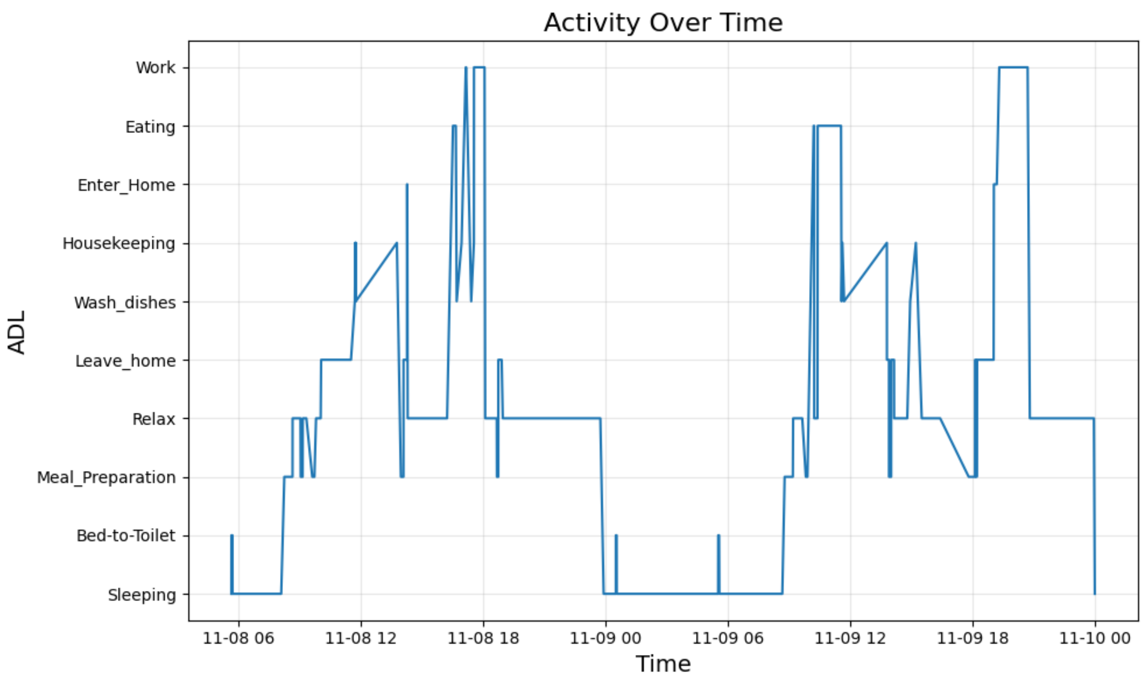


Figure 4.2: The Behavior Trend of the Aruba Resident over Two Days

Table 4.5: Kastaren Dataset Statistics.

Dataset	**Number of Records**	**Number of Activities**
Kastaren	2,120	245

4.1.3 Kastaren

In Kastaren dataset seven different activities are annotated, namely: leave house, toileting, showering, sleeping, preparing breakfast, preparing dinner and preparing a beverage. Table 4.5 presents some overall statistics about this dataset.

4.2 Results

In this section, I introduce the evaluation process and present the results of four main components of my proposed framework as follows:

- BehavGAN model, which is introduced to create synthetic data.
- BERT-based BCD module
- IRL-based BCD module
- Note generation module

4.2.1 BehavGAN Model

In this section, I discuss my experimentation to generate synthetic dataset based on a real dataset. I introduce the proposed combined reward evaluation, and discuss the evaluation process which illustrates the improvement to the quality of generated data compared to those in baseline methods such as MLE, LeakGAN and SeqGAN.

Table 4.6 presents the specification of the high performance computing platform, i.e., the Nvidia's DGX-1 HPC server, that I used for computations in this research project. To develop a GAN for generating synthetic yet realistic dataset, I chose two real daily activity datasets as ground-truth data to test the effectiveness of Behav-GAN. These were: (i) the CASAS-Aruba dataset (4.1.2) [Cook, 2010], which consists of activities that a woman performed in a home during a period of seven months; and (ii) Kastaren dataset (4.1.3) [Van Kasteren et al., 2008] consisting of 28 days of sensor data with annotation of activities.

As discussed in Section 3.2, I first encoded the dataset records and defined a sliding window of size 10 (BehavGAN), or padded sequences until length 20 is reached (BehavGAN_padded) from which behavior tensors were calculated for the model. I

Table 4.6: NVIDIA high performance computing platform used in this research.

HPC Server	
GPU Architecture	NVIDIA Volta
GPU Product	NVIDIA Tesla V100
Driver Version	418.126.02
CUDA Version	10.1
GPU Memory	16 GB HBM2
Memory Bandwidth	900 GB/sec
System Memory	251 GiB
Operating System	
OS Version	Ubuntu 18.04.4
Software	
Programming Language	Python 3.6.9
Libraries	NVIDIA Release 20.01-tf2

chose 10 and 20 for the sequence lengths by analyzing the real data sequences. It turned out that most days have less than 10 ADLs, which suggests defining the sliding window size of 10. Also, the maximum length of daily sequences in the real data is 16, which is why I set the padded sequence length to 20 to allow for generating marginally longer sequences. These tensors are represented in Algorithm 3 as members of the S set. The Algorithm then generates negative samples via the generator network and eventually outputs a final generated dataset. Table 4.7 illustrates the parameters I set for running the BehavGAN algorithm on CASAS-Aruba and Kastaren datasets. I ran the model for the two datasets, separately.

I employed the same architecture for both the generator and the discriminator networks as in the original SeqGAN study. The Tanh activation function is used in the generator's LSTM network. The hidden states are then mapped into the output token distribution via a Softmax output layer. For pre-training, the generator implements Negative Log-Likelihood Loss (MLE pre-training steps). The generator seeks to maximize the reward as well as the discriminator's loss. The discriminator network outputs the likelihood that a given sequence is real using a fully connected Sigmoid layer. Before the final fully connected layer, it adds a highway layer and a dropout layer (0.75). The discriminator uses Cross Entropy loss. All parameters are randomly initialized. Both networks use Adam optimizer. For calculating the BLEU reward ($R_b^{G_\theta}$) I used BLEU-4 as it is usually used for evaluating the similarity of a hypothesis sequence to a reference set.

I also ran SeqGAN and LeakGAN algorithms [Yu et al., 2017, Guo et al., 2018] with real data as input. For the sake of comparison I implemented a MLE model to generate synthetic data, which aims to maximize the log-likelihood of ground-truth sequences. Simply put, it is trained to predict the next token based on the ground-truth tokens that have come before it. This method was also used in the pre-training of SeqGAN and my proposed algorithm, but in this case, I did not use it for pre-training but for the training process. Figure 4.3 shows the distribution of BLEU-4 scores for CASAS-Aruba data as well as generated data using MLE, SeqGAN, LeakGAN, BehavGAN, and BehavGAN_padded (BehavGAN with padding). The purpose of this comparison is to examine if the similarity of the generated data to real data is comparable to what happens in real data. I want it to resemble what happens with real-world data. To compare the distribution of synthetic data with that of real data I calculate BLEU-4 score for each dataset. To calculate BLEU scores for synthetic datasets, I consider the real data as the reference set while the generated data with each model is considered as the candidate set. To calculate BLEU scores for the real data (CASAS), I partitioned the real data into two separate ordered subsets. The first half goes to the candidate set and the second half goes to the reference set. By comparing all candidate sequences with the reference set I calculate the similarity of the first half to the second half.

As shown in Figure 4.3, the generated sequences using SeqGAN, LeakGAN and

Table 4.7: Run parameters for BehavGAN model

Parameter	**CASAS-Aruba**	**Kastaren**
No of Generated Records	10,000	1,000
BL_Threshold	0.85	0.8
Sequence Length	10, 20	10, 20
Pre-training Epochs	50	250
Training Epochs	150	200
Generator's Learning Rate	0.08	0.03

MLE are very similar to the reference set (CASAS data). The issue with data distribution of three baseline methods is that Using baseline models, a major amount of the generated data is too similar to the original data (i.e., Q1, Q2(Median), Q3, and Maximum are all too similar), reducing the diversity of the generated data. My method, as compared to MLE, LeakGAN, and SeqGAN, decreases the proportion of generated sequences that are overly close to the real data (BLEU-4 score near 1). It is important to note that my proposed algorithm can produce sequences that are comparable to the real dataset (median remains around 0.85) while avoiding the generation of a large number of identical records (Q3 is in the neighborhood of 0.9). This feature of BehavGAN makes it a better solution for generating synthetic data. Furthermore, the results suggest that employing padded sequences in the model has no considerable impact on the similarity and diversity of generated sequences. Setting a longer maximum length (20) for generated sequences in BehavGAN_padded may explain why BLEU-4 scores are marginally lower.

Table 4.8 presents comparison metrics in terms of similarity (BLEU-4 score average and variance) and diversity (identical records proportion) for experiments on CASAS and Kastaren datasets. In this table, real data is used as the baseline for comparison. I provide the average BLEU-4 score for the real data, MLE-, LeakGAN-, SeqGAN-, BehavGAN, and BehavGAN_padded-generated data to illustrate that the

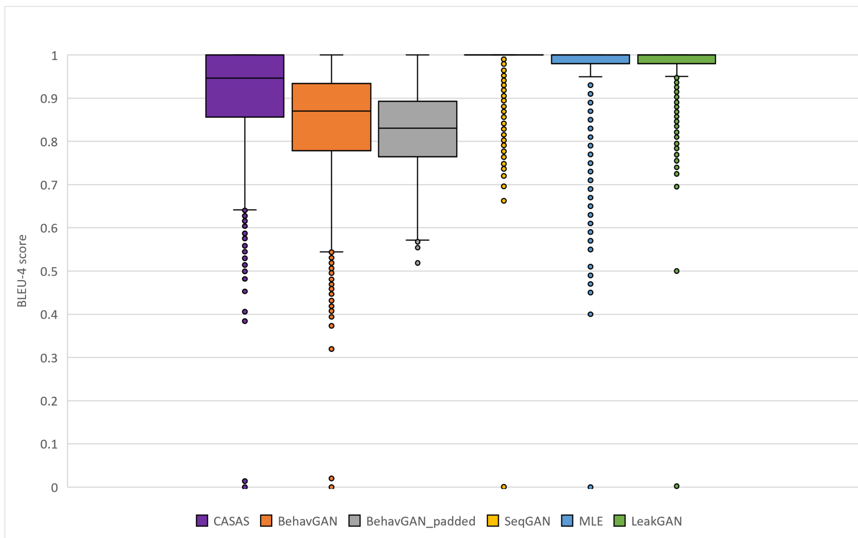


Figure 4.3: Comparison of BLEU-4 Score Distribution for CASAS-Aruba Data with Synthetic Data Generated with MLE, LeakGAN, SeqGAN, BehavGAN, and BehavGAN_padded (BehavGAN with padding) using box plots. A large portion of the generated data using baseline models is too similar to the original data (i.e., Q1, Q2(Median), Q3, and Maximum are all too close), seriously affecting the diversity of the generated data. In comparison to MLE, LeakGAN, and SeqGAN, my technique reduces the proportion of generated sequences that are excessively similar to the real data (BLEU-4 score near 1). An important point to note here is that my proposed algorithm (with and without padding) can generate sequences that are comparable to the real dataset (median is still around 0.85) while avoiding the generation of a significant number of identical records (Q3 is slightly higher than 0.9).
* The circles on the box plots represent outliers.

proposed algorithm is capable of generating a synthetic dataset with a high similarity to the real data. For calculating the BLEU score I consider the real data and the generated data as the reference set and the candidate set, respectively. Moreover, according to this table my proposed reward method improves the output dataset by decreasing identical sequences while maintaining an acceptable similarity rate. I ran each model for five times and report the average value for each reported item. In order to further analyze the ability of BehavGAN in generating interleaved activities, I compute the BLEU-4 score to measure the similarity of generated sequences that include concurrent activities with the reference set. The results (an Average BLEU-4 score of 0.86 with a Variance of 0.08) indicate that generated sequences that include concurrent activities still have high similarity to the reference data. I also investigate the diversity of these sequences by calculating the Identical Record Ratio. Only 14 percent of these sequences are identical to the reference sequences.

Table 4.9 shows how the speed of BLEU score calculation is enhanced by implementing a hash table structure. In this table, the run time of my algorithm (for 150 epochs) with and without the enhancement solution is compared. The runtime of the original SeqGAN algorithm for the same number of epochs and parameters is slightly lower, i.e. 130 mins, which is not a significant difference considering the fact that my proposed algorithm outputs higher quality data.

Effectiveness of BehavGAN

In this section, I perform an experiment to evaluate the effectiveness of BehavGAN in synthesizing behaviour sequences. This experiment is designed to demonstrate the effectiveness of generated data in machine learning tasks. I describe the task and the results of training the model using data augmented by synthesized data vs training the model with real data only.

Bidirectional Encoder Representations for Transformers (BERT) are standard

Table 4.8: Comparison of similarity and diversity metrics on CASAS-Aruba and Kataren datasets.

Dataset	**Algorithm**	**BLEU-4**	**Identical Record Ratio**
	AVG - VAR		
CASAS-Aruba	MLE	97.2% - 0.003	14.2%
	LeakGAN	94.0% - 0.04	49.8%
	SeqGAN	99.4% - 0.00006	47.6%
	BehavGAN	**89.3% - 0.020**	**8%**
	BehavGAN_padded	83.6% - 0.011	8.1%
	Real data	90.8% - 0.013	6%
Kastaren	MLE	91.2% - 0.0025	17.3%
	LeakGAN	95.1% - 0.06	61.2%
	SeqGAN	92.9% - 0.033	56.3%
	BehavGAN	**87.4% - 0.024**	**12%**
	BehavGAN_padded	82.4% - 0.041	7.6%
	Real data	88.5% - 0.027	8.4%

Table 4.9: BLEU score calculation speed

	With hash table	**Without hash table**
Run Time (150 Epochs)	145 mins	4,100 mins

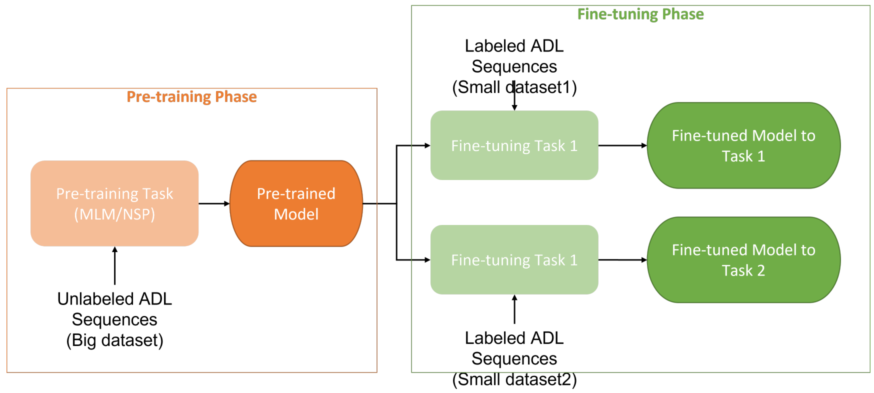


Figure 4.4: Two-Phase training in BERT models using unlabeled and labeled data.

building blocks for training task-specific Natural Language Processing (NLP) models [Devlin et al., 2018]. When fine-tuned utilizing domain-specific labeled data, pre-trained BERT models have been shown to be effective, cost-effective, and time-efficient in addressing downstream tasks [Gu et al., 2021]. This is greatly beneficial since models are pre-trained using general unlabeled data, where labeling is a costly and time-consuming task and little labeled data is available. Subsequently, they can be fine-tuned to a particular supervised task, such as sentiment classification, with a rather small, labeled dataset as illustrated in Figure 4.4.

The input to a BERT model is text/sequence spans, such as sentences divided by special tokens [SEP]. Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) are two tasks used to pre-train the BERT model with unlabeled data to capture the inter-dependencies between words and sentences. BERT can extract several contextual and structural features during pre-training if adequate training data is provided.

Masked Language Modeling (MLM) is the process of masking tokens in a sequence with an arbitrary probability of 15% - 20% with a masking token, [MASK], and instructing the model to fill (predict) that mask with an appropriate token. The goal of the training is to reduce the cross-entropy loss between the original masked tokens and the predicted ones as much as possible. This allows the model to focus on the right (tokens on the right side of the mask) and left (tokens on the left side of the mask) contexts at the same time. Models may learn textual patterns from unlabeled data via MLM, which is employed in pre-training tasks. NSP is used to pre-train the model by having it anticipate the sentence that comes after each one in the training corpus.

In this section, I demonstrate the effectiveness of BehavGAN, my proposed synthetic behavior sequence generation approach, by presenting the results of a MLM task that is trained on both original and synthetic data generated by BehavGAN, as well as synthetic data generated by three baseline methods. For each method, I train separate masked models with training data, which includes 90% of real data augmented by the generated data with the corresponding method. The trained model is then evaluated on test data, which is 10% of the real data. Each sequence is treated as an input to the MLM task. During the training process, 15% of the tokens will be chosen at random and masked. The model is trained to predict masked tokens throughout the training process as shown in Figure 4.5. As a result, the model's low evaluation loss suggests that it has acquired contextual and structural features of the data, making it suitable for use as a pre-trained model for tasks including abnormality detection, next activity prediction, etc. MLM has also been directly used to solve problems like System Log Anomaly Detection [Lee et al., 2021] and Text Denoising [Sun and Jiang, 2019].

For this experiment, I used BERT-base-uncased from Hugging Face library with six attention heads. I run the baseline methods as well as the BehavGAN to generate 10,000 records. Then, I combine synthetic data with 90% of real data. Now, I run the MLM task with each training set. In the evaluation step, the aforementioned trained models will be evaluated using the evaluation set (10 percent of the real data from CASAS). In Figure 4.6, I present the evolution of the masked model during the

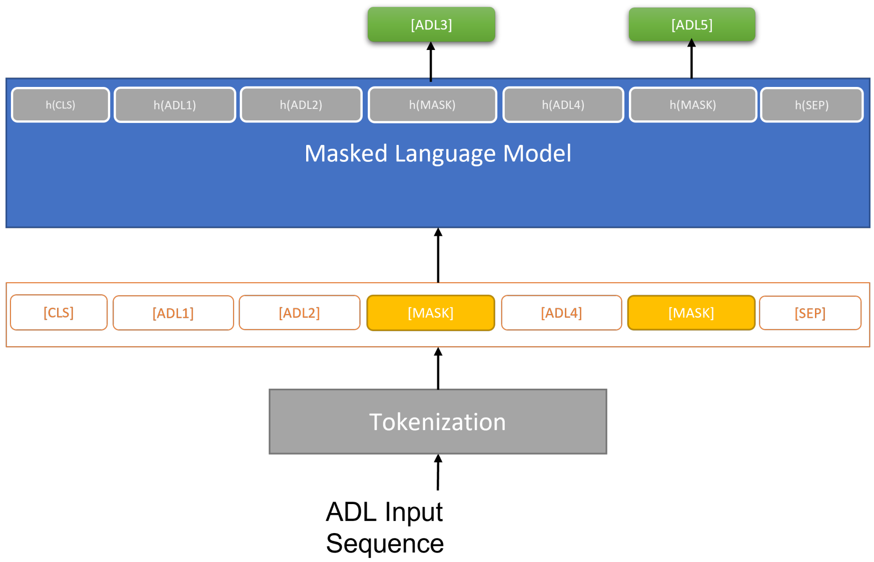


Figure 4.5: An illustration of the Masked Language Modeling task for predicting MASK tokens in ADL sequences.

training steps. As illustrated in this figure, training loss decreases throughout the training process. However, training the model with BehavGAN results in the most consistent and rapid reduction in training loss.

Table 4.10 presents the evaluation loss of the MLM experiment, where the number in front of each method shows the evaluation loss of the MLM on the data generated by that method. Also, the number in front of Real data shows the evaluation loss of the MLM on the original CASAS data. We can deduce from these findings that BehavGan has effectively increased the model's accuracy in predicting masked tokens (0.11 decrease in the Cross-Entropy Loss of the model). Other methods do not show noticeable improvement in training the model.

Human Evaluation

In addition to evaluating the similarity and diversity of generated data using the BLEU score metric, human evaluation was conducted in order to make sure the

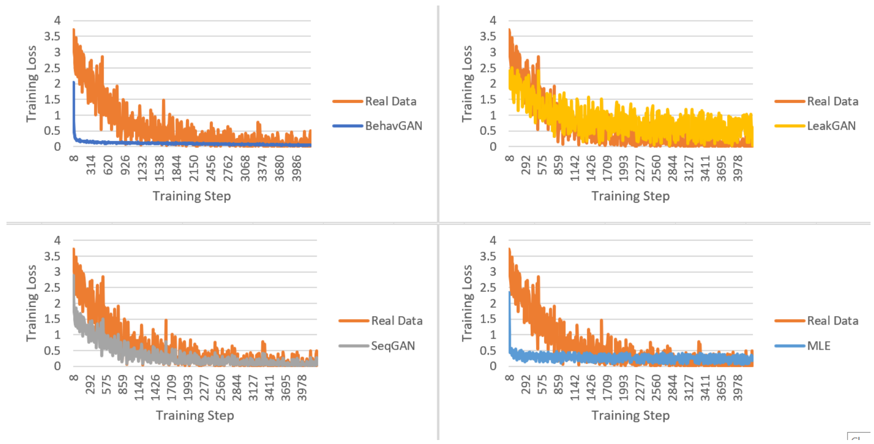


Figure 4.6: A Comparison of Training Loss for the Masked Language Modeling Task using Real Data, MLE-, LeakGAN-, SeqGAN-, and BehavGAN-generated Data.

Table 4.10: The Evaluation Loss of the MLM task - CASAS dataset

Algorithm	**Cross-Entropy Loss**
MLE	0.59
LeakGAN	0.63
SeqGAN	0.62
BehavGAN	**0.53**
Real data	0.64

generated behavior sequences made sense in terms of the order and the duration of activities. To that purpose, I enlisted the help of three raters, who were given 50 sample sequences to rate on a scale of 1 to 5, with the higher the number, the more likely the sequence is perceived to be real. Two raters are Ph.D. students in the field of health management, and one is a Ph.D. student in the field of management sciences. Prior to collecting their judges, I double-checked that they fully comprehended the task. To ensure that they only use their common sense to determine if the presented sequence is real or fake, I anonymized the data and released no detail about the typical ADL patterns in the data. The sample data includes 10 random samples from real data, random samples generated by the baseline methods, 10 samples each, and 10 random samples generated by my proposed method. I excluded generated sequences that are identical to the real data since the BLEU score evaluation revealed that a large proportion of generated sequences by baseline methods was identical to the real dataset. This is not desired when it comes to synthesizing data. Instead, I want the method to be able to generate valid yet diverse sequences. Table 4.11 presents the result of the human evaluation phase. For each model, I report the average score from the three raters. The results indicate that the generated sequences of the BehavGAN make more sense when compared to the MLE-, the SeqGAN- and the LeakGAN-generated sequences. Table 4.12 provides a few samples of sequences that are generated by BehavGAN and the SeqGAN method, the method with best results among the baseline methods, to illustrate the superiority of the BehavGAN output. The SeqGAN-generated sequences simply repeat some frequent sub-sequences, for example, (LeaveHome, EnterHome) or (MealPreparation, Relax), but the method performs weakly in generating various activity sequences and does not follow some common-sense rules such as eating after meal preparation. The reason is the SeqGAN tries to maximize the similarity to ground-truth sequences,

Table 4.11: Human evaluation score

Method	Average Human Score
MLE	2.1
SeqGAN	3.2
LeakGAN	2.6
BehavGAN	**3.7**
Real data	**4.8**

thereby sacrificing its diversity. BehavGAN does show better performance in generating various patterns, and its sequences mostly follow common-sense rules of daily activities. The behaviour sequence in the model I propose can be simulated in the same way as GANs can simulate the meaningful order of words in linguistic models. The reason the generator doesn't generate sequences like "LeaveHome -¿ WashDishes -¿ EnterHome," for example, is that this pattern doesn't appear in the real sequences I supplied the model. Actually, the "LeaveHome" token always follows the "EnterHome" token. However, in comparison to the ground truth sequences, BehavGAN has issues in dealing with sleep duration.

4.2.2 BERT-based BCD Module

In this section, I present the results of running the BERT model for detecting behavior changes of two residents in CASAS-Twor Dataset. First, I train the predictor model with 90 percent of ADL sequences of resident R1. I then evaluate the model using the remaining 10 percent of the unseen ADL sequences of resident R1 (Experiment 1). I also repeat this experiment for resident R2 (Experiment 2).

In two separate experiments, I tested the model for predicting behavior changes of resident R2 without fine-tuning (Experiment 3) and with fine-tuning (Experiment 4) on resident R2 data. In experiment 3, I used a pre-trained model which was trained on ADL sequences of resident R1 to predict behavior changes in resident R2. In experiment 4, I fine-tuned a pre-trained model using ADL sequences of resident R2, where the pre-trained model is trained on ADL sequences of resident R1. The intuition is that while different residents have their unique routines of life, there are commonalities that can be transferred from one model to the other (the Transfer Learning feature in Transformers). I expect the former experiment to show less accurate predictions. The reason is that different residents are supposed to have their unique routine of life which makes it unlikely to precisely predict their behavior change without fine-tuning the model on their specific data.

I also determined the number of training epochs by monitoring training loss and evaluation loss in order to avoid under- or over-fitting. The goal was to train an accurate model with training data (low training loss) that also shows promising performance on the evaluation data (low evaluation loss). The parameters I set for running the BERT-based classifier are listed in Table 4.13.

Table 4.14 shows the results of the four experiments for 10 training epochs. The high values of accuracy, AUC, precision, and recall illustrate the capability of the BERT-based classifier model in predicting behavior changes for both residents (E1 and E2). I acknowledge that training the model from scratch for each new resident is inefficient and possibly impossible. Therefore, experiments E3 and E4 were created to investigate the transfer learning characteristic of Transformers in this specific problem. In experiment E3, I assume that I do not have access to resident R2's ADL data. As a result, I train the model using data from resident R1 and test it using data from resident R2. The model predicts behavior changes well (high accuracy and recall), but it has a significant False Positive rate (noticeably poorer precision than E2), which means the classifier incorrectly labels some normal patterns as abnormal. In experiment E4, I use resident R2's ADL data to fine-tune the trained model from experiment E3. The results show a slight increase in all metrics, which I interpret as the capability of the model to transfer learned knowledge from one resident's behavior to predicting the behavior anomalies of others.

Some sample outputs from the classifier are shown in Table 4.15. The first three samples are correctly predicted as abnormal with relatively high probabilities. The reason is that they have clues of abnormal behavior such as long personal hygiene at night or leaving home without returning, which are not usual behavior of the resident. The next two samples (4 and 5) are also correctly predicted as normal with high probabilities. The last sequence is not classified correctly.

My findings suggest that the BERT-based classifier is capable of detecting behavior changes in ADL sequences. Transfer learning has also proven to be useful in fine-tuning a pre-trained model for a new resident. These results acknowledge the applicability of Transformer models to the behavior change detection problem through analyzing the ADL sequences. It is also a significant finding that transfer learning feature of Transformers is effective in training the models for new residents without requiring a huge amount of data collection and labeling for the new resident.

4.2.3 IRL-based BCD Module

In this subsection, I present the evaluation results of the IRL-based BCD module using the CASAS-Aruba dataset.

I split the CASAS dataset into train and test sets with a 70-30 ratio. Using the train set, I trained the inverse reinforcement learning model to associate reward values to each action (activity class) in a given state (ADL sequence) from sequences of ADLs data. I then evaluated the model's performance on the test set.

Table 4.16 shows the activity codes and their corresponding activity labels. These codes are used to identify different activities that are performed by the individual. For example, the code "4" represents the activity of bathing for a long duration in

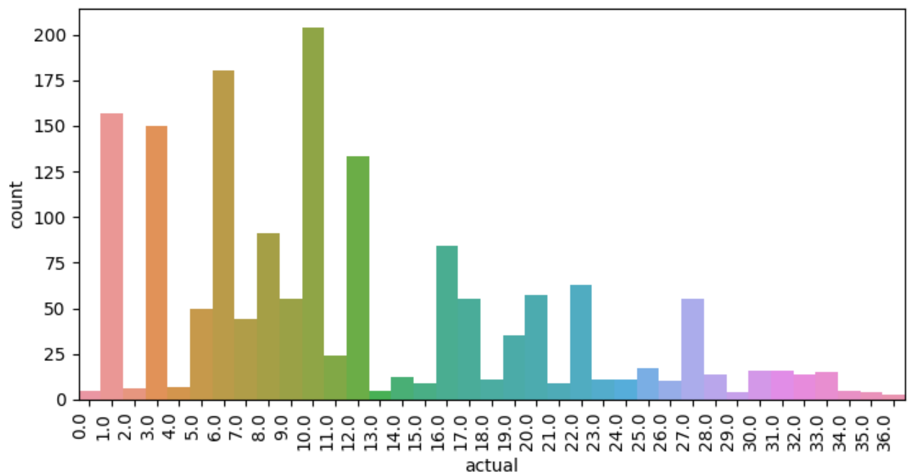


Figure 4.7: The distribution of activity classes in the train set

the morning, and the code "18" represents the activity of transitioning from bed to toilet for a short duration at midnight. Overall, this table serves as a reference to understand the codes that are used to represent these activities in the following graphs.

To provide an overview of the dataset, I first generated a bar chart showing the distribution of data over various activity classes (Figure 4.7). The chart revealed that the dataset is imbalanced, with some activity classes occurring more frequently than others.

I trained the model using **Adam** optimizer, learning rate of 0.001 and **window size** of $W = 10$ on Google Colab environment with a T4 GPU, 14GB of system RAM and 15GB of GPU RAM. I monitored its training progress by tracking the **cross-entropy loss** over the 1000 epochs of training. Figure 4.8 shows the line chart of the model's training loss. As can be seen, the model's loss decreases from 12 to 2.5 over the 1000 epochs of training, indicating that the model is learning to assign higher rewards to next activities that conform to the behavior pattern of the individual and lower rewards to abnormal activities. This suggests that the model is able to capture

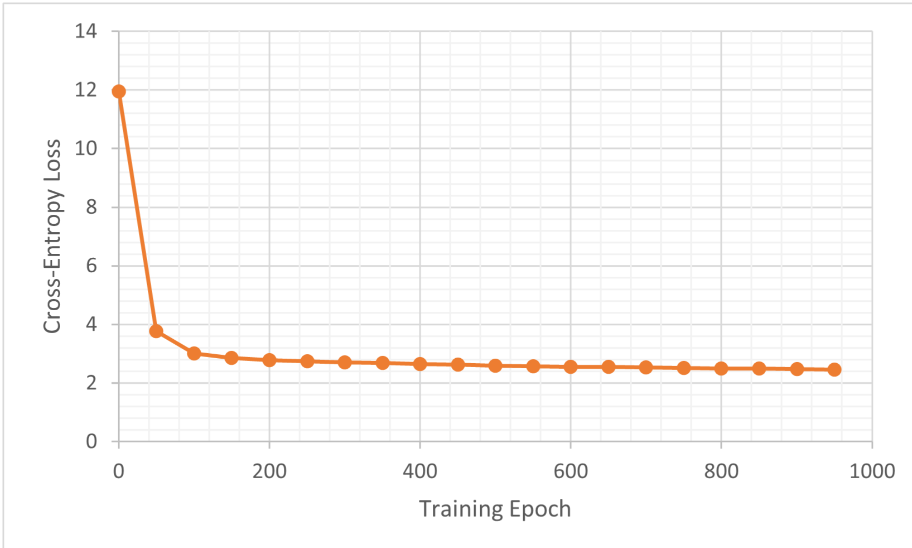


Figure 4.8: Cross-entropy Loss over 1,000 Epochs of Training

the underlying patterns in the data.

The decrease in cross-entropy loss over the training epochs indicates that the model is learning to minimize the difference between its predicted activity classes and the actual activity classes. This is an important feature of the model, as it allows us to detect behavior changes in older adults more accurately and efficiently. Next, I analyzed the model's performance using a HeatMap that displays the normalized average reward for each predicted and actual activity class (Figure 4.9). The X-axis represents the predicted activity class by the IRL model, while the Y-axis represents the actual activity class. The lighter colors in the HeatMap indicate a higher reward, while the darker colors indicate a lower reward. I observed that the Heatmap's diagonal is apparent, which indicates that the trained model is able to give high rewards to activity classes that match the actual activity class. This suggests that the model can correctly identify the majority of activity classes. Additionally, I noticed that there are dark cells corresponding to each actual activity class, which indicates that the model is able to identify activity classes that are not very probable to occur in

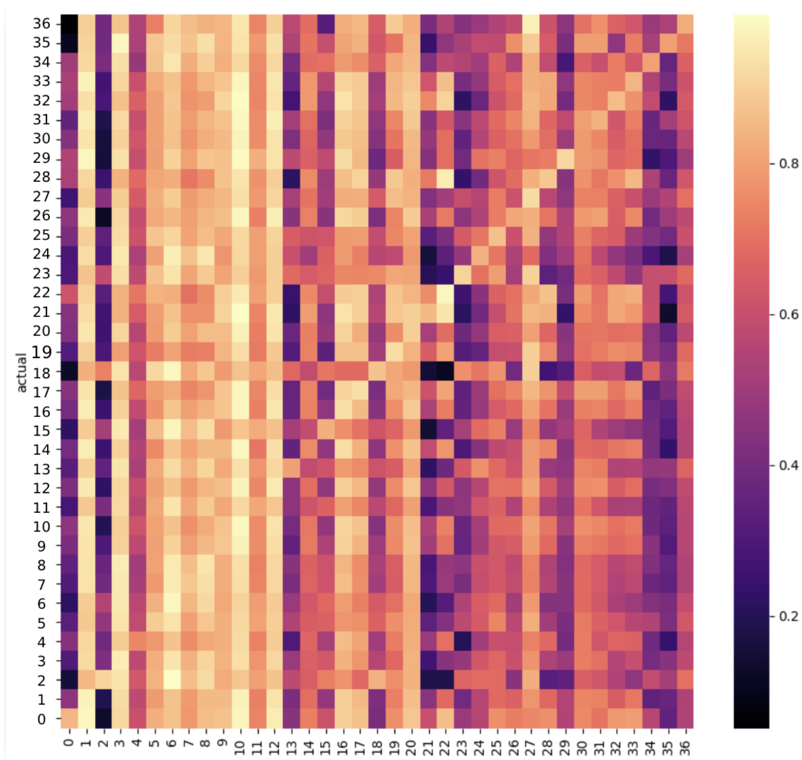


Figure 4.9: The average reward for predicted activity classes in the train set

some states. This is an important feature of the model, as it allows us to identify anomalies in the data that may indicate behavior changes.

However, I also observed that apart from the light cells in the diagonal, there are other light cells present in the HeatMap. This is because, in each state, there is more than one single activity class that is possible to occur due to the diverse nature of the behavior patterns of an individual. This suggests that the model may sometimes predict multiple activity classes with similar probabilities. To further evaluate the model's performance, I define a metric 'Alternative Activity Reward Accuracy (AARA)' to measure how accurately the model assigns high rewards to activity classes that could be alternatives to the true activity classes. I do this by calculating the ratio of high-reward activity classes present in the training set when added to the end of the current state S.

Also, 'Low-Reward Irrelevance Rate (LRIR)' metric calculates the proportion of low-reward activity classes that are not present in the training set when added to the end of the current state S. A higher ratio suggests that a significant portion of low-confidence predictions are correctly identifying irrelevant classes.

$$\text{AARA} = \frac{\sum_{i=1}^{n} \nVdash(r_i > \theta \wedge S\&a_i \in \text{TrainSet})}{\sum_{i=1}^{n} \nVdash(r_i > \theta)} \qquad \text{LRIR} = \frac{\sum_{i=1}^{n} \nVdash(r_i < \theta \wedge S\&a_i \notin \text{TrainSet})}{\sum_{i=1}^{n} \nVdash(r_i < \theta)} \tag{4.2.1}$$

where:

- n is the total number of predictions.
- r_i is the reward for the predicted activity class i.
- θ_h is the threshold for high rewards.
- θ_l is the threshold for low rewards.
- S is the current state (i.e. previous W-1 ADL events)

- a_i is the predicted activity class.
- $\mathbb{1}$ is the indicator function, which is 1 if the condition is true and 0 otherwise.
- TrainSet is the set of ADL sequences in the training data.

θ_h is set to 0.85 while θ_l is set to 0.15 according to reward statistics and the fact that rewards are normalized. An AARA of **0.96** and an LRIR of **0.93** indicate that the model performs significantly well in capturing the behavior patterns.

The HeatMap graph clearly shows that certain activities cannot be substituted with others in a typical situation. For instance, sleeping for a medium duration at night cannot be replaced with bed to toilet transition, eating, or a long personal hygiene. Additionally, less frequently occurring activities like wandering in the room are generally associated with lower average reward values, except when they actually occur. This is reflected in the graph as the diagonal cells for such activities remain light, while almost all other cells in the column are dark.

To ensure that my IRL model was able to generalize well, I evaluated its performance on a separate test set that was not used during the training phase. I used the same bar chart and HeatMap visualizations to demonstrate the test set performance, as I did for the training set. Figure 4.10 shows the bar chart for the test set, which has a similar distribution of data across the various activity classes as the training set. This indicates that the test set is representative of the overall dataset and that the model is able to generalize effectively beyond the training data. Overall, the evaluation of the IRL model on the test set provides further evidence of its robustness and effectiveness in accurately predicting activity patterns and rewards for residents in smart homes. Figure 4.11 shows the HeatMap for the test set, which displays similar results as the HeatMap for the training set. The diagonal is apparent, indicating that the model is able to accurately predict the majority of activity classes, and there are dark cells corresponding to each actual activity class, suggesting that the model is

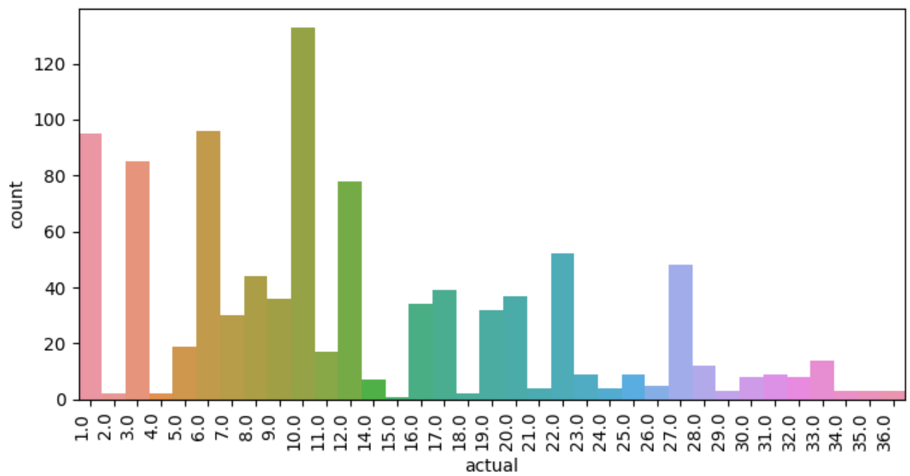


Figure 4.10: The distribution of activity classes in the test set

able to identify activity classes that are not very probable to occur in some states. Additionally, there are light cells present in the HeatMap, which suggests that the model is able to predict multiple activity classes with similar probabilities in some states.

The fact that my IRL model produced similar results for the test set as for the training set suggests that the model is not overfitted to the training data and can effectively generalize to new, unseen data. This is a crucial feature of the model, as it enables us to apply it to new datasets with confidence, thereby improving our ability to detect behavior changes in older adults more accurately and efficiently.

The robustness of my IRL model is particularly important in the context of home-care, where residents' behavior patterns can vary widely and change over time. By accurately predicting these patterns and detecting any changes, my model can help caregivers and researchers to better understand the needs and preferences of individual residents, and to develop tailored interventions that improve their quality of life. Overall, the ability of my IRL model to effectively generalize to new datasets is a significant advantage that enhances its practical utility in real-world care settings.

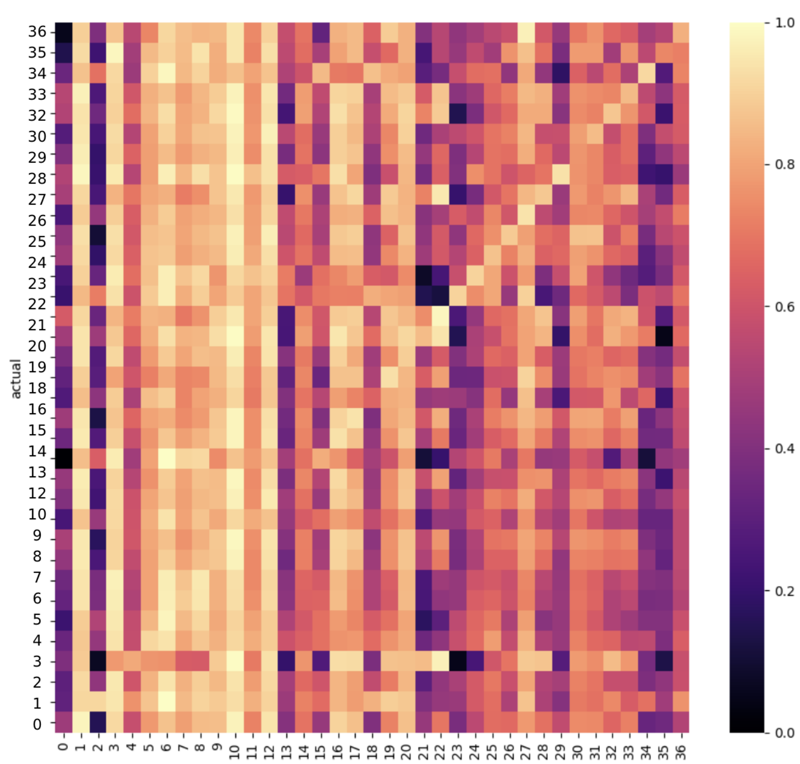


Figure 4.11: The average reward for predicted activity classes in the test set

Data Augmentation

To evaluate the ability of my models to identify behavior changes, I introduced synthetic abnormal sequences into the dataset. Based on existing literature, changes in physical activity levels, alterations in rest periods between tasks, changes in sleep patterns, forgetting to complete tasks, and repeating tasks are all included in the symptom profiles of diseases such as Alzheimer's, heart disease, urinary tract infections, diabetes, and others. I therefore introduced these changes into the ADL sequences in the CASAS dataset. I also inject samples of behavior abnormalities by rearranging ADL and manipulating activity duration. For example, while in the original ADL sequences eating occurs after meal preparation, I reversed the ADL order to inject partially misordered sequences. I also created some abnormalities by randomly shuffling the ADL. After injecting %10 synthetic abnormal sequences, SMOTE method is used to oversample the abnormal sequences to ensure a balanced dataset. I used the augmented labeled dataset to evaluate the performance of the fusion module.

Table 4.17 presents the performance metrics for the proposed approach, including accuracy, precision, recall, and F1 score, at different threshold values. The results show that using lower threshold values increases the number of false positives, indicating that more normal ADL sequences are incorrectly classified as abnormal. Conversely, higher threshold values result in a decrease in recall, indicating that the model is more likely to miss abnormal cases. I recommend selecting a threshold value that balances precision and recall. To aid in this decision, I also report the F1 score, which is the harmonic mean of precision and recall. This score provides a single metric that combines both precision and recall, making it useful for selecting an appropriate threshold.

To compare the performance of my proposed BCD models with baseline models, I implemented an LSTM classifier and a Transformer-based classifier with the same sequence size. In my implementation of the LSTM model, I adopted the architecture proposed by [Zerkouk and Chikhaoui, 2019], utilizing a hidden size of 64 and an embedding size of 200. For the BERT model, I employed the 'AutoModelForSequenceClassification' from the Hugging Face Transformers library, using the pretrained 'bert-base-uncased' model.

According to Table 4.18, in my analysis, I observed that recall is significantly higher than precision for both the LSTM and BERT models. This discrepancy suggests a lower performance on the positive class, as indicated by a higher false positive rate. In contrast, the IRL model demonstrates superior overall performance when evaluating the F1 score, which provides a balanced measure of precision and recall.

The experimental results suggest that both BERT and LSTM are not particularly effective in the given context, despite their demonstrated success in various other applications. This limitation can primarily be attributed to the insufficient size of the training dataset, which is inadequate for effectively training deep neural network models such as LSTM and BERT. Furthermore, based on the performance of the BERT-based model described in the previous subsection, it can be concluded that processing longer input sequences (e.g., 128 events) is necessary for Transformer-based models to achieve satisfactory performance. However, this requirement represents a significant limitation, particularly for real-time (or near real-time) use cases such as the current problem, where it is impractical to wait for the observation of 128 events before detecting abnormalities in ADL. Additionally, the pre-training of BERT involves general-domain text sequences, which may not translate well to the specific nuances of ADL sequences without substantial fine-tuning. LSTM networks also face challenges in managing longer sequences, which can further affect performance.

To enhance the adaptability of Transformer models like BERT to specialized tasks, future research should prioritize fine-tuning with domain-specific data. This approach could potentially mitigate the limitations observed in my study and improve model efficacy in similar applications.

The results demonstrate that my IRL model effectively detects behavior changes in older adults from ADL sequences with high accuracy. The model successfully identifies activity classes unlikely to occur in certain states, aiding healthcare professionals in detecting anomalies and potential behavior changes.

My evaluation on the augmented dataset highlights the effectiveness of my approach in identifying behavior changes linked to various diseases. By accurately detecting these changes, my method has the potential to enhance the quality of care provided to residents in smart homes.

4.2.4 Note Generation Module

In this section, I present the experimental results of note generation using large language models (LLMs). To begin, I provide a summary of various symptoms associated with different health conditions, which can be identified through the monitoring of ADLs and their related changes. Next, I showcase example notes generated by applying various prompting strategies.

Health Abnormalities via Sensors

Based on domain-specific knowledge of diseases and symptoms, abnormalities in the behavior sequences of older adults can be detected using data from ambient sensors monitoring daily activities like eating, sleeping, bathing, and meal preparation. These abnormalities often correspond to common health conditions such as cognitive decline, depression, and mobility issues. For instance, irregular sleep patterns, forgetfulness, or repetitive actions may indicate cognitive impairments, while changes in appetite or reduced activity levels might suggest depression. Similarly, decreased physical activity or difficulty completing tasks could point to arthritis or other mobility challenges.

Understanding these patterns enables early identification of potential health concerns.

Each abnormality is linked to specific health conditions and supported by research references, providing a comprehensive framework for interpreting sensor data. Conditions like Parkinson's disease, cardiovascular issues, or diabetes exhibit distinct behavioral indicators, such as tremors, fatigue, or frequent urination, respectively. Additionally, sleep disorders, malnutrition, and social isolation manifest through disruptions in routines, such as altered sleep patterns or reduced social interactions. By leveraging this knowledge, caregivers and healthcare systems can use sensor data to implement timely interventions, improving the quality of life for older adults. The following table summarizes these health conditions and their associated abnormalities for quick reference.

Engineering Prompts: Translating ADL Logs into Descriptive Notes

Designing an effective prompt for a large language model to interpret ADL logs and generate meaningful notes for health professionals requires a structured approach. The goal is to provide the LLM with sufficient context, baseline data, and specific abnormalities to ensure the output is both accurate and actionable. Below, I outline the key elements of the prompt engineering process used in my research.

- **Contextualizing the Input**

  To enable the LLM to generate insightful notes, the first step is to establish a clear context. This involves framing the ADL logs within the broader scope of the patient's health and daily routine. For example, the prompt begins with a summary of the patient's baseline behavior over a defined period, highlighting normal patterns for key activities such as sleeping, eating, bathing, and mobility. This allows the LLM to differentiate between expected variations and significant deviations.

Example Prompt Section: "Patient X's baseline activity patterns over the past two weeks include 8 hours of sleep nightly, meals prepared and consumed at regular intervals (three times daily), and consistent bathroom visits occurring 6–8 times per day. Mobility patterns show regular movement between kitchen, living room, and bedroom during waking hours."

- **Providing Detected Behavior Changes**

  The second component is the presentation of detected behavior changes. These are deviations from the baseline activity patterns identified by the behavior change detection system. Each behavior change is described with specifics, such as sequence of activities, timing, and duration. This precision helps the LLM understand the nature and extent of the deviations.

  Example Prompt Section: "Here is the sequence of activities on Saturday Jan 12, which is detected as a potential abnormality: 12 AM- 4:11 AM: Sleeping 4:11 AM- 4:30 AM: Personal Hygiene 4:11 AM- 4:18 AM: Wandering in Room ..."

- **Instructing the LLM to Generate Notes**

  The final step involves instructing the LLM to synthesize the provided information into a concise and informative note for health professionals. The instruction emphasizes the need to explain the potential significance of the abnormalities and suggest possible health concerns or conditions that may warrant further investigation.

  Example Prompt Section: "Based on the patient's baseline patterns and the detected behavior changes, generate a note for health professionals. The note should:

  - Summarize the detected behavior changes.

- Explain potential underlying health conditions linked to these deviations.
- Highlight the need for further clinical evaluation, where applicable.
- Use clear and professional language appropriate for healthcare documentation."

Example Output

Generated Note: "The patient has exhibited significant deviations from baseline activity patterns over the past week. Specifically, there is increased nighttime activity (2:00 AM–4:00 AM), which may suggest sleep disturbances or disorientation. Additionally, the absence of meal preparation for three days raises concerns about potential depression, cognitive decline, or malnutrition. Prolonged inactivity during waking hours could indicate fatigue, reduced mobility, or acute illness. These abnormalities warrant further assessment to determine underlying causes and inform appropriate interventions."

- **Few-Shot Prompting for Enhanced Note Generation**

  To improve the accuracy and contextual relevance of the LLM's output, I incorporated a few-shot prompting strategy. This technique involves providing the model with several examples of high-quality input-output pairs that demonstrate how to transform ADL logs and detected behavior changes into descriptive notes for health professionals. By doing so, the model learns the desired style, tone, and structure of the output, increasing its ability to generate consistent and actionable insights.

  Rationale for Few-Shot Prompting

  Few-shot prompting serves to anchor the LLM's responses by offering concrete examples of well-crafted notes. This reduces the likelihood of irrelevant or overly generic outputs and ensures the generated notes are tailored to healthcare professionals' needs. It is particularly valuable when interpreting complex data like ADL logs, where nuanced patterns must be linked to potential health conditions.

Structuring Few-Shot Examples

Each example in the prompt follows a consistent structure:

- Input: context (baseline ADL logs), detected behavior changes, and the task instruction.
- Output: a detailed, professional note that summarizes abnormalities, links them to possible health concerns, and suggests further steps for evaluation.

Benefits of Few-Shot Prompting

By incorporating these examples, the LLM can better:

- Recognize patterns and align them with relevant health conditions.
- Maintain a professional tone suitable for clinical documentation.
- Adapt its responses to reflect variability in patient behaviors and abnormalities.

- **Implementation in Practice**

  For real-world use, the few-shot examples can be tailored to reflect specific patient demographics, healthcare contexts, and data collection methods. This approach ensures that the model generates outputs that are both relevant and actionable, enhancing its utility as a tool for translating ADL logs into meaningful healthcare insights.

Few-shot prompting, when combined with a well-structured base prompt, provides a robust framework for leveraging LLMs in health data interpretation while maintaining consistency and reliability in generated outputs.

By structuring the prompt with these elements, the LLM can effectively bridge the gap between raw sensor data and actionable insights, providing healthcare professionals with a clearer understanding of the patient's condition while preserving their decision-making authority.

Table 4.12: Sample real sequences and sample sequences from the baseline methods and the proposed method

Method	**Sample Sequence**
SeqGAN	Relax133;MealPreparation20;Relax31;LeaveHome2;EnterHome137; MealPreparation36; Relax138;MealPreparation69;Relax199;Sleeping62
SeqGAN	LeaveHome2;EnterHome101;Relax221;MealPreparation26;Relax125; MealPreparation85; Relax282;Sleeping367;BedtoToilet1;Sleeping43
SeqGAN	Relax145;LeaveHome2;EnterHome144;LeaveHome2;EnterHome101; LeaveHome2;EnterHome150;MealPreparation10;Relax140;LeaveHome2
SeqGAN	MealPreparation41;Relax129;MealPreparation62;Relax47; MealPreparation63;Relax84; LeaveHome1;EnterHome159;LeaveHome1; EnterHome138
BehavGAN	Eating9;Relax74;Work27;MealPreparation22;Relax51;Sleeping183; BedtoToilet1;Sleeping180;BedtoToilet3;Sleeping117
BehavGAN	LeaveHome1;EnterHome128;WashDishes6;Relax26;MealPreparation61; Eating32;Relax94;Sleeping452;BedtoToilet2;Sleeping148
BehavGAN	EnterHome153;Eating17;Relax87;Work58;MealPreparation86;Relax26; Eating36;Sleeping171;BedtoToilet3;Sleeping224
BehavGAN	Eating6;Relax141;WashDishes4;Relax70;Sleeping64;BedtoToilet2; Sleeping311;MealPreparation72;Relax103;Eating45
BehavGAN	MealPreparation11;Realx7;MealPreparation45;Relax3;MealPreparation18; Relax19; MealPreparation6; Relax1;Eating&Relax15;Relax9
CASAS-Aruba	LeaveHome2;EnterHome145;Relax234;Housekeeping9;Relax89;Work20; Relax340;Sleeping343;BedtoToilet4;Sleeping377
CASAS-Aruba	Relax141;MealPreparation17;Eating19;WashDishes4;Relax51;LeaveHome1; EnterHome125; MealPreparation84;Relax287;LeaveHome1
CASAS-Aruba	MealPreparation15;Eating9;MealPreparation57;Eating13;Relax72;Eating16; Relax138; Housekeeping19;Work66;MealPreparation33
CASAS-Aruba	Relax313;Sleeping359;BedtoToilet1;Sleeping350;MealPreparation36;Relax87; Eating62;WashDishes7; Relax22;LeaveHome2
CASAS-Aruba	Relax61;Sleeping58;MealPreparation18;Relax1;Eat&Relax7;Relax59; Meal-Preparation6; Relax3;MealPreparation25;Relax4; Eating12

Note: each event is represented by the type of activity performed followed by its duration in minutes. Events are separated by ";".

Table 4.13: Run parameters for BERT model

Parameter	Value
Training Epochs	10
Loss function	Cross Entropy
Learning Rate	2e-5
Warm up Proportion	0.1
Drop-out rate	0.1
Max Sequence Length	128

Table 4.14: Comparison of Experiments.

	Evaluation Metrics				
Experiments	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
E1 : Classifier for R1 ADLs trained on R1 ADLs	0.87	0.89	0.84	0.86	0.88
E2 : Classifier for R2 ADLs trained on R2 ADLs	0.82	0.88	0.75	0.81	0.83
E3 : Classifier for R2 ADLs trained on R1 ADLs	0.81	0.64	0.90	0.75	0.79
E4 : Classifier for R2 ADLs trained on R1 ADLs and fine-tuned on R2 ADLs	0.84	0.69	0.90	0.78	0.82

Table 4.15: Samples of BERT-based ADL classifier output.

			Prediction Probabilities and Label		
ID	**Input ADL Sequence**	**True Label**	**Class ”0”**	**Class ”1”**	**Label**
(1)	PersonalHygieneLongNight WorkShortNight SleepShortNight PersonalHygieneShortMidNight WorkShortMidNight LeaveHomeShortMidNight	1	0.01	0.99	1
(2)	PersonalHygieneMediumMidNight WorkShortMidNight LeaveHomeShortMidNight	1	0.33	0.66	1
(3)	SleepShortNight PersonalHygieneMediumMorning BathingMediumMorning MealPreparationShortMorning LeaveHomeShortMorning	1	0.31	0.69	1
(4)	EnterHomeShortNight WorkShortNight PersonalHygieneShortNight SleepShortMidNight BedToiletTransitionShortMidNight PersonalHygieneShortMorning LeaveHomeShortMorning EnterHomeShortMorning	0	0.9989	0.0010	0
(5)	SleepShortNight WorkShortNight SleepShortNight BedToiletTransitionShortNight SleepShortNight WorkShortMorning SleepShortMorning WatchTVShortNight	0	0.81	0.19	0
(6)	WorkShortNight LeaveHomeShortNight EnterHomeShortMidNight SleepShortMidNight LeaveHomeShortMidNight	1	0.51	0.49	0

”0” is the Normal Class, and ”1” is the Abnormal Class.

Activity Label	**Activity Code**
Bathing (Long, Morning)	4
Bathing (Medium, Morning)	24
Bathing (Short, Morning)	15
Bed to Toilet Transition (Short, Midnight)	18
Bed to Toilet Transition (Short, Night)	2
Eating (Medium, Night)	21
Eating (Short, Night)	32
Enter Home (Short, Midnight)	33
Enter Home (Short, Morning)	14
Enter Home (Short, Night)	9
Leave Home (Medium, Midnight)	36
Leave Home (Short, Midnight)	19
Leave Home (Short, Morning)	8
Leave Home (Short, Night)	30
Meal Preparation (Medium, Night)	26
Meal Preparation (Short, Midnight)	28
Meal Preparation (Short, Morning)	7
Meal Preparation (Short, Night)	20
Personal Hygiene (Long, Morning)	25
Personal Hygiene (Long, Night)	34
Personal Hygiene (Medium, Midnight)	23
Personal Hygiene (Medium, Morning)	5
Personal Hygiene (Medium, Night)	31
Personal Hygiene (Short, Midnight)	27
Personal Hygiene (Short, Morning)	3
Personal Hygiene (Short, Night)	12
Sleep (Medium, Night)	29
Sleep (Short, Midnight)	17
Sleep (Short, Morning)	11
Sleep (Short, Night)	1
Wandering in Room (Short, Morning)	13
Watch TV (Short, Morning)	35
Watch TV (Short, Night)	16
Work (Short, Midnight)	22
Work (Short, Morning)	6
Work (Short, Night)	10
Wandering in Room (Short, Night)	0

Table 4.16: The Mapping of Activity Classes to Activity Codes

Threshold	**Accuracy**	**Recall**	**Precision**	**F1**
0.6	0.73	**0.87**	0.65	0.744
0.75	**0.74**	0.84	0.66	0.739
0.85	0.73	0.81	0.72	**0.762**
0.9	**0.74**	0.76	**0.73**	0.744

Table 4.17: Evaluation metrics for different thresholds.

Model	**Accuracy**	**Recall**	**Precision**	**F1**
LSTM	0.55	**0.89**	0.54	0.672
BERT	0.51	0.86	0.58	0.693
IRL(th=0.85)	**0.73**	0.81	**0.72**	**0.762**

Table 4.18: Evaluation metrics for baseline models.

Table 4.19: Abnormalities Detected and Associated Health Conditions

Condition	**Abnormalities and Explanation**
Cognitive Decline	Irregular sleep patterns, wandering, forgetfulness, repetitive actions, disorientation. Memory loss and confusion affect daily activities [Burns and Iliffe, 2009].
Depression	Changes in appetite, altered sleep, reduced activity, neglect of hygiene. Lack of motivation impacts daily routines [Fiske et al., 2009].
Parkinson's Disease	Bradykinesia, reduced mobility, tremors affecting tasks. Motor symptoms impair speed and coordination [Jankovic, 2008].
Sleep Disorders	Insomnia, sleep apnea, restless leg syndrome. Disrupted sleep leads to fatigue and altered activities [Crowley, 2011].
Malnutrition	Decreased food intake, weight loss indicators. Results from reduced appetite or difficulty preparing meals [Morley, 2012].
UTIs	Increased bathroom visits, restlessness or agitation. Urinary frequency and discomfort disrupt activity [Rowe and Juthani-Mehta, 2013].
Arthritis	Reduced activity, difficulty with physical tasks. Pain and stiffness limit mobility [Hunter, 2011].
Cardiovascular Diseases	Fatigue, reduced activity, shortness of breath. Fatigue and decreased stamina impact routines [Moser and Riegel, 2008].
Diabetes	Frequent urination, changes in eating habits, fatigue. Symptoms affect energy levels and physiological needs [Kirkman et al., 2012].
Social Isolation	Reduced social interactions, increased sedentary behavior. Linked to decreased activity and mental health risks [Nicholson, 2012].
Medication Mismanagement	Inconsistent routines, frequent kitchen visits without meals. Confusion in medication schedules alters daily tasks [Pretorius et al., 2013].
Acute Illness	Sudden activity changes, prolonged inactivity. Immediate physical changes affect activity levels [Clegg et al., 2013].
Dehydration	Reduced fluid intake, confusion, disorientation. Cognitive impairment reduces activity [Schols et al., 2009].
Falls	Sudden stop in movement, irregular patterns. Falls cause injuries, limiting mobility [Rubenstein, 2006].
BPSD	Agitation, restlessness, aggression. Linked to dementia-related behaviors disrupting routines [Kales et al., 2015].

Example 1
Input **Baseline:** Patient Y exhibits regular mobility, with average walking intervals of 5–10 minutes every hour and consistent meal preparation three times daily. **Abnormalities:** Here is the sequence of activities on Saturday Jan 12, which is detected as a potential abnormality: 12 AM- 7:21 AM: Sleeping 7:21 AM- 7:34 AM: Personal Hygiene 7:36 AM- 8:28 AM: Watch TV ... 4:36 PM- 6:20 PM: Sleep not in Bed **Instruction:** Generate a professional note explaining these abnormalities and their potential implications.
Output The patient has shown prolonged periods of inactivity during afternoon hours, deviating from their usual baseline of regular mobility. This could indicate fatigue, early signs of depression, or physical discomfort such as joint pain or stiffness. These observations warrant further clinical evaluation to identify potential contributing factors.

Table 4.20: Example 1 - Input-output formatting for clinical abnormality analysis.

Example 2
Input **Baseline:** Patient Z's baseline activities include preparing meals twice daily and minimal bathroom usage at night (1–2 visits). **Abnormalities:** Here is the sequence of activities on Saturday Jan 12, which is detected as a potential abnormality: 12 AM- 2:11 AM: Sleeping 2:11 AM- 2:24 AM: Personal Hygiene 2:26 AM- 4:28 AM: Sleeping 4:29 AM- 4:35 AM: Personal Hygiene 4:35 AM- 5:48 AM: Sleeping 5:49 AM- 5:53 AM: Personal Hygiene 5:55 AM- 7:28 AM: Sleeping 7:30 AM- 7:35 AM: Watch TV ... 3:36 PM- 4:20 PM: Watch TV **Instruction:** Generate a professional note explaining these abnormalities and their potential implications.
Output Recent data indicates a complete lack of kitchen activity for two days, which may point to decreased appetite, physical discomfort, or cognitive challenges in meal preparation. Additionally, a significant increase in nighttime bathroom usage suggests potential urinary tract infection (UTI) or diabetes-related symptoms. Further assessment is recommended to explore these possibilities and provide appropriate intervention.

Table 4.21: Example 2 - Input-output formatting for clinical abnormality analysis.

Chapter 5

Discussion

As the world's population ages, enabling proactive, personalized healthcare for older adults has become imperative. Our research pursues an end-to-end framework for activity-based monitoring in smart homes, comprising (1) synthetic data generation, (2) behavior change detection, (3) and automated note generation with human oversight. Each component addresses a critical bottleneck, from scarce labeled data to ethical deployment, while collectively advancing toward reliable, scalable AI-driven eldercare.

5.1 Synthetic Data Generation

Publicly available datasets of older adults' ADLs remain limited, hampering research into early health-degradation signals. To bridge this gap, we adapted Generative Adversarial Networks (GANs), proven on images, text, and music, to synthesize realistic ADL sequences. Rather than raw sensor traces, our GAN works on tokenized activity representations, allowing us to generate diverse, privacy-preserving behavior logs.

5.1.1 Key Findings and Hyperparameter Insights

- **Thresholding for Diversity:** we found that setting the discriminator threshold inversely to the token-space size is critical: a lower threshold discourages mere memorization and encourages novel sequence generation.

- **Sequence Volume Control:** limiting the total number of synthetic sequences prevents mode collapse, over-repetition of a few patterns.

- **Similarity Metric Choice:** BLEU-4 proved appropriate for our mid-length sequences; however, it may underperform on very short or very long activity logs, suggesting alternative sequence-distance measures for other contexts.

5.1.2 Limitations and Future Directions

For future research, behavior representation needs further improvement to capture more features of an individual's life. Due to the unavailability of a real dataset containing more features such as vital signs or health status we will await future developments of this nature.

5.2 Behavior Change Detection

Detecting gradual or abrupt deviations in daily routines can flag early health issues, such as cognitive decline or mobility impairment. We proposed transformer-based classifiers trained on tokenized ADL logs, leveraging their capacity for long-range dependency modeling alongside an IRL-based model capable of near real-time behavior change detection without requiring long observational sequences.

5.2.1 Training Paradigms: From-Scratch vs. Fine-Tuning

- **From Scratch (Domain-Specific Pretraining):** building a BERT-style model solely on ADL sequences allows the masked language modeling (MLM) and next-sentence prediction (NSP) objectives to learn domain semantics.
- **Fine-Tuning Pretrained Models:** applying a general-domain BERT (or RoBERTa, GPT-3, BERT-Large) and fine-tuning on ADL data benefits from large-scale language understanding but may misinterpret activity order divorced from natural language logic.

A systematic comparison of these paradigms will reveal whether domain pretraining yields superior sensitivity to subtle behavior shifts.

5.2.2 Data Scarcity and Synthetic Augmentation

Supervised behavior classifiers are bottlenecked by scarce labeled "abnormal" ADL sequences. In our study, we induced artificial irregularities by shuffling activity orders, varying durations, and injecting patterns associated with health-related symptoms. Future experiments should validate transformer performance on real abnormal sequences collected from individuals with known health conditions, ideally leveraging the GAN-generated data to augment rare event classes. Furthermore, while my experiments indicate that learned knowledge can transfer from one resident to another, additional investigations are needed to assess the generalizability of these models. In particular, evaluating the performance of pretrained models on new residents remains an open question, as this study was limited by the lack of multi-resident data.

Another limitation of this study is that inaccuracies in ADL recognition may propagate errors into the behavior change detection module, directly affecting its performance.

5.2.3 Limitations of Sequence Length and State Space

While the proposed IRL-based model represents a significant advancement in behavior change detection, it is important to acknowledge a critical limitation: as the length of ADL sequences increases, the corresponding size of the state space grows exponentially. This phenomenon, rooted in the underlying Markov Decision Process (MDP) formulation, poses significant challenges to model training.

Specifically, longer sequences result in a combinatorial explosion of possible state transitions, which increases the demand for extensive training data to adequately represent and generalize across the expanded state space. This becomes particularly problematic in real-world settings where annotated ADL data is scarce or unevenly distributed.

Despite this limitation, our approach remains promising. Integrating inverse reinforcement learning techniques and transformer-based architectures can be further explored to uncover subtle shifts in resident behavior. These can support earlier interventions and ultimately enhance quality of life in smart home environments.

5.3 LLM-Based Note Generation with Human-in-the-Loop

Automated summarization of ADL logs into clinical notes promises to reduce clinician burden. However, large language models can hallucinate, embed bias, and lack clinical reasoning. By integrating a HITL framework, AI drafts are systematically reviewed, corrected, and approved by healthcare professionals, ensuring safety, accuracy, and accountability.

5.3.1 Enhancing Trust and Reducing Bias

- **Clinical Oversight:** every AI-generated note passes through a clinician before affecting care decisions, guarding against misdiagnoses.
- **Bias Feedback Loops:** providers flag biased or inaccurate phrasing, feeding corrections back into reinforcement learning from human feedback (RLHF) to iteratively de-bias the model.
- **Automation Bias Mitigation:** enforced verification prevents over-trust in AI outputs, fostering a collaborative "AI as assistant" paradigm rather than "AI as replacement."

5.3.2 Ethical, Legal, and Interpretability Considerations

HITL ensures final medical judgments remain human-driven, aligning with HIPAA and GDPR. Context-aware interpretations, accounting for comorbidities or living environments, further ground AI summaries in real patient circumstances.

5.4 Data Availability and Privacy

5.4.1 Sensor-Based vs. Video Monitoring

- **Privacy Advantages:** ambient sensors such as motion detectors and pressure mats can capture ADLs without relying on visual recordings, thereby preserving personal dignity and mitigating ethical concerns. However, their limited granularity compared to video monitoring can reduce the accuracy of ADL detection.
- **Cost and Efficiency:** sensor data demands far less storage and compute than continuous video, enabling scalable, real-time anomaly detection with minimal

infrastructure.

5.4.2 Challenges in Acquiring Paired Training Data for Note Generation

Effective LLM fine-tuning requires parallel corpora of sensor logs and clinician-written notes, yet privacy regulations, fragmented sensor ecosystems, and patient consent barriers severely limit dataset sizes and standardization.

Chapter 6

Conclusion

This thesis presented a comprehensive and interpretable AI-driven framework for monitoring behavior changes in care environments using non-intrusive sensor data and advanced machine learning techniques. The research was motivated by the increasing need for early detection of health-related behavioral anomalies among older adults living independently. The proposed system integrates behavior modeling, anomaly detection, synthetic data generation, and large language model (LLM)-based interpretation, all under a human-in-the-loop (HITL) paradigm. The framework is grounded in Fogg's Behavior Model, ensuring that technological interventions align with theoretical insights on behavior motivation and capability.

6.1 Addressing the Research Questions

This research presents a comprehensive framework for detecting behavior changes in older adults using machine learning, addressing key challenges related to detection efficacy, evaluation methods, model generalizability, interpretability, and system reliability.

- Detection Efficacy(RQ1): this thesis developed and evaluated two complementary BCD modules: an IRL-based model that learned reward structures indicative of normative behavior, and a Transformer-based model that captured temporal dependencies in ADL sequences. Both approaches demonstrated strong performance in identifying deviations from baseline behaviors, validated using CASAS datasets. The models were able to detect subtle anomalies, such as prolonged inactivity or irregular sleep patterns, which often precede health events.

- Evaluation Framework(RQ2): to systematically assess the effectiveness of the approach, I designed simulation experiments using augmented ADL datasets where behavioral deviations—such as frequent restroom use or prolonged inactivity—were synthetically injected to mimic real health events (e.g., urinary tract infections). I introduced BehavGAN, a novel GAN-based model for generating realistic ADL sequences in data-scarce scenarios. This simulation protocol enables consistent evaluation through quantifiable performance metrics and reproducible experimental design.

- Model Generalizability(RQ3): to explore the generalizability of behavior models, I applied transfer learning techniques that adapt models trained on one resident's ADL data to new users. Specifically, fine-tuning a pre-trained BERT classifier on limited labeled data from a new resident showed that high detection accuracy can be achieved with minimal personalization. These results validate the feasibility of cross-user model transfer for real-world deployment.

- LLM Inference(RQ4): to investigate how large language models (LLMs) can be used to derive clinically meaningful insights, I developed a pipeline that transforms behavior change signals and contextual data (e.g., demographics, medication records, ADL logs) into structured prompts. LLMs such as GPT-4, Gemini, or Claude are then used to generate evidence-based clinically relevant explanations, bridging the gap between ML outputs and clinical narratives.

- Reliability(RQ5): finally, to ensure safe and reliable integration of AI into care workflows, I explored risk mitigation strategies, including human-in-the-loop mechanisms. These approaches help maintain accountability and trustworthiness of AI-generated outputs, particularly in sensitive clinical decision-making contexts.

6.2 Research Implications

Our multi-component framework tackles core hurdles in smart-home monitoring: from generating synthetic behavior data and detecting subtle ADL changes, to drafting clinician-ready notes under a human-overseen regime, all while respecting patient privacy. By integrating these modules, the system enhances both the scalability and interpretability of behavior monitoring solutions for aging populations. Key implications include:

- The development of BehavGAN demonstrates that GAN-based data augmentation can be leveraged to generate realistic and diverse ADL sequences, implying that future research can focus on reducing dataset scarcity and improving model robustness by synthetic data generation strategies.
- The effectiveness of ambient sensors in capturing fine-grained activity patterns while preserving individual privacy suggests that non–vision-based monitoring approaches deserve further exploration as a non-intrusive alternative for home-based health monitoring systems.
- By demonstrating the complementary strengths of a Transformer-based approach (which excels at analyzing longer ADL sequences but requires labeled

data) and an IRL-based approach (which operates semi-supervised), the results suggest that combining supervised and semi-supervised paradigms can address individual limitations. This highlights the potential for further research on hybrid architectures that balance data requirements and sequence-length capabilities, enabling more reliable behavior-change detection.

- The introduction of an LLM-driven clinical note generation module with HITL safeguards implies that automating the translation of raw behavioral data into clinically meaningful summaries is feasible, encouraging investigations into risk mitigation strategies. Future work should focus on developing and validating oversight mechanisms to enhance the reliability and accountability of AI-generated clinical notes in decision-making contexts.

6.3 Future Work

As we look ahead, the thesis outlines the following key directions for advancing the framework and supporting long-term impact:

- Incorporating multimodal data (e.g., wearable sensors, vital sign sensors, sleep sensors, environmental metadata) to enrich behavioral context.
- Exploring continual learning approaches to adapt to evolving behavior over time.
- Addressing the exponential growth of the MDP state space in ADL sequence modeling. This can be mitigated using state abstraction techniques such as clustering behavioral motifs or deploying hierarchical reinforcement learning to maintain tractability under limited training data.

- Enhancing the robustness and specificity of LLM-generated clinical notes through domain-specific training.

- Building Robust HITL Workflows. The human–AI handoff process can be optimized by defining thresholds for clinician intervention, reducing cognitive load while preserving safety and trust in automated interpretations.

- Investigating the generalizability of the framework by incorporating ADL logs from multi-resident homes, examining how overlapping or interacting activity patterns impact detection accuracy and model robustness.

- Assessing system usability among older adults, focusing on perceptions of technological paternalism [Voinea et al., 2024]; design participatory studies to elicit user feedback, adapt interfaces, and ensure acceptance in real-world deployments.

- Evaluating the LLM-based clinical note generation module under diverse training scenarios, such as varying LLMs, fine-tuning on domain-specific corpora, and fine-tuning strategies, to identify optimal configurations for accuracy and relevance.

- Comparing the quality of automatically generated notes against human-authored clinical summaries, quantifying differences in completeness, coherence, and clinical utility.

By directly addressing each of the guiding research questions, this thesis establishes a foundation for intelligent, interpretable, and ethical eldercare solutions. It demonstrates how advanced AI systems can augment human judgment in healthcare without compromising trust, privacy, or clinical accountability. The proposed framework offers a path forward for integrating ambient intelligence into aging-in-place strategies, ultimately improving quality of life and care delivery.

Appendix A

Background

In this chapter, the scientific background related to this research is discussed. First, the Markov Decision Process (MDP) is introduced as a basis for formulating problems in Reinforcement Learning. I also present how Reinforcement Learning solves an MDP-formulated problem. Finally, Inverse Reinforcement Learning (IRL), its benefits as well as its differences with Reinforcement Learning is introduced.

A.1 Markov Decision Process and Reinforcement Learning

A process can be considered a Markov Decision Process if the decision to be taken is only dependent on the current state of the environment (Markov property [Frydenberg, 1990]). In other words, regardless of previous states of the environment, the agent should be able to take proper action (make a decision) at any point in time.

Reinforcement Learning (RL) problems are commonly modeled using Markov Decision Processes (MDPs) [Russell, 1998]. An MDP is defined as $M = \langle S, A, T, R, \gamma \rangle$, where S represents the set of possible states, A denotes the set of available actions, T is the transition function, R is the reward function, and γ is the discount factor.

A *state* represents the situation of the environment. It is important to represent the state of the environment in a way that the Markov property is not violated. To be more specific, the state representation must contain all the necessary information for the agent to understand the current situation of the environment in order to take the appropriate action. The environment offers a set of actions available in each state, forming an action space from which the agent selects its actions. The agent interacts with the environment by performing these actions, which may cause changes in the environment's state in response.

In the context of a Markov Decision Process (MDP), the likelihood of transitioning to the next state S_{t+1} depends solely on the current state S_t and the chosen action A_t at time step t, regardless of any previous states or actions.

The transition function determines the resulting state the agent moves to. This environment may be either deterministic or stochastic. In a deterministic setting, the transition from state S_t to S_{t+1} upon taking action A_t occurs with a probability of 1. Conversely, in a stochastic environment, the transition function assigns a probability value (p) to each possible transition tuple $\langle S_t, A_t, S_{t+1} \rangle$, representing the chance of moving from state S_t to state S_{t+1} after executing action A_t.

As a part of the interaction between the agent and the environment, upon the agent's action, the environment passes a reward on to the agent using a reward function. The reward gives the agent feedback about its performance in order to reinforce the agent's behavior positively or negatively. Guiding the agent via feedback can be done by providing an immediate reward (discount factor γ of 0) or discounted reward (discount factor γ of 1). While an immediate reward only takes into account the reward associated with the current action, a discounted reward considers a trajectory of rewards that are possible throughout the journey toward the terminal state.

Therefore, a discounted reward is affected by rewards in the distant future, which makes it a suitable approach where the agent's actions have long-term consequences.

The agent's primary objective is to select actions that maximize the total discounted reward accumulated over a series of steps. The *policy* is referred to as a function that determines what action to take in order to maximize the accumulated reward given the current state of the environment. In order to find the optimal policy of the agent, a value function Q is defined to estimate the expected reward that can be obtained when following the policy, given a state-action pair (s, a). The optimal policy yields the highest possible values for every state and is found by solving the Bellman equation:

$$\underbrace{\text{New}Q(s,a)}_{\text{New Q-Value}} = (1-\alpha)Q(s,a) + \underset{\text{Learning rate}}{\alpha}\,[\underbrace{R(s,a)}_{\text{Reward}} + \underset{\text{Discount rate}}{\gamma}\,\overbrace{\max Q'(s',a')}^{\substack{\text{Maximum predicted reward, given} \\ \text{new state and all possible actions}}}] \tag{A.1.1}$$

where the value function (Q) is approximated based on the cumulative expected reward for (s,a).

If all the elements of an MDP are known, the solution can be computed before ever actually executing an action in the environment. Otherwise, the agent will need to experience the environment (i.e. trial and error) in order to understand the environment dynamics and to estimate the optimal policy. In this case, the agent starts in the first place with a random policy to choose the action given the current state. Then, it receives the reward from the environment and updates its policy so that in the next occurrences of a similar state, the agent would take a better action leading to a higher reward.

When it comes to making a decision in stochastic environments, decision-making turns into a tricky problem to be solved as model-based approaches cannot be applied due to the fact that the environment is not deterministic. In a deterministic environment, given the current state of the environment, the action that is taken by the agent solely determines the next state without any uncertainty. However, in a stochastic environment, an action in a given state can put the environment in different next states due to the stochastic nature of the environment.

A.2 Inverse Reinforcement Learning (IRL)

The goal of Inverse Reinforcement Learning (IRL) is to model an agent's preference based on observed behavior, avoiding the need to manually specify the reward function. The observed agent's interaction with its environment is commonly modeled as a Markov Decision Process (MDP), where the solution is a policy that assigns actions to states. Because the reward function of this MDP is not known, it is assumed that the agent follows the optimal policy of the MDP. IRL has drawn a lot of interest from researchers in the fields of artificial intelligence and machine learning because it satisfies the needs listed below [Arora and Doshi, 2021].

- Demonstration replaces manual reward specification

The requirement to pre-specify the reward function restricts the use of RL and optimal control to issues where a reward function can be simply stated. Given that a policy or example of intended behavior is known, IRL offers a technique to expand the applicability of RL and minimize manual task specification design. While obtaining the entire desired policy is typically impractical, we have simpler access to demonstrations of behavior, frequently in the form of recorded data.

- Improved Generalization

A reward function can be transferred to another agent and provide a concise representation of an agent's preferences. If the subject agent and the other agent have similar environments and purposes, the learned reward function can be employed exactly as is; otherwise, it continues to serve as a valuable foundation even when the agent specifications are slightly different. In fact, compared to the observed agent's policy, the reward function is naturally more transferrable, as Russell [Russell, 1998] has pointed out.

A.2.1 Definition of IRL

In order to define IRL, I focus on the most popular framework for modeling the observed agent's behavior, which is the MDP. I use the standard terminology in Inverse Reinforcement Learning (IRL), where the observed agent is called the expert and the agent trying to learn is called the learner. IRL typically assumes that the expert acts according to some underlying policy, denoted as π_E, which is often unknown. When the policy is not directly accessible, the learner observes sequences of state-action pairs from the expert, known as trajectories. Although the reward function is unknown, the learner usually assumes it follows a certain structure to facilitate learning. Common assumptions include representing the reward as a linear combination of features, modeling it as a probability distribution over reward functions, or using a neural network. With this background, we can now formally define the IRL problem [Arora and Doshi, 2021].

Let an MDP without reward, model the interaction of the expert E with the environment. Let $\mathcal{D} = \langle (s_0, a_0), (s_1, a_1), ..., (s_j, a_j) \rangle_1, ..., \langle (s_0, a_0), (s_1, a_1), ..., (s_j, a_j) \rangle_{i=2}^N, s_j \in S, a_j \in A, and i, j, N \in \mathbb{N}$ be the set of demonstrated trajectories. A trajectory in $\mathcal{D}$ is denoted as τ. We may assume that all $\tau \in \mathcal{D}$ are perfectly observed. Then, determine $\hat{R_E}$ that best explains the observed behavior in the form of demonstrated trajectories.

We may express the reward function as a linear sum of weighted features:

$$R(s,a) = w_1\phi_1(s,a) + w_2\phi_2(s,a) + ... + w_k\phi_k(s,a) = w^T\phi(s,a) \tag{A.2.1}$$

where $\phi_k : S \to R$ is a feature function and weight $w_k \in \mathbb{R}$.

A.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] consist of two neural networks: a generator and a discriminator, which compete in a minimax game. The generator creates samples starting from random noise, while the discriminator learns to differentiate between these generated samples and real data, both of which are input into a supervised learning model. The generator receives feedback from the discriminator and aims to produce samples that closely resemble real data, making it harder for the discriminator to tell them apart. Specifically, the discriminator D is trained to maximize the accuracy of labeling both real and generated samples correctly, whereas the generator $G$is trained simultaneously to confuse the discriminator. Together, D and G engage in a two-player minimax game defined by the value function $V(D,G)$:

$$\begin{aligned} \underset{G}{min} G \underset{D}{max}\, V(D,G) =& \mathbf{E}_{Y \approx p_{data}(Y)}[log D(Y)] \\ &+ \mathbf{E}_{z \approx p_z(z)}[log(1 - D(G(z)))]. \end{aligned} \tag{A.3.1}$$

where p_{data} is the real data distribution and $p_z(z)$ is input noise used to learn p_g. The value function V is defined so as to maximize the discriminator's error by minimizing the generator's error. According to the above formula, a good generator generates samples similar to real data so that $D(G(z))$ would be close to 1 and $D(Y)$ would be close to 0 leading to maximizing $log(D(Y) + log(1 - D(G(z)))$.

Although several versions of generative adversarial networks such as conditional GANs, DCGAN, and InfoGAN [Mirza and Osindero, 2014, Radford et al., 2015, Chen et al., 2016] have been presented and successfully used for generating verisimilar images, generating sequences of discrete tokens has not received much study. SeqGAN is an effort to close this gap by providing an algorithm that leverages reinforcement learning to calculate a reward based on the discriminator's judgment on complete generated sequences. The authors have also utilized a Monte Carlo search to calculate the reward for partial sequences using rollout mechanisms. They have tested the efficiency of their proposed algorithm using text and music datasets [Yu et al., 2017]. LeakGAN [Guo et al., 2018] is also an effort to address the issue of long text generation. The authors propose to allow the generator receive leaked information on the discriminator's high-level features and incorporate such signals into generation steps. We (Akbari et al.) [Akbari et al., 2022] propose BehavGAN, a model-free behavior sequence generator algorithm (Appendix C), by extending the original SeqGAN method. BehavGAN proposes a new use for GANs in the simulation of older people's behavior by learning the characteristics of a target dataset. By including n-gram-based similarity metrics in the reinforcement mechanism, BehavGAN gains an efficient reward function for GAN backpropagation.

In SeqGAN, the discriminator reward, which is backpropagated to the generator, is formulated as:

$$R_{D_\phi}^{G_\theta}(a = y_T, s = y_{1:T-1}) = D_\phi(y_{1:T}). \tag{A.3.2}$$

In this formula, $D_\phi(y_{1:T})$ is the discriminator's judgment on a complete sequence. This stands for the discriminator's estimate of the probability that the sequence is real. It is then backpropagated to the generator as the reward in reinforcement. For further details of SeqGAN please see the original paper [Yu et al., 2017].

Appendix B

Metrics

The purpose of this appendix is to introduce the metrics that we will use for evaluating the results of the proposed models for detecting bahvior changes. Accuracy is a widely-used metric for measuring how much accurate a prediction is. It's computed by the sum of true predictions divided by the total predictions (See Formula B.0.1(a)). While the model's ability to distinguish positive and negative classes can be measured by accuracy, it is not merely enough to measure the efficiency of a predictor model.

The first issue with accuracy metric is that it gives equal importance to all classes. In problems that predicting one class is of more importance than the other/s, such as anomaly detection, it is required to use other evaluation metrics such as recall and precision.

In Precision, the focus is on the positive class predictions as shown in Formula B.0.1(b). If the model predicts negative class poorly, it would not be caught by Precision. Also, if the data is imbalanced, Precision would not be sufficient for evaluation. Recall, which can be calculated from Formula B.0.1(c), takes into account the false negatives, which are super important in fraud detection, anomaly detection, etc. Finally, F1 measure is a combined metric which can be computed according to Formula

B.0.1(d).

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} && (a)\\ Precision &= \frac{TP}{TP+FP} && (b)\\ Recall &= \frac{TP}{TP+FN} && (c)\\ F1 &= \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} && (d) \end{aligned} \tag{B.0.1}$$

Another useful metric for evaluating binary classifiers is AUC that reflects the area under the ROC curve. Unlike accuracy, AUC is independent from prediction threshold. This feature makes it an excellent statistic for evaluating the model in an unbiased way.

A key deficiency of the aforementioned metrics is that they do not consider the model confidence in predicting classes. For These metrics do not reflect if a model predicts a true class with a high probability or a marginal probability. This is why the model uses Cross Entropy or Log Loss for training (See Formula B.0.2, where p is the prediction probability and y is class label, 0 or 1). Predictions that are closer to the class label receive a lower Cross Entropy loss while the accuracy is a binary true/false for a certain sample.

$$CrossEntropy = -(y\log(p) + (1-y)\log(1-p)) \tag{B.0.2}$$

Appendix C

BehavGAN Algorithm

The BehavGAN algorithm for data generation is provided here.

$$
\begin{aligned}
&R_{D_\phi}^{G_\theta}(a = y_t, s = B'_{1:t-1}) = \\
&\begin{cases}
\frac{1}{N}\sum_{n=1}^{N} D_\phi({B'_{1:t}}^n), {B'_{1:t}}^n \in MC^{G_\beta}(B'_{1:t}; N) & for\ t < T \\
D_\phi(B'_{1:t}) & for\ t = T
\end{cases}
\end{aligned}
\tag{C.0.1}
$$

$$
\begin{aligned}
&R_b^{G_\theta}(a = y_t, s = B'_{1:t-1}) = \\
&\begin{cases}
\frac{1}{N}\sum_{n=1}^{N} R_b({B'_{1:t}}^n), {B'_{1:t}}^n \in MC^{G_\beta}(B'_{1:t}; N) & for\ t < T \\
R_b(B'_{1:t}) & for\ t = T
\end{cases}
\end{aligned}
\tag{C.0.2}
$$

$$
\begin{aligned}
&R_{comb} = f(R, R_b) = \\
&\qquad\qquad\begin{cases}
max(R) - R & if\ R_b > Threshold \\
R & otherwise
\end{cases}
\end{aligned}
\tag{C.0.3}
$$

Algorithm 3: Behavior GAN (BehavGAN).)

Require: Generator Policy: G_θ; Roll-out Policy G_β; Discriminator Policy D_ϕ; Real Sequence Dataset (Positive Samples) $S = X_{1:T}$
Ensure: Synthetic Sequence Data (Negative Samples)

Initialize G_θ, D_ϕ with random weights θ, ϕ.
Pre-train G_θ using MLE on S
$\beta \leftarrow \theta$
Generate negative samples using G_θ for training D_ϕ
Pre-train D_ϕ using negative and positive(S) samples via minimizing cross entropy
repeat

  for g-steps **do**
    Generate a sequence $B'_{1:T} = (y_1, ..., y_T) \approx G_\theta$
    for t in 1 : T **do**
      Compute $R^{G_\theta}_{D_\phi}(a = y_t, s = B'_{1:t-1})$ by Eq.C.0.1
      Compute $R^{G_\theta}_{b}(a = y_t, s = B'_{1:t-1})$ by Eq.C.0.2
      Compute R_{comb} by Eq.C.0.3
    end for
    Update generator parameters via policy gradient
  end for
  for d-steps **do**
    Use Current G_θ to generate negative samples and combine with given positive samples S
    Train discriminator D_ϕ for k epochs
  end for
  $\beta \leftarrow \theta$
until BehavGAN converges

Bibliography

Fateme Akbari, Kamran Sartipi, and Norm Archer. Synthetic behavior sequence generation using generative adversarial networks. *ACM Transactions on Computing for Healthcare*, 2022.

Ane Alberdi, Alyssa Weakley, Maureen Schmitter-Edgecombe, Diane J Cook, Asier Aztiria, Adrian Basarab, and Maitane Barrenechea. Smart home-based prediction of multidomain symptoms related to alzheimer's disease. *IEEE journal of biomedical and health informatics*, 22(6):1720–1731, 2018.

Paula Alhola and Päivi Polo-Kantola. Sleep deprivation: Impact on cognitive performance. *Neuropsychiatric disease and treatment*, 2007.

Afsaneh Ardebili, Selcuk Eken, and Dilek Küçük. Context-aware rule-based anomaly detection for smart homes. In *2020 International Conference on Artificial Intelligence and Data Processing (IDAP)*, pages 1–6. IEEE, 2020.

Damla Arifoglu and Abdelhamid Bouchachia. Abnormal behaviour detection for dementia sufferers via transfer learning and recursive auto-encoders. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 529–534. IEEE, 2019a.

Damla Arifoglu and Abdelhamid Bouchachia. Detection of abnormal behaviour for dementia sufferers using convolutional neural networks. *Artificial intelligence in medicine*, 94:88–95, 2019b.

Damla Arifoglu and Abdelhamid Bouchachia. Abnormal Behaviour Detection for Dementia Sufferers via Transfer Learning and Recursive Auto-Encoders. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 529–534. IEEE, 2019c. ISBN 1-5386-9151-5.

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

Hadi Banaee, Gibson Chimamiwa, Marjan Alirezaie, and Amy Loutfi. Explaining habits and changes of activities in smart homes. In *Artificial Intelligence for Health, Personalised Medicine and Wellbeing (HELPLINE), in conjunction with ECAI 2020, Santiago de Compostela, Spain (Digital Conference), August 29-September 8, 2020*, 2020.

Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019. Publisher: Oxford University Press.

Anjanava Biswas and Wrick Talukdar. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *arXiv preprint arXiv:2405.18346*, 2024.

Nathan Brake and Thomas Schaaf. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency? *arXiv preprint arXiv:2404.06503*, 2024.

A. Burns and S. Iliffe. Alzheimer's disease. *BMJ*, 338:b158, 2009.

Fabien Cardinaux, Simon Brownsell, Mark Hawley, and David Bradley. Modelling of behavioural patterns for abnormality detection in the context of lifestyle reassurance. In *Iberoamerican Congress on Pattern Recognition*, pages 243–251. Springer, 2008.

Flávia Dias Casagrande, Jim Tørresen, and Evi Zouganeli. Sensor event prediction using recurrent neural network in smart homes for older adults. In *2018 International Conference on Intelligent Systems (IS)*, pages 662–668. IEEE, 2018.

Hapugahage Thilak Chaminda, Vitaly Klyuev, and Keitaro Naruse. A smart reminder system for complex human activities. In *2012 14th International Conference on Advanced Communication Technology (ICACT)*, pages 235–240. IEEE, 2012. ISBN 89-5519-163-4.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

Veerle Claes, Els Devriendt, Jos Tournoy, and Koen Milisen. Attitudes and perceptions of adults of 60 years and older towards in-home monitoring of the activities of daily living with contactless sensors: an explorative study. *International journal of nursing studies*, 52(1):134–148, 2015.

A. Clegg et al. Frailty in elderly people. *The Lancet*, 381(9868):752–762, 2013.

Diane J. Cook. Learning setting-generalized activity models for smart spaces. *IEEE intelligent systems*, 2010(99):1, 2010. Publisher: NIH Public Access.

Diane J Cook and Maureen Schmitter-Edgecombe. Assessing the quality of activities in a smart environment. *Methods of information in medicine*, 48(05):480–485, 2009.

Diane J Cook and Maureen Schmitter-Edgecombe. Fusing ambient and mobile sensor features into a behaviorome for predicting clinical health scores. *IEEE Access*, 9: 65033–65043, 2021.

ANDREA Cove-Smith and MK Almond. Management of urinary tract infections in the elderly. *Trends in Urology, Gynaecology & Sexual Health*, 12(4):31–34, 2007.

Aaron S Crandall and Diane J Cook. Behaviometrics for identifying smart home residents. In *Human Aspects in Ambient Intelligence*, pages 55–71. Springer, 2013.

K. Crowley. Sleep and sleep disorders in older adults. *Neuropsychology Review*, 21 (1):41–53, 2011.

Siddharth Dalal, Majd Alwan, Reza Seifrafi, Steven W Kell, and Donald E Brown. A rule-based approach to the analysis of elders' activity data: Detection of health and possible emergency conditions. In *AAAI Fall Symposium: Caring Machines*, pages 29–36, 2005.

Trisha Das, Dina Albassam, and Jimeng Sun. Synthetic patient-physician dialogue generation from clinical notes using llm. *arXiv preprint arXiv:2408.06285*, 2024.

Athanasios Dasios, Damianos Gavalas, Grammati Pantziou, and Charalampos Konstantopoulos. Wireless sensor network deployment for remote elderly care monitoring. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–4, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Vincent Mark Durand and David Barlow. Abnormal psychology, 2003.

Laura Erhan, M Ndubuaku, Mario Di Mauro, Wei Song, Min Chen, Giancarlo Fortino, Ovidiu Bagdasar, and Antonio Liotta. Smart anomaly detection in sensor systems: A multi-perspective review. *Information Fusion*, 67:64–79, 2021.

Labiba Gillani Fahad and Syed Fahad Tahir. Activity recognition and anomaly detection in smart homes. *Neurocomputing*, 423:362–372, 2021.

A. Fiske, J. L. Wetherell, and M. Gatz. Depression in older adults. *Annual Review of Clinical Psychology*, 5:363–389, 2009.

Brian J Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, pages 1–7, 2009.

Abdur Rahim Mohammad Forkan, Ibrahim Khalil, Ayman Ibaida, and Zahir Tari. Bdcam: Big data for context-aware monitoring—a personalized knowledge discovery framework for assisted healthcare. *IEEE transactions on cloud computing*, 5 (4):628–641, 2015a.

Abdur Rahim Mohammad Forkan, Ibrahim Khalil, Zahir Tari, Sebti Foufou, and Abdelaziz Bouras. A context-aware approach for long-term behavioural change detection and abnormality prediction in ambient assisted living. *Pattern Recognition*, 48(3):628–641, 2015b.

Roschelle Fritz, Katherine Wuestney, Gordana Dermody, and Diane J Cook. Nurse-in-the-loop smart home detection of health events associated with diagnosed chronic conditions: A case-event series. *International Journal of Nursing Studies Advances*, page 100081, 2022.

Morten Frydenberg. The chain graph markov property. *Scandinavian Journal of Statistics*, pages 333–353, 1990.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

WHO Guidance. Ethics and governance of artificial intelligence for health. *World Health Organization*, 2021.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. ISBN 2374-3468. Issue: 1.

André Hajek and Hans-Helmut König. Longitudinal predictors of functional impairment in older adults in europe–evidence from the survey of health, ageing and retirement in europe. *PloS one*, 11(1):e0146967, 2016.

Jiyeon Han, Jimin Park, Jinyoung Huh, Uran Oh, Jaeyoung Do, and Daehee Kim.

Ascleai: A llm-based clinical note management system for enhancing clinician productivity. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2024.

Yongkoo Han, Manhyung Han, Sungyoung Lee, A. M. Sarkar, and Young-Koo Lee. A framework for supervising lifestyle diseases using long-term activity monitoring. *Sensors*, 12(5):5363–5379, 2012. Publisher: Molecular Diversity Preservation International.

Simon Himmel and Martina Ziefle. Smart home medical technologies: users' requirements for conditional acceptance. *i-com*, 15(1):39–50, 2016.

Enamul Hoque and John A Stankovic. Poster abstract: A context-sensitive anomaly detection framework for smart home environment. In *Proceedings of the ACM/IEEE 3rd International Conference on Cyber-Physical Systems*, pages 207–208, 2012.

Aadel Howedi, Ahmad Lotfi, and Amir Pourabdollah. An entropy-based approach for anomaly detection in activities of daily living in the presence of a visitor. *Entropy*, 22(8):845, 2020.

D. J. Hunter. Osteoarthritis. *Best Practice & Research Clinical Rheumatology*, 25(6): 801–814, 2011.

Walaa N Ismail, Mohammad Mehedi Hassan, and Hessah A Alsalamah. Context-enriched regular human behavioral pattern detection from body sensors data. *IEEE Access*, 7:33834–33850, 2019.

J. Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376, 2008.

Théo Jourdan, Antoine Boutet, Amine Bahi, and Carole Frindel. Privacy-preserving iot framework for activity recognition in personal healthcare monitoring. *ACM Transactions on Computing for Healthcare*, 2(1):1–22, 2020.

HyoJe Jung, Yunha Kim, Heejung Choi, Hyeram Seo, Minkyoung Kim, JiYe Han, Gaeun Kee, Seohyun Park, Soyoung Ko, Byeolhee Kim, et al. Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. *arXiv preprint arXiv:2404.05144*, 2024.

H. C. Kales, L. N. Gitlin, and C. G. Lyketsos. Assessment and management of behavioral and psychological symptoms of dementia. *BMJ*, 350:h369, 2015.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*, 2021.

Anastasios Karakostas, Alexia Briassouli, Konstantinos Avgerinakis, Ioannis Kompatsiaris, and Magda Tsolaki. The dem@ care experiments and datasets: a technical report. *arXiv preprint arXiv:1701.01142*, 2016.

R Kavitha and Sumitra Binu. Performance evaluation of area-based segmentation technique on ambient sensor data for smart home assisted living. *Procedia Computer Science*, 165:314–321, 2019.

Susann M Keohane, Scott N Gerard, Aliza Heching, Sam S Adams, and Peri Tarr. Reflections on the effectiveness of a high density ambient sensor deployment for monitoring healthy aging. In *International Conference on Human Aspects of IT for the Aged Population*, pages 315–333. Springer, 2018.

M. S. Kirkman et al. Diabetes in older adults. *Diabetes Care*, 35(12):2650–2664, 2012.

Antonios Konios, Mario Garcia-Constantino, and Ahmad Lotfi. Personalized anomaly detection for activities of daily living using ambient sensors. In *UK Workshop on Computational Intelligence (UKCI)*, pages 1–6. IEEE, 2019.

Maria Koutli, Stathis Konstantinidis, Panagiotis D. Bamidis, and Andreas Billis. Probabilistic modeling of human activity patterns in a smart home using a hidden semi-markov model. *IEEE Journal of Biomedical and Health Informatics*, 23(6): 2492–2503, 2019.

Kundan Krishna, Deepali Jain, Sanket V Mehta, and Sunav Choudhary. An lstm based system for prediction of human activities with durations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–31, 2018.

Stein Kristiansen, Konstantinos Nikolaidis, Thomas Plagemann, Vera Goebel, Gunn Marit Traaen, Britt Øverland, Lars Aakerøy, Tove-Elizabeth Hunt, Jan Pål Loennechen, Sigurd Loe Steinshamn, et al. Machine learning for sleep apnea detection with unattended sleep monitoring at home. *ACM Transactions on Computing for Healthcare*, 2(2):1–25, 2021.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. Medsyn: Llm-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer, 2024.

Paula Lago, Claudia Jiménez-Guarín, and Claudia Roncancio. Contextualized behavior patterns for change reasoning in ambient assisted living: A formal model. *Expert Systems*, 34(2):e12163, 2017.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, and Zehan Wang. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

Yukyung Lee, Jina Kim, and Pilsung Kang. Lanobert: System log anomaly detection based on bert masked language model. *arXiv preprint arXiv:2111.09564*, 2021.

Chen Li, William K Cheung, Jiming Liu, and Joseph K Ng. Automatic extraction of behavioral patterns for elderly mobility and daily routine analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–26, 2018.

Dingwen Li, Jay Vaidya, Michael Wang, Ben Bush, Chenyang Lu, Marin Kollef, and Thomas Bailey. Feasibility study of monitoring deterioration of outpatients using multimodal data collected by wearables. *ACM Transactions on Computing for Healthcare*, 1(1):1–22, 2020.

Qiang Lin, Daqing Zhang, Dongsheng Li, Hongbo Ni, and Xingshe Zhou. Extracting intra-and inter-activity association patterns from daily routines of elders. In *International Conference on Smart Homes and Health Telematics*, pages 36–44. Springer, 2013.

Inna Lisko, Jenni Kulmala, Martin Annetorp, Tiia Ngandu, Francesca Mangialasche, and Miia Kivipelto. How can dementia and disability be prevented in older adults: where are we today and where are we going? *Journal of internal medicine*, 289(6): 807–830, 2021.

Ahmad Lotfi, Caroline Langensiepen, Sawsan M Mahmoud, and Mohammad Javad Akhlaghinia. Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour. *Journal of ambient intelligence and humanized computing*, 3(3):205–218, 2012a.

Ahmad Lotfi, Caroline Langensiepen, Sawsan M. Mahmoud, and Mohammad Javad Akhlaghinia. Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour. *Journal of ambient intelligence and humanized computing*, 3(3):205–218, 2012b. ISBN: 1868-5137 Publisher: Springer.

Weina Ma and Kamran Sartipi. Synthesizing scenario-based dataset for user behavior pattern mining. *International Journal of Computer and Information Technology*, 4(6):855–866, 2015.

Shahin Mahmoud, Ahmad Lotfi, and Claus Langensiepen. Anomaly detection in activities of daily living using fuzzy logic. In *2011 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*, pages 1–7. IEEE, 2011.

Michael Marschollek, Marcus Becker, Jürgen M Bauer, Petra Bente, Lena Dasenbrock, Katharina Elbers, Andreas Hein, Gerald Kolb, Harald Künemund, Christopher Lammel-Polchau, et al. Multimodal activity monitoring for home rehabilitation of geriatric fracture patients–feasibility and acceptance of sensor systems in the gal-natars study. *Informatics for Health and Social Care*, 39(3-4):262–271, 2014.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

M Moallem, H Hassanpour, and AA Pouyan. Anomaly detection in smart homes using deep learning. *Iranian (Iranica) Journal of Energy & Environment*, 10(2): 126–135, 2019.

J. E. Morley. Undernutrition in older adults. *Family Practice*, 29(suppl_1):i89–i93, 2012.

D. K. Moser and B. Riegel. Heart failure. *The Journal of Cardiovascular Nursing*, 23(4):345–351, 2008.

Ivan Moura, Ariel Teles, Luciano Coutinho, and Francisco Silva. Towards identifying context-enriched multimodal behavioral patterns for digital phenotyping of human behaviors. *Future Generation Computer Systems*, 131:227–239, 2022.

Mohamed Tarik Moutacalli, Abdenour Bouzouane, and Bruno Bouchard. The behavioral profiling based on times series forecasting for smart homes assistance. *Journal of Ambient Intelligence and Humanized Computing*, 6(5):647–659, 2015. Publisher: Springer.

Ghaith Mustafa, Antonios Konios, and Mario Garcia-Constantino. Temporal anomaly detection in human behaviour using iot data. In *UK Workshop on Computational Intelligence (UKCI)*, pages 1–6. IEEE, 2020.

Viswam Nathan, Sudip Paul, Temiloluwa Prioleau, Li Niu, Bobak J Mortazavi, Stephen A Cambone, Ashok Veeraraghavan, Ashutosh Sabharwal, and Roozbeh Jafari. A survey on smart homes for aging in place: Toward solutions to the specific needs of the elderly. *IEEE Signal Processing Magazine*, 35(5):111–119, 2018.

Ehsan Nazerfard. Temporal features and relations discovery of activities from sensor data. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–16, 2018.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

N. R. Nicholson. A review of social isolation: an important but underassessed condition in older adults. *The Journal of Primary Prevention*, 33(2-3):137–152, 2012.

Liqiang Nie, Luming Zhang, Yi Yang, Meng Wang, Richang Hong, and Tat-Seng Chua. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 591–600, 2015.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Hamidreza Pazhoumand-Dar, Martin Masek, and Chai Kiat Lam. Anomaly detection in smart home data: A hidden markov model approach. *Applied Soft Computing*, 33:224–233, 2015.

Antoine Piau, Benoit Lepage, Carole Bernon, Marie-Pierre Gleizes, Fati Nourhashemi, et al. Real-time detection of behavioral anomalies of older people using artificial intelligence (the 3-pegase study): protocol for a real-life prospective trial. *JMIR Research Protocols*, 8(11):e14245, 2019.

R. W. Pretorius et al. Reducing the risk of adverse drug events in older adults. *American Family Physician*, 87(5):331–336, 2013.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Parisa Rashidi and Diane J Cook. Keeping the resident in the loop: Adapting the smart home to the user. *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans*, 39(5):949–959, 2009.

Daniele Riboni, Claudio Bettini, Gabriele Civitarese, Zaffar Haider Janjua, and Rim Helaoui. Fine-grained recognition of abnormal behaviors for early detection of mild cognitive impairment. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 149–154. IEEE, 2015a.

Daniele Riboni, Claudio Bettini, Gabriele Civitarese, Zaffar Haider Janjua, and Rim Helaoui. Fine-grained recognition of abnormal behaviors for early detection of mild cognitive impairment. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 149–154. IEEE, 2015b. ISBN 1-4799-8033-1.

T. A. Rowe and M. Juthani-Mehta. Urinary tract infection in older adults. *Aging Health*, 9(5):519–528, 2013.

L. Z. Rubenstein. Falls in older people: epidemiology, risk factors and strategies for prevention. *Age and Ageing*, 35(suppl_2):ii37–ii41, 2006.

Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 101–103, 1998.

Jérémie Saives, Clément Pianon, and Gregory Faraut. Activity discovery and detection of behavioral deviations of an inhabitant from binary sensors. *IEEE Transactions on Automation Science and Engineering*, 12(4):1211–1224, 2015.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Alberto Sanfeliu and Manuel Lazo Cortés. Progress in pattern recognition, image analysis and applications. In *9th Iberoamerican Congress on Pattern Recognition, CIARP*. Springer, 2004.

J. M. Schols et al. Defining chronic organ failure in frail elderly patients: a conceptual paper. *Nursing Outlook*, 57(6):317–324, 2009.

Kyarash Shahriari and Mana Shahriari. Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 197–201. IEEE, 2017.

Cuijuan Shang, Chih-Yung Chang, Jinjun Liu, Shenghui Zhao, and Diptendu Sinha Roy. Fiid: feature-based implicit irregularity detection using unsupervised learning from iot data for homecare of elderly. *IEEE Internet of Things Journal*, 7(11): 10884–10896, 2020.

Sanjeet Singh, Shreya Gupta, Niralee Gupta, Naimish Sharma, Lokesh Srivastava, Vibhu Agarwal, and Ashutosh Modi. Generation and de-identification of indian clinical discharge summaries using llms. *arXiv preprint arXiv:2407.05887*, 2024.

Yifu Sun and Haoming Jiang. Contextual text denoising with masked language models. *arXiv preprint arXiv:1910.14080*, 2019.

Nagender Kumar Suryadevara, Subhas C. Mukhopadhyay, Ruili Wang, and R. K. Rayudu. Forecasting the behavior of an elderly using wireless sensors data in a smart home. *Engineering Applications of Artificial Intelligence*, 26(10):2641–2652, 2013a. ISBN: 0952-1976 Publisher: Elsevier.

Nagender Kumar Suryadevara, Subhas C Mukhopadhyay, Ruili Wang, and RK Rayudu. Forecasting the behavior of an elderly using wireless sensors data in a smart home. *Engineering Applications of Artificial Intelligence*, 26(10):2641–2652, 2013b.

NK Suryadevara, Anuroop Gaddam, RK Rayudu, and SC Mukhopadhyay. Wireless sensors network based safe home to care elderly people: Behaviour detection. *Sensors and Actuators A: Physical*, 186:277–283, 2012.

Jonathan Synnott, Chris Nugent, and Paul Jeffers. Simulation of smart home activity datasets. *Sensors*, 15(6):14162–14179, 2015. Publisher: Multidisciplinary Digital Publishing Institute.

Kok Weng Tay, Nur Izzati A. Bakar, Nazlia Norwawi, Kamarul Azmi, Jeffrey K. Liew, and Zurina M. Hanapi. A survey on abnormal human behaviour detection in smart environments. *IEEE Access*, 11:3843–3863, 2023.

Claire M Timon, Pamela Hussey, Hyowon Lee, Catriona Murphy, Harsh Vardan Rai, and Alan F Smeaton. Automatically detecting activities of daily living from in-home sensors as indicators of routine behaviour in an older population. *Digital health*, 9:20552076231184084, 2023.

Md Zia Uddin, Weria Khaksar, and Jim Torresen. Ambient sensors for elderly care and independent living: a survey. *Sensors*, 18(7):2027, 2018.

Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 1–9, 2008.

Gilles Virone, Majd Alwan, Siddharth Dalal, Steven W Kell, Beverely Turner, John A Stankovic, and Robin Felder. Behavioral patterns of older adults in assisted living. *IEEE transactions on information technology in biomedicine*, 12(3):387–398, 2008.

Cristina Voinea, Tenzin Wangmo, and Constantin Vică. Paternalistic ai: the case of aged care. *Humanities and Social Sciences Communications*, 11(1):1–9, 2024.

Yanan Wang, Mengmeng Liu, and Lin Zhao. Smart home anomaly detection for elderly care: A review. *IEEE Access*, 11:21901–21915, 2023.

Wiktoria Wilkowska, Julia Offermann, Susanna Spinsante, Angelica Poli, and Martina Ziefle. Analyzing technology acceptance and perception of privacy in ambient assisted living for using sensor-based technologies. *Plos one*, 17(7):e0269642, 2022.

Jingyu Xiao, Zhiyao Xu, Qingsong Zou, Qing Li, Dan Zhao, Dong Fang, Ruoyu Li, Wenxin Tang, Kang Li, Xudong Zuo, et al. Make your home safe: Time-aware unsupervised user behavior anomaly detection in smart homes via loss-guided mask. *arXiv preprint arXiv:2406.10928*, 2024.

Bibi M. Yahaya, Claus Langensiepen, and Ahmad Lotfi. Outlier detection for activity of daily living pattern using one-class classification. *Journal of Ambient Intelligence and Humanized Computing*, 10(3):1185–1197, 2019a.

Bibi M. Yahaya, Ahmad Lotfi, and Mohamad Mahmud. An ensemble method for abnormal behaviour detection in smart homes. *IEEE Access*, 7:144888–144897, 2019b.

Salisu Wada Yahaya, Ahmad Lotfi, and Mufti Mahmud. A consensus novelty detection ensemble approach for anomaly detection in activities of daily living. *Applied Soft Computing*, 83:105613, 2019c.

Salisu Wada Yahaya, Ahmad Lotfi, and Mufti Mahmud. Towards a data-driven adaptive anomaly detection system for human activity. *Pattern Recognition Letters*, 145:200–207, 2021.

Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeff Ward. A continued pretrained llm approach for automatic medical note generation. *arXiv preprint arXiv:2403.09057*, 2024.

Dorsaf Zekri, Thierry Delot, Marie Thilliez, Sylvain Lecomte, and Mikael Desertot. A framework for detecting and analyzing behavior changes of elderly people over time using learning techniques. *Sensors*, 20(24):7112, 2020.

Meriem Zerkouk and Belkacem Chikhaoui. Long short term memory based model for abnormal behavior prediction in elderly persons. In *How AI Impacts Urban Living and Public Health: 17th International Conference, ICOST 2019, New York City, NY, USA, October 14-16, 2019, Proceedings 17*, pages 36–45. Springer, 2019.

Tingzhi Zhao, Hongbo Ni, Xingshe Zhou, Lin Qiang, Daqing Zhang, and Zhiwen Yu. Detecting abnormal patterns of daily activities for the elderly living alone. In *International Conference on Health Information Science*, pages 95–108. Springer, 2014.

Yan Zhao, Baoqiang Ma, Pengbo Jiang, Debin Zeng, Xuetong Wang, and Shuyu Li. Prediction of alzheimer's disease progression with multi-information generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 25(3): 711–719, 2021. doi: 10.1109/JBHI.2020.3006925.

Chun Zhu, Weihua Sheng, and Meiqin Liu. Wearable sensor-based behavioral anomaly detection in smart assisted living systems. *IEEE Transactions on automation science and engineering*, 12(4):1225–1234, 2015.