BILSTM AND EVOLUTIONARY COMPUTATION FOR SURGICAL GESTURE RECOGNITION

ENHANCING SURGICAL GESTURE RECOGNITION USING BIDIRECTIONAL LSTM AND EVOLUTIONARY COMPUTATION: A MACHINE LEARNING APPROACH TO IMPROVING ROBOTIC-ASSISTED SURGERY

By YIFEI ZHANG, BASc

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

Lay Abstract

Advancements in artificial intelligence (AI) are transforming medicine, particularly in robotic surgery. This thesis focuses on improving how robots recognize and classify surgeons' movements during operations. Using a special AI model called a bidirectional Long Short-Term Memory (BiLSTM) network, which looks at data both forwards and backwards, the study aims to better understand and predict surgical gestures.

By applying this model to a dataset of surgical tasks, specifically suturing, and optimizing its settings with advanced techniques, the research shows significant improvements in accuracy and efficiency over traditional methods. The enhanced model is not only more accurate but also smaller and faster.

These improvements can help train surgeons more effectively and advance robotic assistance in surgeries, leading to safer and more precise operations, ultimately benefiting both surgeons and patients.

Abstract

The integration of artificial intelligence (AI) and machine learning in the medical field has led to significant advancements in surgical robotics, particularly in enhancing the precision and efficiency of surgical procedures. This thesis investigates the application of a single-layer bidirectional Long Short-Term Memory (BiLSTM) model to the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) dataset, aiming to improve the recognition and classification of surgical gestures. The BiLSTM model, with its capability to process data in both forward and backward directions, offers a comprehensive analysis of temporal sequences, capturing intricate patterns within surgical motion data. This research explores the potential of BiLSTM models to outperform traditional unidirectional models in the context of robotic surgery.

In addition to the core model development, this study employs evolutionary computation techniques for hyperparameter tuning, systematically searching for optimal configurations to enhance model performance. The evaluation metrics include training and validation loss, accuracy, confusion matrices, prediction time, and model size. The results demonstrate that the BiLSTM model with evolutionary hyperparameter tuning achieves superior performance in recognizing surgical gestures compared to standard LSTM models.

The findings of this thesis contribute to the broader field of surgical robotics and

human-AI partnership by providing a robust method for accurate gesture recognition, which is crucial for assessing and training surgeons and advancing automated and assistive technologies in surgical procedures. The improved model performance underscores the importance of sophisticated hyperparameter optimization in developing high-performing deep learning models for complex sequential data analysis.

Dedicated to my family and mentors for their unwavering support and guidance.

Dedicated to myself for persevering through all the challenges.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Thomas Doyle, for his invaluable guidance, support, and mentorship throughout my research journey. I am also sincerely thankful to my defense committee, Dr. Michael Noseworthy and Dr. Anita Acai, for their insightful feedback and encouragement, which greatly improved the quality of my work. I would also like to thank McMaster University for providing the opportunity and funding to pursue my graduate studies.

I would also want to extend my appreciation to my fellow lab members for their camaraderie and contributions, which made this experience more rewarding. I am especially grateful to my parents for their unwavering support and belief in me, and to my girlfriend, Judy, for her constant love and encouragement. A special thanks to my friends for always lifting my spirits when I encountered obstacles.

Lastly, a heartfelt thanks to xxsk for providing entertaining streams that kept me motivated through the long grind, and to Uber Eats drivers working at midnight providing services.

Table of Contents

List of Figures

List of Tables

Notation, Definitions, and Abbreviations

Notation

b_t	Bias term at time t
C_t	Cell state at time t
h_t	Hidden state at time t
i_t	Input gate at time t
o_t	Output gate at time t
σ	Sigmoid activation function
tanh	Hyperbolic tangent activation function
W_f, W_i, W_C, W_o	Weight matrices for forget, input, cell, and output gates
$\tilde{C}_t$	Candidate cell state at time t

f_t	Forget gate at time t
h_{t-1}	Hidden state at previous time step
x_t	Input vector at time t

Definitions

Challenge With respect to video games, a challenge is a set of goals presented to the player that they are tasked with completing; challenges can test a variety of player skills, including accuracy, logical reasoning, and creative problem solving

Bidirectional LSTM (BiLSTM)

A type of recurrent neural network that processes data in both forward and backward directions to capture dependencies in both temporal directions

Deep Learning

A subset of machine learning involving neural networks with many layers, capable of learning complex patterns in large datasets

Evolutionary Computation

A family of optimization algorithms inspired by natural selection that iteratively improve candidate solutions based on a fitness function

Hyperparameter Tuning

The process of selecting the best configuration of hyperparameters to optimize the performance of a machine learning model

Recurrent Neural Network (RNN)

A type of neural network designed to recognize patterns in sequences of data by using loops within the network to maintain information

Surgical Gesture Recognition

The process of identifying and classifying the movements of surgeons during operations using machine learning techniques

Abbreviations

AI	Artificial Intelligence
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
EC	Evolutionary Computation
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
JIGSAWS	JHU-ISI Gesture and Skill Assessment Working Set
HRI	Human-Robot Interaction

Declaration of Academic Achievement

This thesis represents the culmination of my work towards my Master's degree, and I declare that the work presented herein is my own, except where acknowledged otherwise. I would like to acknowledge the contributions of the following individuals:

- Dr. Thomas Doyle, my supervisor, for his guidance, advice, and expertise in the development of this research. His invaluable feedback shaped the methodology and analysis presented in this thesis.
- Dr. Michael Noseworthy and Dr. Anita Acai, my defense committee members, for their insightful comments and suggestions, which greatly improved the quality of this work.
- My fellow lab members for their collaboration and support throughout the research process.

I performed all data collection, analysis, and interpretation, and I am solely responsible for the writing of this thesis. The assistance provided by my supervisor and committee members was advisory in nature, and the conclusions and findings are my own.

Chapter 1

Introduction

The advancement of robot systems and machine learning has led to the employment of robots in various industries to restructure labor. According to a report by the International Federation of Robotics (IFR), the adoption of human-robot collaboration is on the rise, with an 11% increase in collaborative robot (cobot) installations compared to 2019 [56]. Cobots are designed to work alongside humans, enhancing productivity and safety in various tasks. In the medical field, these advancements have been particularly transformative, leading to the development of sophisticated robotic systems designed to assist surgeons in performing complex procedures. One notable example is the Da Vinci Surgical System, which enhances a surgeon's capabilities by providing greater precision, flexibility, and control during operations [23].

Intention-based systems are a new class of user-centered assistance systems that recognize the user's intention and act upon it to take on both active and passive roles in the interaction [55]. This results in a more natural interaction between the user and the robot as the system synchronizes with the interacting entity during the process. In surgical settings, these systems can significantly enhance the surgeon's ability to

perform delicate procedures by anticipating and responding to their needs in real-time. However, the implementation of these systems necessitates the integration of multiple sensors and sophisticated algorithms to accurately interpret the surgeon's intentions from limited information such as voice commands, gestures, and eye movements.

The rapid advancement of machine learning and artificial intelligence has profoundly impacted various research fields, leading to significant progress in data analysis, pattern recognition, and predictive modeling. Among these advancements, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have emerged as powerful tools for handling sequential data. Their success in capturing temporal dependencies makes them particularly well-suited for time-series analysis, natural language processing, and complex tasks like surgical gesture recognition in robotic surgery.

A key aspect of these advancements is the growing partnership between humans and AI. In the medical field, this partnership aims to leverage the strengths of both humans and machines to achieve outcomes that neither could accomplish alone. For instance, AI can process vast amounts of data with high precision, while human surgeons bring expertise, intuition, and decision-making skills that are crucial in complex medical scenarios. This synergy enhances the overall efficiency and effectiveness of medical procedures, leading to improved patient outcomes and advancing the field of surgery.

However, traditional LSTM models are limited in their ability to fully capture the nuances of sequential data that contain important features at both the beginning and end of the sequence. This limitation stems from their unidirectional processing of data, which can result in the loss of crucial contextual information from future events

within a sequence. To address this issue, Bidirectional LSTM (BiLSTM) models have been developed [48]. These models process data in both forward and backward directions, offering a more comprehensive analysis of temporal sequences and improving the ability to understand and predict complex patterns.

In the context of robotic surgery, the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) has become a pivotal dataset for developing and testing machine learning models aimed at recognizing and classifying surgical gestures. This dataset includes kinematic and video data from surgeons performing fundamental surgical tasks such as suturing, knot-tying, and needle-passing. Accurate recognition of these gestures is crucial for assessing and training surgeons and advancing automated and assistive technologies in surgical procedures.

Despite these advancements, there are still significant challenges in effectively analyzing and modeling complex, sequential data. The inherent complexity of surgical gestures, combined with the variability in individual surgeon techniques, necessitates innovative approaches to improve accuracy and reliability. Advanced models capable of capturing subtle temporal patterns and dependencies within the data are essential for accurate surgical gesture recognition.

The introduction of intention-based systems in human-robot interaction (HRI) represents a significant advancement in the field of robotics and artificial intelligence. These systems are designed to recognize and predict human intentions, facilitating a more natural and efficient interaction between humans and robots. Intention-based systems utilize various sensors and algorithms to understand user intentions and act upon them, enhancing the collaborative capabilities of robots in diverse scenarios. However, trust in these systems remains a critical factor that has not been extensively

studied. Understanding the role of trust in the implementation of AI within healthcare and other domains is essential for optimizing human-robot collaboration and ensuring the effective use of these technologies.

This thesis introduces a novel application of a single-layer bidirectional LSTM model to the JIGSAWS dataset to enhance the recognition and classification of surgical gestures. The hypothesis is that the BiLSTM architecture, with its dual-direction processing capability, will outperform traditional unidirectional models in capturing the intricate patterns of surgical motion sequences. Additionally, the study explores optimizing the model through hyperparameter tuning, employing evolutionary computation techniques to systematically search for optimal configurations. This approach not only seeks to improve model performance but also contributes to the broader discussion on efficient hyperparameter optimization strategies in deep learning.

By leveraging the strengths of BiLSTM models and advanced hyperparameter optimization, this research aims to provide valuable insights into the applicability of these models for complex sequential data analysis, specifically in surgical gesture recognition. The methodology and findings presented in this thesis have the potential to enhance surgical training programs, improve surgical outcomes, and advance the field of robotic surgery.

note

Chapter 2

Literature Review

2.1 Overview

The current robots involved in human-robot interaction scenarios range from taking and serving orders in restaurants to assembling sophisticated parts in factories. However, most interactions with robots require people to approach the robot and initiate the interaction, reflecting their belief in the robot's ability to complete a successful social encounter [2]. In contrast, humans are both initiators and responders in social interactions, relying on rich sensory input and experience to anticipate the other's actions. This is where intention-based systems come into play in HRI scenarios.

In the field of Human-Robot Interaction, the definition of a "robot" or "intention-based system" has been a subject of debate, particularly with the integration of AI technologies blurring the traditional boundaries. A robot is conventionally considered as an autonomous or semi-autonomous system, capable of perceiving its environment, processing information, and performing actions to achieve specific goals [23]. However,

the advent of AI has extended this definition to include systems that were not traditionally considered robots. For instance, autonomous vehicles, which have the ability to perceive their environment and operate without human intervention, could be classified as robots within the broader understanding [23]. Similarly, exoskeletons, which enable or enhance human capabilities through intelligent design and control, can also be included under this umbrella [23]. Moreover, a prime example of this expansive definition is the Da Vinci Surgical System, a robot-assisted platform designed to facilitate complex surgery using a minimally invasive approach in healthcare domain [23]. Although it doesn't operate autonomously, the system enhances the surgeon's capabilities, enabling more precise movements and greater control, with the need for more communication and teamwork during robotic assisted surgery (RAS) [4]. This further illustrates how AI-driven systems, even those requiring substantial human operation, can be classified as robots within the context of their intention-based operation. This expanded definition recognizes that as technologies advance, the line distinguishing robots from other systems becomes increasingly ambiguous.

Previous literature has proposed various methods of determining the user's intention in HRI scenarios by utilizing different sensor data and algorithms. Some of the designs are already employed in working environments such as rehabilitation [35], life-support [36], assembly [41], driving [15], etc. Like machine learning algorithms, intention-based systems are task-oriented in implementation, leading to variability in the choice of sensor combination and algorithm, given the tasks spanning different areas of the industry.

Aside from designing the system, trust is also fundamental in HRI, especially in healthcare. It significantly affects the adoption and optimal use of AI technologies

by influencing users' confidence in the system's capabilities, reliability, and safety [9]. Trust involves not only belief in the AI's technical competencies but also understanding its operations, transparency, and risk management [9]. Hence, cultivating trust in HRI is paramount to the successful integration of AI in healthcare and vital for ensuring beneficial interactions between users and AI systems. However, despite the nature of intention-based systems where factors like trust, which influences human interaction with these systems, are pivotal, the exploration of this aspect remains scarce in the existing literature.

2.2 Scoping Review Methodology

This section outlines the methodology used to conduct the scoping review for this thesis. The aim of the scoping review is to provide a comprehensive overview of the existing research on intention-based systems in human-robot interaction (HRI), particularly in the context of surgical gesture recognition. The review follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure a systematic and unbiased approach. The methodology includes the formulation of research questions, literature search strategy, selection criteria, and data extraction process.

2.2.1 Objective

As AI and robotics continue to advance, the use of intention-based systems in working environments is becoming increasingly common [56]. However, there is currently no literature providing a comprehensive overview of the design characteristics of

intention-based systems. Therefore, a scoping review is needed to gain a better understanding of the field before impactful designs can be made. The aim of this study is to provide a basic understanding of the current state of intention-based systems and assist in future implementations. Specifically, the objectives are to:

1. Provide an overview of the sensors and algorithms used in intention-based systems in the collected experimental research and describe the HRI scenarios in which each system is used.

2. Explore the possible effects of human trust when working with intention-based systems.

3. Identify gaps in literature for future research and establish a foundation for subsequent design.

By compiling these different aspects, this study can help researchers implement more comprehensive and user-friendly intention-based systems in HRI scenarios. The review process will be conducted according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [44] to minimize bias and provide a broad understanding of the current state of the field.

This systematic review adheres to the PRISMA guideline throughout the entire process. This guideline outlines a systematic approach to collecting and synthesizing data while having a well-formulated research question. By following this structure, the review aims to provide a comprehensive and unbiased overview of the current status and design characteristics of intention-based systems.

2.2.2 Strategy

The literature search for this review was conducted between October and November 2022 using three databases: Ovid MEDLINE, Ovid Embase, and IEEE Xplore. The search query used for each database is shown in Table 2.1. While the syntax of the search string may vary depending on the database, the terms were chosen to capture a similar set of research literature. Both title and abstract, as well as subject headings, were searched and reviewed based on the availability of search methods for each database. The search was limited to English language publications but not restricted by publication date. The aim of the search was to identify as many relevant studies as possible to ensure the comprehensiveness of the review.

Table 2.1: Database and respective search strings.

Database	**Search string**
Ovid MEDLINE, Ovid Embase	1. Exp artificial intelligence/ 2. (Machine intelligence OR Computer intelligence OR Cognitive computing OR Robot* OR Expert system* OR Intelligent system* OR Autonomous agent* OR Artificial* intelligen* OR Machine learning OR Deep learning OR Neural network OR Computational intelligence).ti,ab,sh 3. (Intent* OR intent* predict* OR move* predict* OR act* predict* OR Prediction algorithm).ti,ab,sh 4. 1 AND 2 AND 3
IEEE	("Document Title":"machine intelligence" OR "Document Title":"computer intelligence" OR "Document Title":"cognitive computing" OR "Document Title":"robot" OR "Document Title":"expert system" OR "Document Title":"intelligent system" OR "Document Title":"autonomous agent" OR "Document Title":"artificial intelligence" OR "Document Title":"machine learning" OR "Document Title":"deep learning" OR "Document Title":"neural network" OR "Document Title":"computational intelligence") AND ("Document Title":"intent*" OR "Document Title":"intent* predict*" OR "Document Title":"move* predict*" OR "Document Title":"act* predict*" OR "Document Title":"prediction algorithm")

2.2.3 Participants

Studies with human aged above or equal to 18 years old are included. The demographics of participants drawn from the surveyed literatures present a diverse range. In totality, they comprise of more than 200 individuals spanning various age groups, sex, and physical abilities. The age of participants largely ranged from young adults in their early twenties to individuals in their late sixties, with a few studies focusing on specific age ranges from 21 to 35 years old. In terms of sex, a majority of the subjects were male, though a substantial number of females were also included. The handedness of participants was also considered in some studies in hand gesture recognition, and a few in lower and upper-limb intention recognition included subjects with specific physical conditions, such as amputations. Overall, the participant pool was diverse, providing a broad perspective on the interaction between humans and intention-based systems across different demographic groups.

2.2.4 Intervention

The inclusion criteria for this review were studies that proposed the design of sensors or algorithms for intention-based systems, as well as evaluations of such systems. For the purposes of this review, any system that utilized human intention to provide feedback or judgments was considered an intention-based system, as long as it was implemented in an HRI scenario that directly involved human interaction. Additionally, studies that used Wizard of Oz testing were included in order to provide a comparison, where users believed they were interacting with intention-based systems, but the system was actually controlled by a human [26].

2.2.5 Inclusion and Exclusion Criteria

Exclusion criteria was chosen to prune out less relevant literature to this study. Three exclusion criteria (E) and three inclusion criteria (I) are identified as following before screening and assessing the search results from databases:

- E1: Studies that are review articles, dissertations, and conference abstract.
- E2: Studies focusing on financial, cryptocurrency, and brain-computer interface.
- E3: Studies that do not meet the requirement stated in participants, that is, with humans under 18 years of age, or do not meet any inclusion criteria.
- I1: Studies focusing on intention, intention-based system, human-robot interaction.
- I2: Studies that include implementation of sensors or algorithms.
- I3: Studies evaluating effect of intention-based system on team dynamics or human perceptions and attitudes when working with one.

Inclusion criterion I1 required that any study searched had to focus on the topics of intention, intention-based system, or human-robot interaction to be considered for inclusion. Additionally, studies needed to meet at least one of the other inclusion criteria: I2, which included studies that discussed the implementation of sensors or algorithms in intention-based systems, and I3, which included studies that evaluated the effect of intention-based systems on team dynamics or human perceptions and attitudes. However, studies that met any of the exclusion criteria were excluded from the review. E1 excluded all review articles, dissertations, and conference abstracts to ensure that the review was based only on primary sources. E2 excluded studies that

focused on irrelevant topics such as financial or cryptocurrency markets, or derivation of human intention in non-HRI scenarios such as brain-computer interfaces. While these topics related to AI methodologies are often incorporated and can be included in the initial search, given that they do not specifically focus on HRI, they will be omitted from the scope of this review. Finally, E3 excluded any study that included a participant of under-aged or did not meet any of the inclusion criteria.

2.2.6 Information Extraction

After applying the inclusion and exclusion criteria, full-text articles were reviewed, and data were extracted from the selected references. To ensure the collected data is relevant to the purpose of the review, several crucial aspects were considered, such as the design's purpose, the sensors and algorithms utilized, the data collection process, the data size for training, testing, and validation, the performance evaluation, demographic information of the participants involved(age, sex, height, weight, right-handed or left-handed), and the method used for identifying intentions. Once the data extraction was completed, the identified design characteristics were compiled into a single file and categorized accordingly. A matrix was created to show the statistical outcomes from the references. The designs were then grouped together based on the recognized intention and were analyzed in comparison.

2.3 Conducting the Review

2.3.1 Analysis

The entire process is shown in Figure 2.1. With the search string in Table 2.1 applied to the following databases: Ovid MEDLINE, Ovid Embase, and IEEE Xplore. The search identified a total of 1296, 428, and 223 articles, respectively. After removing duplicates and retracted articles, the number of literatures reduced to 1293 before screening. All remaining studies were screened by title, and 1015 articles (78.50% of deduplicated search) were ineligible for not discussing social impacts, sensor, or algorithm implementation, or is application in financial, cryptocurrency, or brain-computer interface; 19 articles (1.47% of deduplicated search) that are under different author names (problem of naming convention). In the remaining 259 articles, abstract screening excluded 196 articles (75.68%) due to not discuss intention-based system, no sensor or algorithm mentioned, not in HRI context, or not original research. Finally, the remaining 63 articles were screened as full-text, leaving 35 articles (51.61%) as included in this review. The exclusion is more specific compared to the two before: discussed about psychology of intention rather than intention recognition of systems, trials but limited social impact mention, design for recognizing vehicle intention instead of human, etc.

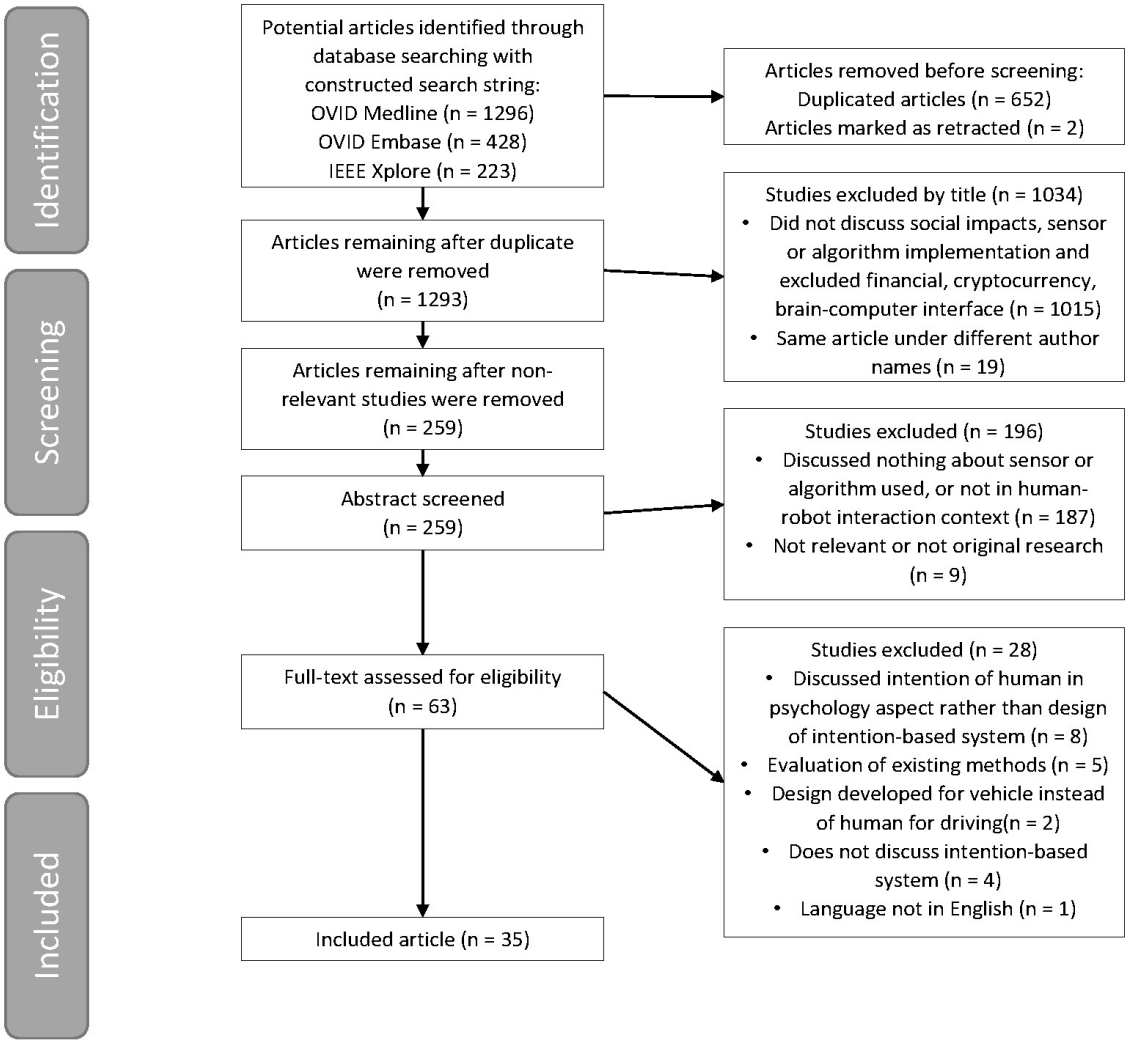

Figure 2.1: PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart.

Table 2.2 presents a summary of the demographic information of the studies included in the review that proposed new designs for sensors and algorithms in intention-based systems. Since the identification of human intention varies depending on the task, the designs proposed in the studies were categorized based on the type of intention, mentioned in Figure 2.2. It shows the ontology created to better visualize the structure of the literature review. The intention types are separated into whole body and localized body parts, where interaction, motion, and activity are the former, and hand gesture, upper/lower limb movement, facial gesture are the latter. Each would be introduced in the later sections. The use of sensor clusters also varies depending on the specific intention, especially in upper-limb and lower-limb detection, where the placement of electrodes on muscles may differ. Thus, it is difficult to rank the sensors and algorithms used in the designs as being best or worst. Instead, they have their own advantages and disadvantages. Table 2.3 provides the complete reference to the intention types, and Table 2.4 for reference to algorithms, which both will be discussed in detail in the later sections. Figure 2.3 depicts the distribution of included studies from 2017 to 2022, and indicates that interest in intention-based systems has peaked in 2020, although it remains consistent throughout the years.

Table 2.2: Summary of study characteristics of sensor and algorithm design literature (n = 35).

Study characteristics	**Value, n (%)**	
Year		
2017	4 (11)	
2018	4 (11)	
2019	6 (17)	
2020	11 (31)	
2021	6 (17)	
2022	4 (11)	
Involved participants		
1–4	6 (17)	
5–9	3 (9)	
10–15	9 (26)	
16–20	2 (6)	
21–100	2 (6)	
Unspecified	13 (37)	
Sensor		
RGB camera	10 (29)	
Inertial Measurement Unit (IMU)	6 (17)	
Surface electromyography (sEMG)	6 (17)	
Depth camera	4 (11)	
Force sensor	4 (11)	
Myo armband	3 (9)	
Custom	5 (14)	
Other (mentioned once in article)	11 (31)	
Algorithm		
CNN-based	15 (43)	
LDA	3 (9)	
CNN + ConvLSTM	2 (6)	
NN	2 (6)	
Other (mentioned once in article)	13 (37)	
Data type used		
Image	11	
sEMG	10	
IMU	3	
Force	2	
Other (once in article)	9	

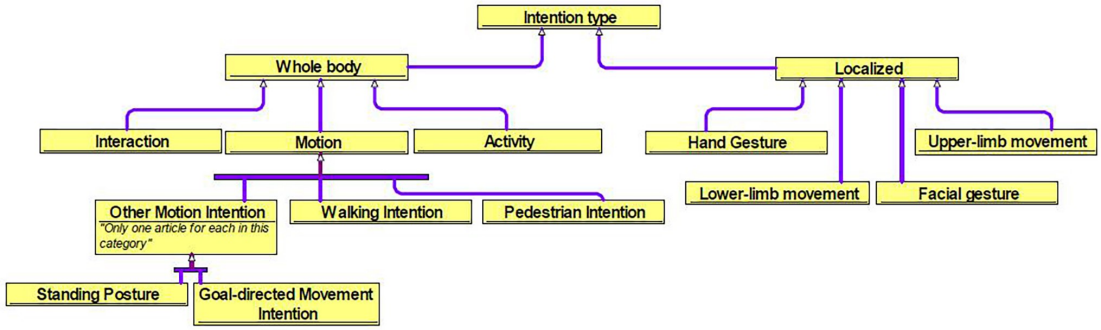

Figure 2.2: Ontology of intention-based system.

Table 2.3: Literature to intention type.

Intention type	**Study**
Motion	Ding and Zheng (2020), Follmann et al. (2018), Foerster et al. (2020), Golumbic et al. (2020), Gao et al. (2021), Thakur et al. (2020), Chiu et al. (2020)
Hand gesture	Chao et al. (2019), Chen et al. (2020), Dely et al. (2020), Gao et al. (2021), Liang et al. (2020), Nguyen et al. (2020), Novak et al. (2020), Popa et al. (2020)
Facial gesture	Cha et al. (2019)
Upper-limb movement	Benatti et al. (2019), Chen et al. (2020), Kim et al. (2020), Lin et al. (2020), Wang et al. (2020), Yousefi et al. (2020)
Activity	Singh et al. (2020), Takizawa et al. (2020)
Category	Abedini et al. (2020), Alavizadeh et al. (2020), Gao et al. (2021), Nishizawa et al. (2022)

Table 2.4: Literature to algorithm.

Algorithm	**Study**
CNN-based	15 Fang et al. (2017), Owoyemi and Hashimoto (2017), Su et al. (2019), Chen et al. (2020a), Chen et al. (2020b), Janušonis et al. (2020), Kumar and Michmizos (2020), Li et al. (2020), Mohammadi Amin et al. (2020), Chin et al. (2021), Velash et al. (2021), Wang (2021), Wen and Wang (2021), Ding and Zheng (2022), Poulos et al. (2022)
LDA	3 Lamini et al. (2018), Huang et al. (2020), Kopke et al. (2020)
CNN + ConvLSTM	2 Cha et al. (2019), Zhang et al. (2022)
NN	2 Moon et al. (2019), Coker et al. (2021)
Other (mentioned once in article)	13 Kılıç and Doğan (2017), Liu et al. (2017), Liu et al. (2018), Masalin et al. (2018), Ren et al. (2018), Lin et al. (2019), Cole-Albarran (2019), Young et al. (2019), Gardner et al. (2020), Goldhammer et al. (2020), Liu et al. (2020), Fedot et al. (2021), Tsitso et al. (2022)

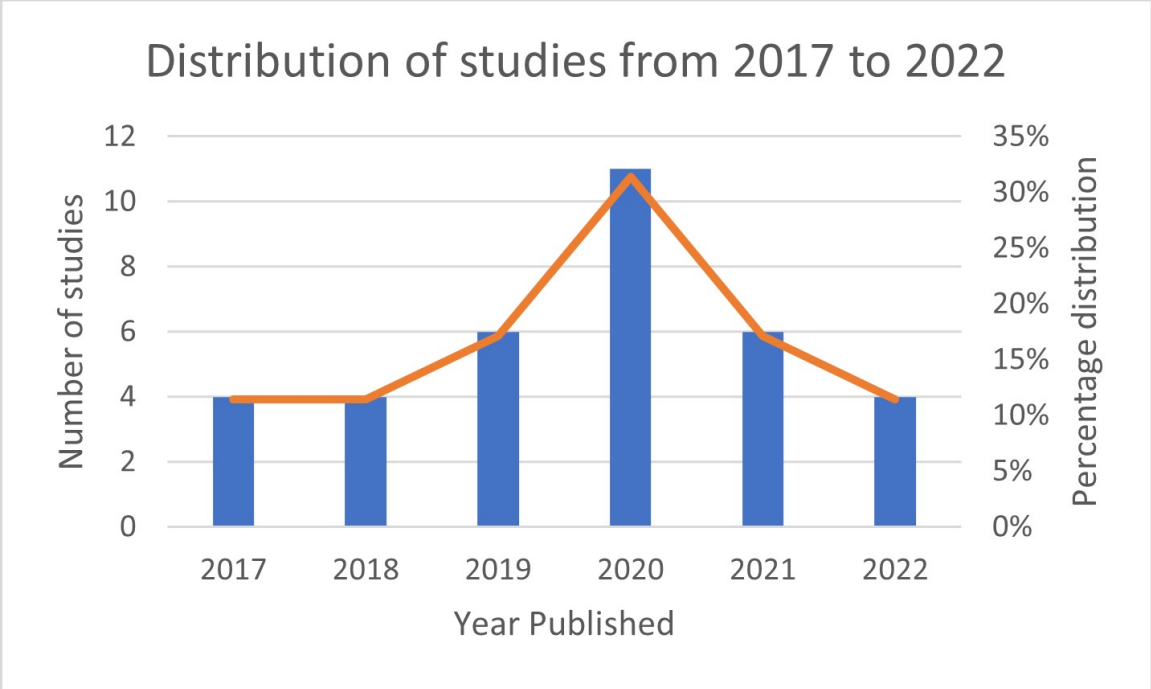

Figure 2.3: Distribution of studies from 2017 to 2022.

2.4 Reporting the Review

2.4.1 Whole Body Intention

Motion

The intention of human body movement, including standing posture, gait pattern, and walking, is covered in the section on motion intention. Instead of focusing on specific body parts such as the upper-limb or lower-limb, the studies discussed in this section prioritize whole body movements that are more general. The determination of motion intention can ensure human safety in HRI scenarios by avoiding collisions and increasing efficiency. It differs from activity, as activity focuses on different types

of action (e.g., push-up vs. sit-up vs. walking with one algorithm) whereas motion determines the occurrence of one type of action (e.g., only walking). There are a total of 7 articles that falls in the category, a summary of citations can be found in Table 2.3.

Pedestrian Intention One aspect of motion intention is predicting the movement of pedestrians in the context of automated vehicles. Goldhammer et al. [19] proposed a method called "PolyMLP" which uses artificial neural networks to predict the future movement of pedestrians and cyclists. The method employs a multilayer perceptron (MLP) with sigmoid activation functions and polynomial approximation of time series to recognize the current motion state and future trajectory of vulnerable road users (VRUs). The network is trained using offline learning approach, meaning the model would have no further learning input after training. Since it only requires information regarding VRU's past position in any coordinate system to make prediction, the sensor choice is widely flexible, and no additional information such as map data is needed. MLP was chosen due to its ability to handle multi-dimensional data input and output and learn complex patterns through several hidden layers. However, MLP requires a large amount of training data and can be prone to overfitting. In this study, the model was trained with video data of pedestrians and cyclists, and the resilient backpropagation (RPROP) algorithm was used to optimize the number and sizes of hidden layers and improve generalization [47]. The trained model classified four motion intentions, including moving, waiting, starting, and stopping, organized into connected states. The method does not require additional information such as map data and is widely flexible in sensor choice.

Fang and López [15] proposed a method to detect a pedestrian's intention to cross

the road or turn in front of the vehicle using stereo camera images as input. The system employs a two-branch multi-stage convolutional neural network (CNN) for recognition of human posture. The CNN is trained on the Microsoft COCO 2016 keypoints challenge dataset for skeleton fitting [34], and then a support vector machine (SVM) and random forest (RF) are compared as binary classifiers to determine the motion intention. Both classifiers output a normalized score, with SVM using Platt scaling on radial basis function kernel scores and RF using probability values. The system achieves an identification of crossing intention in under 750 milliseconds when transitioning from other actions such as standing still and bending. The study identifies a problem when encountering pedestrians at a far distance where the skeleton fitting may confuse left and right side body parts. This issue could potentially be minimized with a larger or more specific dataset including these cases.

Walking Intention The ability to identify the walking direction of humans is critical in several human-robot interaction scenarios, such as walking support, object manipulation, and exoskeleton control. Liu and Yang [36] proposed a design that can detect the walking direction intention of a human when using a walking support robot (WSR). The design employs a smartphone as a 3-axis accelerometer, along with force sensors embedded in the armrest of the WSR to detect pressure exerted by the human. The accelerometer is placed on the chest of the subject. To classify the intention, SVM is used, which has the advantage of being robust to noise and having global optimization. The training and validation dataset is collected from four participants when using the WSR to walk in eight directions, including forward, back, left, right, left front, left back, right front, and right back. The data is split into 80% for training and 20% for testing. The combined sensors can achieve an accuracy

of 89.4% at a data collection window width of 0.1s, and an accuracy of 95.9% at a window width of 0.5s.

LANINI et al. [30] proposed a model for human-robot collaboration when carrying heavy objects together. The model identifies the motion state of humans and enables the robot to perform synchronous movement. The study used 3D force sensors (Optoforce) and a motion capture system (Optitrack) with 15 markers to collect training data, and only used the force sensor during testing and imaginary work environment. The data collection involved 16 individuals with fair distribution, where one subject always acted as a leader in motion, and the others were followers. The followers were blindfolded and wore earmuffs to prevent visual and acoustic feedback, and the leader was equipped with Bluetooth earphones from which an audible beat was played to minimize disturbances in the data. The feature extraction was performed using single variable classification (SVC) and multivariable classification (MVC) models. SVC was used to investigate the effect of a single threshold approach on features such as force and position, and MVC was not limited in the number of features used. For classification of the four types of identified intentions (Stationary State SS, Walking Forward State WFS, and Walking Backward State WBS), linear discriminant analysis (LDA) classifier was used as it has interpretability on which are the most discriminative features and is fast for training and testing. As a result of supervised learning, SVC performs well on SS, achieving a 96% accuracy, but less satisfactory on WFS with only 79%, while MVC has a 92.3% accuracy for WFS. When the model is implemented on the COMAN robot, it performs well on the starting and stopping of synchronized motion but poorly on acceleration and deceleration, possibly due to misclassification of deceleration as stopping with slow walking speed (0.25m/s).

Another usage of walking intention detection is in dynamic gait generation for exoskeleton. Ren and Liu [46] designed a on-line dynamic gait generation model to plan real-time gait trajectories in continuous motion process according to user intentions. The exoskeleton used in the study is lightweight lower-limb exoskeleton robot (LLEX), with inertial measurement unit (IMU) in the backpack and angle sensors in the joints. Since the users all have unique stride length, the study adopted a strategy to utilize real-time spatial position planning, then use inverse kinematics to calculate the joint angle trajectory. The walking process is divided into four distinct patterns, start, normal gait, transition, and end, each with different constraint conditions. A two-state state machine is used to distinguish the two-leg support phase and other phases, which the movement intention recognition would focus on. With multi-sensor fusion of the data of IMU and angle sensors and rules developed, the four patterns can be correctly identified, with a 5% difference between generated gait and natural gait collected at the same time of generating. Comparing with existing method proposed by Kagawa et al. [25], it shows higher accuracy, naturalness, and continuity, among varies stride length tested.

Other Motion Intention In addition to the previously mentioned motion intentions, there are other categories that require identification.

Rehabilitation Kumar and Michmizos [29] proposed a design to assess motor learning by identifying the intent of initializing a goal-directed movement and the reaction time (RT) of the movement, which can be used in the rehabilitation of sensorimotor impairment. A deep CNN consisting of five layers was trained using 128-channel EEG signals to predict movement intention, and four layers were used for RT

classification. Data collection was performed in two separate tasks: the first is active mode, where the subject performs the motion, and the second is passive mode, where motion is performed by the robot with the subject's arm affixed to the robotic end effector. The training and testing dataset ratio is 4:1, and the mean accuracy achieved for movement intent and RT classification were $87.34\% \pm 2.83\%$ and $84.68\% \pm 3.68\%$, respectively. In the future, the proposed model could be used to target specific treatment and provide assistance according to the percentage of voluntary movement, and RT could be used as an indicator of functional motor recovery.

Standing Posture Li et al. [31] proposed a design for the recognition of standing posture using a pressure-sensing floor. The floor design includes a pressure buffer layer, a pressure sensor array, and a supporting plate, along with a data collection unit that gathers foot-pressure distribution over the sensor matrix. The foot-pressure distribution is then converted into a grayscale image for further usage. The proposed multi-classifier fusion algorithm includes a CNN similar to lenet-5, a SVM classifier, and a KNN classifier. The latter two were selected after comparing the training results within a group of classifiers that included SVM, KNN, RF, decision tree (DT), Naïve Bayes (NB), and backpropagation (BP) neural network. The trained network can classify between nine standing postures with an average accuracy of 99.96% across testing data. However, this study is limited to static standing postures, and future implementations could focus on identifying dynamically moving subject's posture.

Activity

This literature review focuses on activity recognition, which involves classifying whole body movements into distinct categories. The studies analyzed in this section utilize

image sensors, which are a common method as they convey more comprehensive information. [33]. Moreover, the sensors do not need to be placed directly on humans, which allows activities to proceed without interference. Detecting activities is critical when it comes to collaborating on multiple tasks, as the robot can then identify the specific task the human partner is performing and provide the corresponding assistance. There are a total of 3 articles that falls in the category, a summary of citations can be found in Table 2.3.

Jaouedi et al. [24] proposed a novel approach for human activity recognition using a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) with a Kalman filter. The CNN model used in this study is a combination of Inception V3 and MobileNet, while the RNN model is used for activity classification. The approach was applied to videos captured using an RGB-D camera and depth camera, and the spatio-temporal features of the human skeleton were extracted for feature presentation. The CAD-60 dataset was used for training and testing the model. The dataset consists of RGB-D video sequences of humans performing activities, recorded using the Microsoft Kinect sensor. The study achieved an accuracy of 95.50% for activity recognition, demonstrating the effectiveness of the proposed approach.

Poulose et al. [45] proposed a novel approach for Human Activity Recognition (HAR) systems that use a smartphone camera to capture human images and subsequently perform activity recognition. The proposed approach, referred to as the Human Image Threshing (HIT) machine-based HAR system, uses Mask R-CNN for human body detection and ResNet for classification. The HIT machine-based HAR system relies on images captured from a smartphone camera for activity recognition,

which has the potential to significantly lower the cost and complexity of HAR systems. The accuracy of the proposed system was evaluated using a dataset of 9 activities, including sitting, standing, walking, dancing, sit-up, running, jumping, push-up, and lying. The model accuracy was reported as 98.53%, with a model loss of 0.20. The precision, recall, and F1 scores were also reported as 98.56%, 98.53%, and 98.54%, respectively. The HIT machine-based HAR system achieved high accuracy in activity recognition, indicating its potential to serve as a cost-effective and efficient solution for HAR systems.

Li et al. [32] proposes a novel gaze-based intention inference framework for robots. The framework consists of three main components: head pose estimation, eye center localization, and eye model and gaze tracking. By analyzing the gaze data, the system predicts the user's intention, allowing the robot to provide appropriate assistance or interaction. Existing frameworks mainly focus on establishing the relationship between gaze points and objects, but lack the ability to predict the user's intentions. The proposed framework aims to address this limitation by enabling the robot to understand the user's intentions and provide more personalized assistance. The input to the system is gaze data captured by a camera and a fixed monitor of scene image observed by the robot.

Interaction

This literature review underscores the significance of recognizing interaction intention to ensure safety in human-robot interaction (HRI) scenarios. There are instances when humans have no intention of interacting with robots, and it is vital for the robot to identify these moments and halt the collaboration to avoid any potential

risks. This category differs from the others, focuses on multimodal approaches, similar to a human-human interaction where multiple sensors (eyes, ears, hands, etc.) are utilized to express intent. There are a total of 1 articles that falls in the category, a summary of citations can be found in Table 2.3.

Amin et al. [41] proposed a method to enhance safety by combining visual and tactile perception in human-robot interaction. To achieve this, the study employs a camera system consisting of two Kinect V2 cameras, with RGB and depth cameras. The study utilizes a 3D CNN for human action recognition and 1D CNN for contact recognition. The input for the system includes RGB and depth images captured by the camera system. The dataset for the study consists of 33,050 images divided into five classes of human action recognition and 1,114 samples divided into five classes of contact recognition. The study achieved an accuracy of 99.72% for human action recognition and 93% for contact recognition.

2.4.2 Localized Body Intentions

Hand Gesture

In this literature review, the topic of hand gestures is explored, encompassing both hands and wrist movements. Various studies discussed in this section focus on recognizing different hand gestures and their intended actions. The recognition process involves the use of several sensors and algorithm combinations, such as sEMG, MMG, RGB, and depth sensors. The primary objective of recognizing hand gestures is to enable robots to interpret human actions accurately or follow commands during collaboration, thereby enhancing the efficiency of human-robot interaction (HRI) scenarios. There are a total of 11 articles that falls in the category, a summary of citations can

be found in Table 2.3.

Chen et al. [6] proposed a design that utilizes a compact deep neural network called EMGNet for gesture recognition using sEMG data collected by Myo armband. The network has four convolutional layers and a max pooling layer, without a full connection layer as final output. EMGNet has reduced the parameters to 34,311, which is significantly lower than other models, such as CNN_LSTM, LCNN, and ConvNet. Moreover, the accuracy of EMGNet is also higher, with 98.81% on the Myo dataset and 69.62% on the NinaPro DB5 dataset. However, the NinaPro DB5 dataset suffers from low accuracy due to a relatively small amount of data with a large number of gesture categories and similar gestures representing different categories. The Myo dataset has 19 subjects performing 7 gestures, with 2280 samples for each gesture by each person, while the NinaPro DB5 dataset has 10 subjects performing 12 gestures, with 1140 samples for each gesture by each person.

In another study, Chiu et al. [8] proposed a design for recognizing human intention to open automatic doors by detecting and interpreting hand gestures. The proposed system aims to address privacy concerns and reduce the spread of infection during pandemics by enabling non-contact intention recognition. The authors utilized both thermal and camera sensors to collect data, but only the thermal data was used for actual recognition. The data consists of 6,000 images of RGB and thermal data that were masked into 64x48 pixels for "open" and "close" classes. The masked images were then fed into a Mask R-CNN, implemented with the Detectron2 library, to extract human masking. A U-Net structure was subsequently employed to identify the intention of the detected human.

Cote-Allard et al. [11] presented a novel 3-D printed armband called the 3DC

armband for sEMG hand gesture recognition. This armband features a custom SoC that can record 10 sEMG channels in parallel, a 9-axis IMU, a wireless transceiver, a MCU for interfacing the components, and a power management unit (PMU) for low-power consumption. The system is powered by a 100-mAh LiPo battery and employs a Molex connector for connecting with the armband and programming the MCU. For comparative experiment, the performance of the 3DC armband was compared to the widely used Myo armband as an sEMG measurement sensor. The Myo and 3DC armbands were worn simultaneously on the dominant arm of participants, and a total of 8 cycles of 11 hand gestures were collected for testing and training. The ConvNet architecture was then used to classify each gesture. Results showed an accuracy of 89.47% for the 3DC armband and 86.41% for the Myo armband.

Ding et al. [12] proposes a dual-channel VGG-16 CNN for gesture recognition using CCD RGB-IR and depth-grayscale images. The authors collected 30,000 CCD RGB-IR and 30,000 depth-grayscale images using a Kinect depth camera, with 10 actions to recognize in total. They fused the images using three different wavelet fusion techniques (max-min, min-max, and mean-mean), resulting in 30,000 of each fused image. The dataset was split into 15,000 for training and 15,000 for testing. The results showed that the fusion of the min-max type had the highest accuracy of 83.88%, while CCD RGB-IR only had an accuracy of 75.33% and depth-grayscale only had an accuracy of 72.94%. The mean-mean type fusion had an accuracy of 80.95%, which was also relatively high. Overall, the proposed method achieved high accuracy in gesture recognition by combining CCD RGB-IR and depth-grayscale images through wavelet fusion.

In Feleke et al. [16], a recurrent fuzzy neural network (RFNN) is proposed to map

sEMG signals to 3D hand positions without considering joint movements. The aim is to predict human motor intention for robotic applications. The study analyzed the effects of slow and fast hand movements on the accuracy of the RFNN model. Two complex tasks were performed, one involving picking up a bottle from the table and pouring it into a cup, and the other a manipulation task with multiple obstacles, resembling intelligent manufacturing. Each task was performed at both slow and fast speeds, with 9 trials for each scenario, making a total of 36 trials. The accuracy for slow tasks are 83.02%, 81.29% and for fast tasks 85.29%, 82.38%, both in the order of task 1 and task 2. The results showed that RFNN could predict hand positions with high accuracy, regardless of the speed of motion. This approach could be useful for developing human-robot interaction systems. In Young et al. [58], a simplified pipeline system for hand gesture recognition is proposed for prosthetic hand users. The system utilizes a Myo armband as the sEMG sensor and a random forest (RF) algorithm for classification. The system was tested on a dataset consisting of five hand gestures: wave in, wave out, spread fingers, fist, and pinch, from the Myo dataset. The results show that the proposed system achieved an accuracy of 94.80%, indicating high performance.

In this study by Gardner et al. [18], a low-cost multi-modal sensor suite is proposed for shared autonomy grasping, which includes a custom mechanomyography (MMG) sensor, an IMU, and a camera. The system is used to estimate muscle activation, perform object recognition, and enhance intention prediction based on grasping trajectory. The proposed KNN grasp classifier achieved high accuracy for bottle (100%), box (88.88%), and lid (82.46%) grasping tasks. The study aims to overcome limitations of commercially available systems, which often employ indirect mode-switching

or limited sequential control strategies. The proposed system allows for simultaneous activation of multiple degrees of freedom (DoF) during grasping. Three different grasp patterns were tested for a box and two different grasp patterns were tested for a bottle and a lid. Each object was tested at 18 different locations, with a 3-second time window provided for the user to reach and grasp the object and 2.5 seconds to return to a resting position. The results demonstrate the feasibility and potential for shared autonomy grasping using the proposed multi-modal sensor suite.

Tsitos et al. [51] used a single RGB-D camera to observe human behavior and predict intentions in real-time for robotic actions using logistic regression (LR) algorithm. The aim was to evaluate the feasibility of the proposed approach. The results showed 100% accuracy for both touching and distant objects when compared to human performance in all scenarios. The input was an image, and the dataset consisted of two subjects who performed grasping movements towards two identical objects in two different scenarios. Each participant completed 50 movements towards each object (left or right) for each scenario, making a total of 400 movements. An 85% to 15% split was used for training and testing. Further studies are needed to evaluate the proposed approach with a larger sample size and a wider range of scenarios to determine its generalizability and practicality.

From Chen et al. [7], a CNN-based algorithm was proposed for stiffness estimation and intention detection using sEMG data collected from the Myo armband. The algorithm consisted of six 2D convolutional layers (Conv2D), a 2D max-pooling layer (MaxPool2D), and three 2D Global Average Pooling layers (GAP2D). The output of the last GAP2D layer was concatenated with the output of the previous layer, and the process was repeated three times. The accuracy of the algorithm was 96% for

each type of wrist configurations in several trials.

The ability to identify hand movement intention is prominent for robots on collaborated assembly line in HRI scenario. Zhang et al. [59] proposed a method to predict human hand motion during an assembly task to improve collaboration flow and efficiency. The design utilizes an RGB camera mounted over the robot, facing downwards at the working area on the table. The study proposed a state-enhanced ConvLSTM network that combines the flexibility and effectiveness of regular ConvLSTM with improved accuracy [37]. The experiment involved six sub-tasks, each requiring the installation of one part of a seat. Using an extended Kalman filter (EKF) to track and a separate CNN to recognize the human intended part, together with the ConvLSTM to predict intention, the robot arm can assist the human in assembly with only image data. After training, the recognition accuracy was higher than 99%. Comparing this method with using speech recognition to detect and deliver intended parts, the prediction method saved 36.43 seconds in completing all sub-tasks. This approach reduces idle time during the process and improves the efficiency and quality of collaboration since longer idle time reflects worse collaboration between human and robot.

Owoyemi and Hashimoto [43] developed an approach to identify intention by utilizing collections of point clouds. The study used a 3D sensor mapped into 3D occupancy grids and input the processed data into a 3D CNN to recognize arm and hand motion. The evaluation of the model was to recognize subject's pick and place action from four boxes. Using 119,102 datasets for training and 14,802 separate datasets for offline testing, the model achieved 100% accuracy in identifying the intention of

the subject. When compared to LSTM and 1D convolution model variants, the proposed model showed better accuracy with significantly fewer parameters. However, the positioning of the camera used to generate the point cloud can affect the accuracy of the final result. Additionally, generalization can be a problem since only one subject was used in the training and testing datasets, which requires further data collection with different participants and more complex motions to enhance the model's generalizability.

Upper-limb Movement

This literature review delves into the topic of upper-limb movement, which encompasses the entire arm and shoulder, excluding hand gestures and wrist configurations. Various sensors, including sEMG and muscle shape change (MSC), as well as torque and limb position detection by exoskeletons, are utilized to detect upper-limb movements. The primary objective of recognizing upper-limb movement intentions is to enable robots to anticipate the trajectory of human arms and provide assistance in movement, thus reducing the workload for humans in specific muscle areas with effective designs . There are a total of 5 articles that falls in the category, a summary of citations can be found in Table 2.3. In Liu et al. [35], the focus is on upper-limb rehabilitation using exoskeleton robots, and a study proposing a sensorless control scheme with human intention estimation is discussed. The study aims to address the control problem of upper-limb rehabilitation by utilizing a self-built exoskeleton robot called NTUH-II, which detects shoulder horizontal abduction/adduction (HABD), shoulder flexion/extension (SF), and elbow flexion/extension (EF) joints. The proposed

control scheme employs a deep neural network (DNN) for human intention estimation, with the input being the upper limb torque. The accuracy of the scheme was evaluated using root mean square error (RMSE) and normalized RMSE (NRMSE) metrics. The dataset used for training and testing consisted of 28,000 data points and was split into 85% to 15% for training and testing, respectively.

Lu et al. [38] focuses on a study that proposes a novel controller for a Franka Emika robot. The purpose of the study is to enhance the efficiency of human-robot interaction by improving trajectory prediction accuracy. The controller combines variable admittance control and assistant control, and utilizes fuzzy Q-learning and LSTM algorithms for optimization. The input for the controller is human limb dynamics, and the dataset consists of 30 trajectories sampled at 1000 Hz. The trajectory prediction accuracy after model training was less than 1mm with the actual trajectory. The fuzzy Q-learning algorithm optimizes the damping value of the admittance controller by minimizing the reward function. The LSTM algorithm is utilized to predict the trajectory of the robot based on the human limb dynamics input.

The design and control of an active wrist orthosis that is mobile, powerful, and lightweight is proposed in Kilic et al. [27] as a means to avoid the occurrence and/or for the treatment of repetitive strain injuries. The study utilizes two sEMG sensors at the extensor carpi radialis (ECR) and flexor capi radialis (FCR) muscles, and a force sensor. The control system is based on a fuzzy logic controller. The study aims to reduce the workload of the FCR muscle while maintaining the accuracy of the orthosis. The study recorded the deviation from the intended trajectory for wrist movement and found that it increased from 1.794 degrees to 2.934 degrees on average, but reduced the workload by half for the FCR muscle. The dataset for the study

consisted of a healthy person performing an isometric test to record maximum torque and sEMG at the forearm. The orthosis generated three levels of torque according to EMG signals detected at the FCR and ECR muscles.

Kopke et al. [28] focuses on investigating the effectiveness of pattern recognition of sensor data to identify user intent for various combinations of 1- and 2-degree-of-freedom shoulder tasks. The sensors used in this study include load cells and sEMG electrodes. The dataset consists of different max joint torque lifting or depressing conditions determined by isometric testing. The conditions include 0%, 25%, and 50% of maximum joint torque, with a minimum of three and a maximum of ten trials of each condition completed. Two sets of LDA classifiers were developed for each dataset type, including sEMG, raw load cell data, and a combined dataset, with one set using 0% and ±25% lifting condition data and the other using 0% and ±50%. The accuracy of the combined set was found to have a 9.7% error rate.

Huang et al. [22] presents the development of a novel sensor for the acquisition of muscle shape change (MSC) signals in order to decode multiple classes of limb movement intents. The sensor is custom made using nanogold and is both flexible and stretchable. The study utilized a linear discriminant analysis (LDA) classifier to classify seven classes of targeted upper-limb movements, including hand close, hand open, wrist pronation, wrist supination, wrist extension, wrist flexion, and rest state. The dataset was collected with a video prompt for each movement, followed by a rest session, with each prompt lasting 5 seconds. The accuracy achieved was 96.06% ±1.84%, demonstrating the potential of using MSC signals for multi-class limb movement intent recognition.

Lower-limb Movement

This literature review explores the topic of lower-limb movement, which pertains to leg motions. As walking intention is considered a part of the overall body movement, this section places greater emphasis on detecting signals from the lower body to determine related activities and partial movement intentions of the thigh, knee, and other such areas. The primary objective of recognizing lower-limb movement intentions is to enable robots to predict and assist humans in various movements, including walking, standing, and stair ascending/descending, among others. There are a total of 7 articles that falls in the category, a summary of citations can be found in Table 2.3.

In Moon et al. [42], the development of a single leg knee joint assistance robot with motion intention detection using a sliding variable resistor to measure length between the knee center of rotation and the ankle (LBKA) was investigated. The aim of the study was to enhance the control of exoskeletons by incorporating the detection of motion intention. The algorithm used in the study was a neural network with 15 hidden layers and one input/output layer. The dataset used in the study consisted of three motions: stairs ascending, stairs descending, and walking. 40 training datasets were used for each motion, and 200 training datasets were used for exception state training. The performance of the algorithm was evaluated based on the ROC curve. The results showed that the algorithm had good performance, indicating that it is a promising approach for motion intention detection in exoskeletons.

Su et al. [49] proposed a novel method for training an intent recognition system that provides natural transitions between level walk, stair ascent/descent, and ramp ascent/descent. The study utilizes three IMUs (thigh, shank, and ankle of the healthy leg) and a CNN algorithm for motion intent recognition. The input to the system is

lower limb IMU data, and the dataset consists of 13 classes of motion intent. The able-bodied individuals performed ten trials each, with at least five steps, while the amputees performed ten locomotion modes, including level ground, stairs, and ramp, and any transition between them. The dataset comprises 1300 samples from able-bodied individuals and 130 from amputees. The accuracy of the system is 94.15% for able-bodied individuals and 89.23% for amputees.

In Wen et al. [54], a lower-limb motion intention recognition algorithm is proposed that utilizes multimodal long-term and short-term spatiotemporal feature fusion for accurate recognition. The input used for the algorithm is sEMG data, and the proposed algorithm consists of a 3D CNN for extracting short term spatiotemporal features in segments, Le Net and shape context to extract features of the target motion trajectory, and an LSTM network for time-series modeling of the extracted features. The purpose of the study is to develop a robust motion intention recognition system that can accurately interpret human motion in real-time scenarios. The accuracy achieved by the proposed algorithm is 90%, indicating that the fusion of long-term and short-term spatiotemporal features has significantly improved the recognition performance. However, the study does not provide any information regarding the dataset used for testing the proposed algorithm.

The study of Massalin et al. [39] proposes a user-independent intent recognition framework using depth sensing for five activities: standing, walking, running, stair ascent, and stair descent. The objective of the study is to develop and validate a framework that can accurately recognize user intention without relying on user-specific data. The sensor framework consists of a depth camera on the shank and an action camera for class labeling, with support vector machine (SVM) as the algorithm

for classification. The dataset includes 5 activities with 20 trials per subject, resulting in a total of 402403 depth images. The study concludes that the proposed framework can accurately recognize user intention in real-time, which could have potential applications in various fields such as sports training, gait analysis, and rehabilitation. The framework's user-independent nature makes it particularly useful in scenarios where user-specific data cannot be obtained, such as in public spaces or medical facilities. The accuracy of the proposed framework is reported to be 94.5%, which is achieved using 8 subjects' data for training and 4 subjects for testing.

Wang et al. [53] proposes the use of a convolutional neural network (CNN) model to reconstruct the motion pattern of a lower limb prosthesis. The input to the CNN model is the data collected from the IMU sensor attached to the prosthesis. The dataset used in this study includes four different motion patterns: heel strike, support, swing, and tippy toes touchdown. The CNN model used in this study includes seven layers, which are used to extract the features from the input data and classify the motion pattern. The accuracy of the system is measured using a recognition rate, which is found to be 98.2%. The use of a single sensor and the high accuracy of the system make it a practical and convenient solution for motion pattern recognition in lower limb prostheses.

Coker et al. [10] presents a method for predicting knee flexion angle using surface electromyography (sEMG) signals from thigh muscles and knee joint angle data. The purpose of this research is to develop a framework for predicting human intent for control purposes in exoskeleton technology. Twelve sEMG electrodes and a 10-camera Vicon motion capture system were used to collect data from ten subjects during walking trials. A nonlinear input-output time series neural network trained using

Bayesian regularization was used to predict the knee's flexion angle at 50, 100, 150, and 200 ms into the future. The neural network consisted of a single hidden layer of ten nodes with a feedback delay set to two. The accuracy of the predictions was evaluated using RMSE, which was found to be 0.68 for 50 ms, 2.04 for 100 ms, 3.38 for 150 ms, and 4.61 for 200 ms. The results showed good accuracy in predicting the knee joint angle up to 100 ms in advance, which is promising for real-time control of exoskeletons. However, the accuracy decreased for longer prediction horizons, which may be due to the complexity of the underlying muscle activation patterns. The dataset included ten subjects with no history of chronic pain in the spine or lower extremities, which suggests that the results may not be generalizable to individuals with injuries or pathologies.

Viekash et al. [52] presents a new approach for controlling and actuating a continuous passive motion (CPM) machine using a deep learning-based control strategy that integrates CNNs. The sensor inputs include sEMG and thigh IMU data, which are used to train three 1D-CNN models. Each 1D CNN algorithm is employed to analyze the sensor data, and 40 trials are conducted for each motion (forward, backward, and rest) during the training phase. The training and testing datasets are split at an 80% to 20% ratio. The accuracy of the proposed approach is reported to be 97.40%, indicating good performance.

Facial Gesture

This literature review highlights the importance of facial gestures in controlling augmented reality/virtual reality (AR/VR) systems. These gestures can potentially be utilized to collaborate with robots and enhance the efficiency of such collaborations.

In Cha et al. [5], an infrared (IR) camera and laser diode were used to capture skin deformation data as input for a spatial-temporal autoencoder (STAE) that recognizes facial gestures for hands-free user interaction (UI) with an augmented reality (AR) headset. The use of skin deformation as input for gesture recognition is a novel approach to hands-free UI for AR headsets. The STAE consists of two 3D convolutional neural networks (CNN), two convolutional long-short-term memory (ConvLSTM) networks, and one 3D CNN. The goal was to achieve high accuracy in recognizing user intentions based on facial gestures. The results showed an accuracy of 95.4% on 10 subjects during data collection. Future work could investigate the use of this approach in real-world AR applications, as well as explore the potential for combining facial gesture recognition with other types of input, such as voice or gaze, to further enhance hands-free UI for AR.

2.4.3 Discussion

As robotics become more integrated into our working and living environments, ensuring the safety and efficiency of human-robot interaction has become increasingly important. Intention-based systems have emerged as a promising approach to achieve this, as they allow robots to anticipate and respond to human movements and intentions. This literature review provides an overview of the current methods used in implementing intention-based systems, with a specific focus on the sensors and algorithms used in the process. Various studies have proposed designs for different task environments to react to different determined intentions. However, due to the current limitations of sensors and algorithms, it is premature to assume that one combination of sensors and algorithms is the best choice for all tasks. Therefore, further research is

needed to determine the most effective sensor and algorithm combinations for specific human-robot interaction scenarios.

When analyzing the data extracted from the literature and presented in Table 2, it becomes clear that the most popular sensor used in designs is the RGB camera [24, 45, 32, 8, 12, 18, 51, 41, 59, 15]. This is likely due to the widespread use of image recognition applications in recent years, as well as the relative affordability and accessibility of RGB cameras in work environments. RGB cameras have a distinct advantage in intention recognition due to their ability to capture detailed color information. This allows AI systems to accurately perceive and understand human actions in real-world scenarios, enhancing the robot's ability to infer human intent. By capturing rich visual data, RGB cameras enable machine learning models to interpret nuanced human behaviors and gestures, improving the robot's ability to anticipate human actions and interact more naturally and efficiently. However, they also have notable disadvantages. RGB cameras may struggle with recognizing intentions in low light conditions or when the subject is at a distance. Additionally, they can be affected by occlusion, where objects in the foreground block those in the background. Lastly, there are significant privacy concerns associated with using RGB cameras for intention recognition, as they can capture identifiable and sensitive visual data.

The second most commonly used sensors are IMUs and sEMG sensors. IMUs are often found in exoskeleton designs and commercialized armbands, such as the Myo armband [11, 18, 49, 53, 52, 46], while sEMG sensors are mainly used to measure upper-limb movements with armbands [6, 7, 11, 58] and lower-limb movements with sEMG electrodes [16, 10, 54, 52, 27, 28]. IMUs, which measure body acceleration and angular rate, offer the advantage of being unobtrusive, portable, and relatively

easy to use, making them ideal for real-time, dynamic motion tracking. However, their accuracy may be affected by sensor drift over time, and they may not capture subtle movements or muscular activities that do not result in noticeable motion. On the other hand, sEMG sensors, which record muscle activation, offer high temporal resolution and can detect subtle muscle contractions that might not result in visible motion, potentially improving the detection of intended movements. They can provide detailed information about the degree of muscle activation, which can be useful in assessing user intent in tasks that require fine motor control. However, sEMG signals can be sensitive to variations in sensor placement, skin condition, and muscle fatigue, which can affect their reliability. Also, the setup of sEMG sensors can be more obtrusive and uncomfortable for the user, which might limit their use in certain scenarios.

In contrast, force sensors and depth cameras are less popular. Force sensors have limitations regarding the area and tasks that require contact, which may account for their relatively low usage in designs. Additionally, depth cameras are often used in conjunction with RGB cameras rather than being employed as a standalone sensor in designs. They may also struggle with distant subjects. Moreover, they can have difficulties with transparent or reflective surfaces, and their depth accuracy decreases as the distance from the camera increases.

When compared to other algorithms, those based on CNN have been used most frequently in intention recognition research. This is likely due to the popularity of RGB cameras, which capture visual input that can be processed by CNNs to identify patterns or features indicative of specific motions or actions. They excel in feature extraction from images, which makes them ideal for interpreting complex patterns and

details in RGB images. The result can be a more accurate, real-time prediction of intentions based on visual cues, gestures, and behaviors captured by the RGB camera and analyzed by the CNN. Furthermore, CNNs are well-suited to handling complex and high-dimensional datasets, which are often generated when multiple sensors are used simultaneously in intention recognition.

The majority of studies are focused on motion intention-based system. There appears to be an even spread in more specific directions such as walking intention, pedestrian intention, and hand motion. The most common sensor and algorithm used to determine motion intents are cameras and CNN. Since motion intention is about whole body movement, sEMG and IMU are not as descriptive as camera sensor with the same effort, which builds up to a lot of CNN usage.

Several studies also focus on recognizing hand gestures, which are often used to issue commands to robots or signal collaboration intent. In these studies, sEMG and RGB sensors are used in relatively equal numbers. This is because sEMG sensors, when used on armbands, can accurately predict hand motion, while cameras can capture detailed visual information about hand gestures. Additionally, there has been an increasing usage of sEMG in upper-limb and lower-limb intention determination, making it an ideal choice when only partial intention needs to be determined. This occurs when only a small group of related muscles are used to complete an intended action. While CNN-based algorithms remain the majority in intention recognition research, various other algorithms and classifiers are also used, including RFNN, RF, and KNN. This is because some designs employ sEMG sensors, which can classify hand gestures without requiring complex data input. As a result, CNNs may not always be necessary for these specific applications, leading researchers to explore

alternative algorithms and classifiers.

While motion-based systems have been widely explored in intention recognition research, there are other domains that have received less attention and present opportunities for further study. For example, interaction and facial gestures are relatively unexplored areas that could benefit from more research.

Interaction can be studied in both social robot and industrial robot contexts, although the literature review focused primarily on the latter, with the expanded definition described in the introduction. One included study [41] proposed a design for ensuring safety during interactions with robots using visual and tactile perception, which initiated research on combining tactile cues with intention-based systems. In addition, as collaboration between humans and robots is most efficient when communication is bidirectional, it is also important to explore methods for recognizing the intentions of robots, as this will enable more effective collaboration.

Facial gesture can be explored further for integration with other intention recognition in working environment. For example, a robot could be programmed to recognize specific facial expressions, such as frustration or confusion, and use this information to adjust its actions accordingly. This could be used as an additional factor for robot reaction, improving the robot's ability to support human operators in various tasks. Additionally, facial gesture recognition could be used in assistive robots, allowing users to control the robot's behavior using facial expressions and gestures, leading to more intuitive communication between the user and the robot.

The impact of intention-based systems on trust in HRI scenario has not been extensively studied. None of the articles related to sensor and algorithms included in this review have considered the effect of trust on participants. While there is a

lack of specific research on the effects of these systems on trust, general research on robots and AI in healthcare domain can provide some insight. The studies conducted by Choudhury et al. [9], Esmaeilzadeh et al. [14], Asan et al. [1] underscore the multifaceted importance of trust in the implementation of AI within healthcare scenarios, as it emerges as a pivotal factor influencing the acceptance and effective use of these technologies, as well as Torrent-Sellens et al. [50], which highlights the factors affecting trust in RAS.

Choudhury et al.'s study [9] offers a targeted perspective by examining clinicians' trust in AI systems and how this influences their willingness to adopt such technologies. Notably, trust appears to serve as a mediator between perceived risk and expectancy in the decision to use the AI tool. This study underscores the importance of striking a balance between trust and over-reliance, suggesting that an informed and rational level of trust leads to an optimal utilization of AI, whereas blind trust can lead to overdependence and potential misuse.

Esmaeilzadeh et al. [14] broaden the perspective by focusing on patients' perceptions and how they interact with trust. Their study identifies an array of factors - from privacy concerns to communication barriers - that could influence patients' trust in AI and subsequently their intention to use AI in their healthcare. The study particularly emphasizes the importance of physician involvement in healthcare delivery involving AI tools, suggesting a co-existence model where AI augments rather than replaces human care providers.

Another study by Asan et al. [1] discussed a similar aspect. The study reflected that trust varies between patients and clinicians. With the rise of patient-centered care, understanding the role of AI in patient-clinician decision-making is essential.

Concerning medical responsibility, clinicians could face accountability if AI recommendations deviate from standard care, leading to negative outcomes. Thus, it's crucial to find a balance of trust between human judgment and AI recommendations, considering the evolving nature of AI and individual human factors.

The study by Torrent-Sellens et al. [50] examines factors affecting trust in robot-assisted surgery in Europe. Trust initially increases with more experience with robots but declines as this experience grows, suggesting a nuanced relationship. Sociodemographic factors play a pivotal role; men, those aged 40-54, and higher-educated individuals show pronounced trust based on their experience. Access to detailed, accurate information about procedures significantly impacts trust. The study calls for public policy should address the fluctuating trust by funding research on regulatory, ethical, and legal aspects and emphasizing clinical efficacy, as current design and model lacks attention on the importance of trust during RAS.

While the studies mentioned above shed some light on trust of human with robots and AI in healthcare, they do not specifically address the potential impact of intention-based systems in human-robot interaction scenarios. Intention-based systems are designed to be more "intelligent" and responsive to human intention and behavior, which may lead to different characteristics and perceptions of the robot by the human team members. As robots become more integrated into various aspects of human society, it is crucial to examine the effect of intention-based systems on trust and cooperation in different settings, such as in industrial or medical contexts. Further research on intention-based systems can provide insight into how to design and implement such systems to gain adequate trust and cooperation between humans and robots in various contexts.

2.5 Literature Review Summary

This literature review follows the PRISMA guidelines and examines the sensors and algorithms used in intention-based systems, as well as their potential impact on trust and team dynamics. The studies included in this review propose various designs for intention-based systems based on the given task environment, with the choice of sensors and algorithms being dependent on the task at hand. RGB cameras and CNN-based algorithms are the most commonly used sensors and algorithms, respectively. In contrast, sEMG measurements in electrodes and armbands are more commonly used for determining partial body intention, such as for upper-limb and lower-limb.

Despite the advancements in intention-based systems, there are still several areas where further research is needed. For instance, interaction intention can be further explored to improve bidirectional communication and increase the efficiency of collaboration. Facial gesture recognition could be integrated with other intention recognition methods to create a more intuitive interaction environment. Additionally, the effect of intention-based systems on trust and team dynamics in HRI scenarios has not been well studied. Finally, there is a need to investigate the impact of anthropomorphism on the perception of robots in moral interactions.

This literature review provides a foundation for future research and development of intention-based systems, as well as analysis of their social impact factors. By exploring the gaps in the existing literature, future research can help improve the effectiveness and safety of human-robot interactions in various industries.

Chapter 3

Domain Data

3.1 Overview of JIGSAWS Dataset

The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) is a pioneering dataset designed to facilitate research in the domain of computer-assisted surgical training and evaluation [17]. Collected through a collaboration between Johns Hopkins University and Intuitive Surgical, Inc., JIGSAWS provides a comprehensive collection of synchronized video and kinematic data from robotic surgical systems during various surgical tasks. This dataset is integral to this study, providing the foundation for analyzing and modeling surgical gestures.

The JIGSAWS dataset comprises recordings from eight surgeons performing three fundamental surgical tasks: suturing, knot-tying, and needle-passing. Each task involves complex hand-eye coordination and precision, which are critical components of surgical training. The kinematic data includes positional, orientational, and grasp information of the surgical instruments manipulated by the da Vinci Surgical System. The suturing task, in particular, is the focus of this study, as it involves a series of

predefined gestures that are essential for evaluating surgical skills.

3.2 Detailed Description of JIGSAWS Data

The JIGSAWS dataset is highly relevant for surgical skill analysis and the development of intention recognition models due to its detailed and structured nature. The data collection involved eight right-handed subjects with varying levels of robotic surgical experience, ranging from those with over 100 hours of experience to those with less than 10 hours. Each subject performed each task five times, resulting in a set of trials, each uniquely identified by a specific format. In total, the dataset includes 39 trials of suturing, 36 trials of knot-tying, and 28 trials of needle-passing, although some trials were unusable due to corrupted data recordings.

The suturing task requires the subject to pick up a needle and proceed to an incision designated as a vertical line on the bench-top model. The needle is passed through the "tissue," entering at a marked dot on one side of the incision and exiting at the corresponding dot on the other side. After the first needle pass, the subject extracts the needle, passes it to the right hand, and repeats this process for three more passes. This task is fundamental for surgical training as it involves precise hand-eye coordination and dexterity.

The dataset contains three main components: kinematic data, video data, and manual annotations. The kinematic data is captured using the API of the da Vinci Surgical System at a frequency of 30Hz and includes motion data for the left and right Master Tool Manipulators (MTMs) and the first and second Patient Side Manipulators (PSMs). This data is described by 19 kinematic variables per manipulator,

resulting in a total of 76 variables for the four manipulators. These variables include Cartesian positions, rotation matrices, linear velocities, angular velocities, and gripper angles. The kinematic data is synchronized with the video data, ensuring a consistent temporal alignment.

The video data consists of stereo video recordings from both endoscopic cameras of the da Vinci Surgical System, also captured at 30Hz with a resolution of 640x480 pixels. These video frames are synchronized with the kinematic data, providing a visual context for the recorded movements. The video files are encoded in the FOURCC format with the DX50 codec and are named based on the left and right endoscopic cameras. However, the dataset does not include calibration parameters for the cameras.

Manual annotations in the dataset provide crucial information for supervised learning. Each annotation includes the name of the gesture and the start and end frames in the video. Surgical gestures are defined by watching the video in consultation with a surgeon after data collection. A list of surgical gestures is shown in Table 3.1 All frames for each trial are assigned a gesture label, except for the final frames when the task is finished. Additionally, surgical technical skills are rated by a surgeon using a modified Objective Structured Assessments of Technical Skills (OS-ATS) scale, where each of the six terms is scored on a Likert scale from 1 to 5.

The JIGSAWS dataset inherently supports a Leave-One-User-Out (LOUO) evaluation scheme, where the model is trained on data from all but one surgeon and then tested on the excluded surgeon's data. This evaluation method is critical for assessing the generalizability and robustness of the machine learning models across different users, which is essential for real-world applications in surgical training and

Gesture Index	Gesture Description
G1	Reaching for needle with right hand
G2	Positioning needle
G3	Pushing needle through tissue
G4	Transferring needle from left to right
G5	Moving to center with needle in grip
G6	Pulling suture with left hand
G7	Pulling suture with right hand
G8	Orienting needle
G9	Using right hand to help tighten suture
G10	Loosening more suture
G11	Dropping suture at end and moving to end points
G12	Reaching for needle with left hand
G13	Making C loop around right hand
G14	Reaching for suture with right hand
G15	Pulling suture with both hands

Table 3.1: Gesture list from the JIGSAWS dataset.

skill assessment. The dataset also includes two cross-validation schemes to account for its structure: Leave-One-Supertrial-Out (LOSO) and Leave-One-User-Out (LOUO). LOSO involves creating five folds, each comprising data from one of the five supertrials, while LOUO creates eight folds, each consisting of data from one of the eight subjects.

3.3 Data Preprocessing

The preprocessing of kinematic data from the JIGSAWS dataset is a critical step in preparing the data for training a bidirectional LSTM model. This process involves several stages to ensure that the data is in a suitable format for effective learning.

3.3.1 Data Cleaning and Segmentation

Kinematic data from the suturing task is first extracted from the dataset, focusing on essential features such as positional coordinates, orientation angles, and grip angles of the surgical instruments. This data is then segmented according to the specific suturing gestures being performed, based on timestamps and gesture labels provided in the dataset's annotations. Such segmentation is vital for associating each data snippet with the corresponding surgical gesture, facilitating gesture recognition.

3.3.2 Data Trimming and Label Generation

Following extraction, the kinematic data undergoes trimming to align with the start and end times of each suturing gesture, as indicated in the dataset's transcription files. This step ensures that only the relevant portions of the kinematic data are considered, removing any extraneous information that could potentially hinder the model's learning process. Concurrently, gesture labels are generated and matched with the trimmed kinematic data segments. These labels are crucial for supervised learning, providing the model with the necessary ground truth for each data snippet.

3.3.3 Snippet Generation and Normalization

To accommodate the LSTM model's requirements, the kinematic data is divided into smaller, fixed-length snippets. This division helps in maintaining uniformity across all data samples, ensuring that each input to the model has the same shape and size. Given the temporal nature of the LSTM model, which relies on sequences of data, this step is crucial for capturing the dynamic aspects of surgical gestures. Furthermore, the data is normalized to have a consistent scale, enhancing the model's ability to

learn from the kinematic features.

3.3.4 Training and Validation Split

The final step in the preprocessing pipeline involves splitting the processed kinematic data and corresponding labels into training and validation sets. This split is performed randomly, ensuring a diverse mix of data samples in both sets. Such a division is essential for evaluating the model's performance and its ability to generalize beyond the data it was trained on.

The preprocessing steps described above transform the raw kinematic data from the JIGSAWS dataset into a format that is suitable for training and validating the bidirectional LSTM model. By cleaning, segmenting, and normalizing the data, and by generating appropriate labels for supervised learning, the preprocessing pipeline lays the foundation for developing a robust model capable of recognizing surgical gestures with high accuracy.

Chapter 4

Machine Learning Model

4.1 Prior Knowledge

The domain of surgical gesture recognition, particularly through the analysis of kinematic data from robotic surgery, has garnered significant attention in recent years. This interest is driven by the potential to enhance surgical training, improve the automation of surgical procedures, and offer novel insights into the surgical skill assessment.

4.1.1 Neural Networks for Recognizing Surgical Activities

DiPietro et al. [13] applied recurrent neural networks (RNNs), notably LSTM networks, to recognize surgical activities from robot kinematics. This approach was innovative, diverging from the traditional use of Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), which were limited to recognizing short, low-level activities. DiPietro et al. extended the application to both gestures and

higher-level maneuvers, showcasing RNNs' capability to capture complex temporal dependencies inherent in surgical tasks. Their work established a new benchmark for gesture recognition accuracy and maneuver recognition, emphasizing the effectiveness of RNNs in processing sequential data for surgical applications.

Menegozzo et al. [40] introduced Time Delay Neural Networks (TDNN) applied to kinematic data for surgical gesture recognition, proposing a method that incorporates temporal modeling. By validating their approach on the JIGSAWS dataset and a novel dataset from virtual training simulators, they highlighted TDNN's generalization capability and computational efficiency, marking an important step towards real-time applications in surgical systems.

4.1.2 Bidirectional LSTM

The exploration of bidirectional LSTM (BiLSTM) models, capable of processing data in both forward and backward directions, offers a promising avenue for capturing the full spectrum of temporal dynamics in surgical gestures. This approach aims to leverage the strengths of LSTM models in recognizing long-term dependencies while addressing the limitations of unidirectional models in capturing patterns and dependencies that emerge across the entire sequence of surgical activities.

4.1.3 Evolutionary Computation for Hyperparameter Optimization

The optimization of machine learning models, including BiLSTMs, remains a critical challenge. The application of evolutionary computation techniques for hyperparameter tuning presents an innovative solution. By systematically exploring the hyperparameter space, this approach seeks to identify optimal configurations that enhance model performance, addressing a gap in the existing literature on surgical gesture recognition.

The collective insights from these studies underline the progress and challenges in the field of surgical gesture recognition. DiPietro et al. and Menegozzo et al. have laid the groundwork by demonstrating the potential of LSTM and TDNN models in this domain. The exploration of BiLSTM models and evolutionary computation techniques for hyperparameter tuning represents the next frontier in advancing the accuracy and efficiency of these models. This study contributes to this evolving landscape by harnessing the capabilities of BiLSTMs, augmented with advanced hyperparameter optimization, to offer new perspectives on analyzing complex kinematic data for surgical gesture recognition.

4.2 Model

4.2.1 Long Short-term Memory (LSTM) Neural Network

Recurrent Neural Networks (RNNs) are a class of neural networks that are particularly effective for sequential data because they include loops that allow information to persist. Unlike feedforward neural networks, which process inputs independently,

RNNs take the sequence's previous states into account, making them suitable for tasks like time-series prediction and natural language processing. However, traditional RNNs suffer from the vanishing gradient problem, which limits their ability to learn long-term dependencies.

LSTMs, a specialized form of Recurrent Neural Networks (RNNs), are designed to address the vanishing gradient problem inherent in traditional RNNs, enabling the model to learn and retain long-term dependencies in sequential data more effectively. This capability is critical in surgical gesture recognition where the context and sequence of actions play a significant role in accurate classification.

LSTMs overcome these limitations through the use of a more complex architecture that includes memory cells and gates. The key to LSTM's success is its cell state and the use of gates—forget gate f_t, input gate i_t, and output gate o_t—which regulate the flow of information. These gates decide which information is important to keep or discard as the sequence progresses, making LSTMs particularly adept at modeling time-series data like kinematic sequences in surgery.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4.2.1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4.2.2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4.2.3}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4.2.4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4.2.5}$$

$$h_t = o_t * \tanh(C_t) \tag{4.2.6}$$

In these equations, f_t is the forget gate's activation vector, deciding what information to discard from the cell state; i_t is the input gate's activation vector, determining which new information is added to the cell state; $\tilde{C}_t$ is the candidate vector for addition to the cell state; C_t is the new cell state, updated by forgetting old information and adding new information; o_t is the output gate's activation vector, deciding what part of the cell state makes it to the output; h_t is the output vector of the LSTM unit, based on the cell state and the output gate's activation; σ denotes the sigmoid function; and tanh is the hyperbolic tangent function, both of which are activation functions that help regulate the flow of information [21, 20].

4.2.2 Bidirectional LSTM

BiLSTMs enhance the LSTM architecture by processing data in both forward and backward directions, thus capturing context from both past and future data points as shown in Figure 4.1. This bidirectional processing is particularly beneficial in surgical gesture recognition, where the context before and after a specific gesture can provide valuable cues for accurate classification. The forward pass $\overrightarrow{h_t}$ and the backward pass $\overleftarrow{h_t}$ are computed independently and their outputs are concatenated to form the final output h_t. This allows BiLSTMs to have a more comprehensive understanding of the sequence, offering an advantage over traditional LSTMs which only consider past context. BiLSTMs are particularly adept at recognizing patterns and dependencies that are not immediately apparent, enhancing the model's predictive capabilities.

$$\overrightarrow{h_t} = \overrightarrow{\text{LSTM}}(x_t) \tag{4.2.7}$$

$$\overleftarrow{h_t} = \overleftarrow{\text{LSTM}}(x_t) \tag{4.2.8}$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{4.2.9}$$

Here, $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ represent the hidden states from forward and backward LSTMs at time t, respectively, and h_t is the concatenated hidden state combining information from both directions, providing a comprehensive view of the sequence from both past and future perspectives.

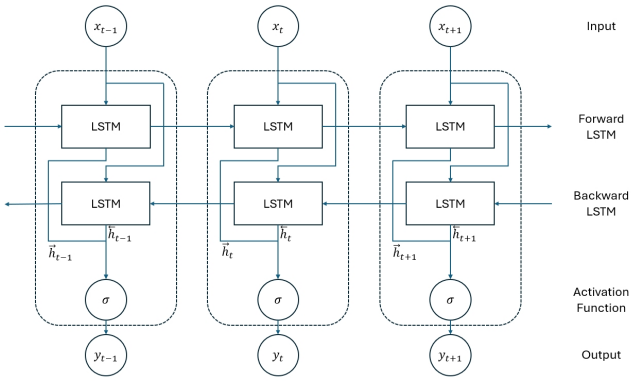

Figure 4.1: Conceptual diagram of a BiLSTM

4.2.3 Hyperparameter Tuning with Evolutionary Computation Technique

Evolutionary computation techniques, inspired by natural selection, offer a robust alternative to traditional hyperparameter tuning methods such as grid search and random search. These traditional methods, while straightforward, can be inefficient

and may not effectively explore the hyperparameter space due to their exhaustive or random nature, which can lead to excessive computational costs and suboptimal performance [3]. Evolutionary algorithms, on the other hand, use mechanisms like mutation, crossover, and selection to iteratively evolve the hyperparameters towards optimal configurations. This process begins with an initial population of hyperparameter sets as demonstrated example in Figure 4.2, which is then evaluated based on model performance. The best-performing sets are selected and used to generate new sets through crossover and mutation. This cycle repeats until the algorithm converges on a set of hyperparameters that yield the best model performance, typically determined by a convergence criterion such as a predefined number of generations or a threshold for performance improvement. This approach is particularly advantageous in complex models like LSTMs and BiLSTMs, where the hyperparameter space is vast and the relationships between hyperparameters and model performance are nonlinear. Evolutionary computation not only enhances the efficiency of the tuning process but also increases the likelihood of discovering optimal configurations that might be missed by more conventional tuning methods [57].

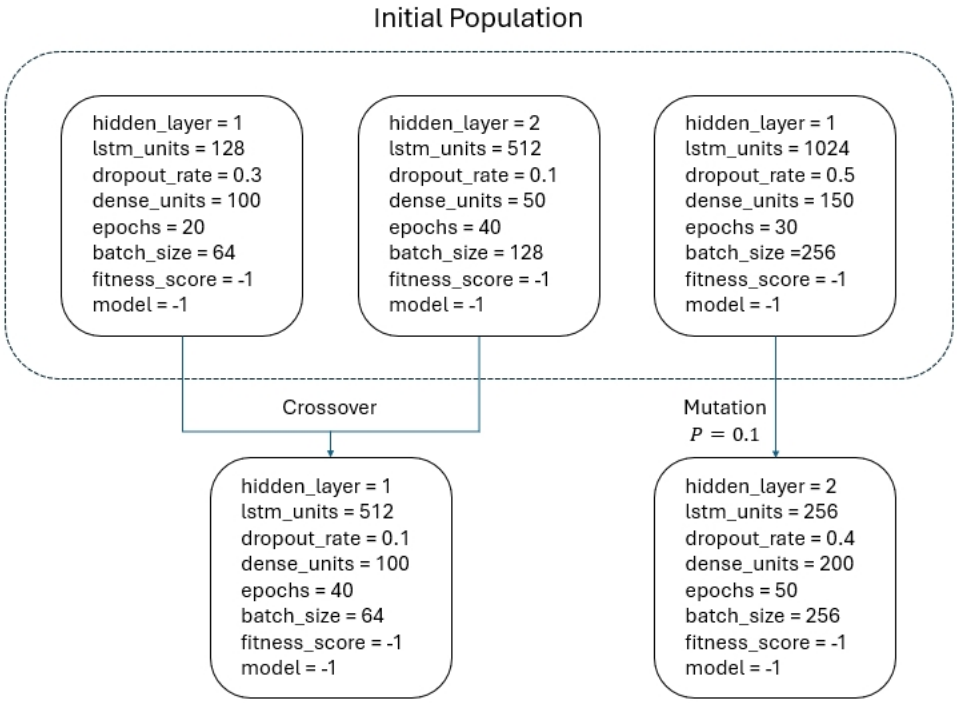

Figure 4.2: An example of initial population

Chapter 5

Methodology

5.1 Experimental Setup

The methodology for this study involves a systematic approach to validate the effectiveness of the bidirectional LSTM (BiLSTM) model optimized with evolutionary computation techniques for recognizing surgical gestures. This chapter details the experimental setup, including the data preprocessing, model training, evaluation metrics, and validation methods. Figure 5.1 shows the overall process of the setup.

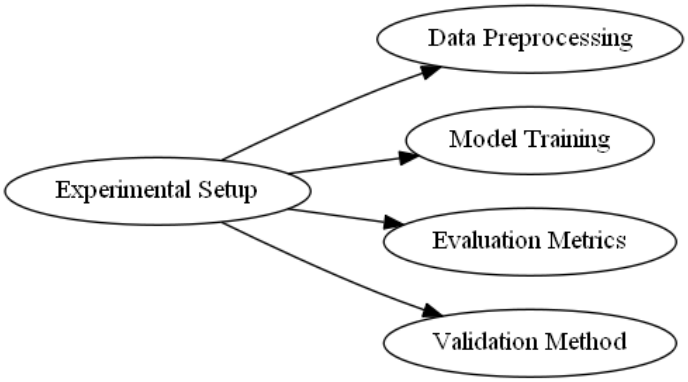

Figure 5.1: Experimental setup flow diagram

5.2 Data Preprocessing

The first step in the methodology is the preprocessing of kinematic data from the JIGSAWS dataset. This involves several stages to ensure that the data is in a suitable format for training the BiLSTM model. The kinematic data from the suturing task was extracted, focusing on essential features such as positional coordinates, orientation angles, and grip angles of the surgical instruments. These features were then segmented according to the specific suturing gestures being performed, based on timestamps and gesture labels provided in the dataset's annotations. This segmentation is vital for associating each data snippet with the corresponding surgical gesture, facilitating gesture recognition.

Following extraction, the kinematic data underwent trimming to align with the start and end times of each suturing gesture, as indicated in the dataset's transcription files. This step ensured that only the relevant portions of the kinematic data were considered, removing any extraneous information that could potentially hinder the model's learning process. Concurrently, gesture labels were generated and matched with the trimmed kinematic data segments, providing the necessary ground truth for supervised learning.

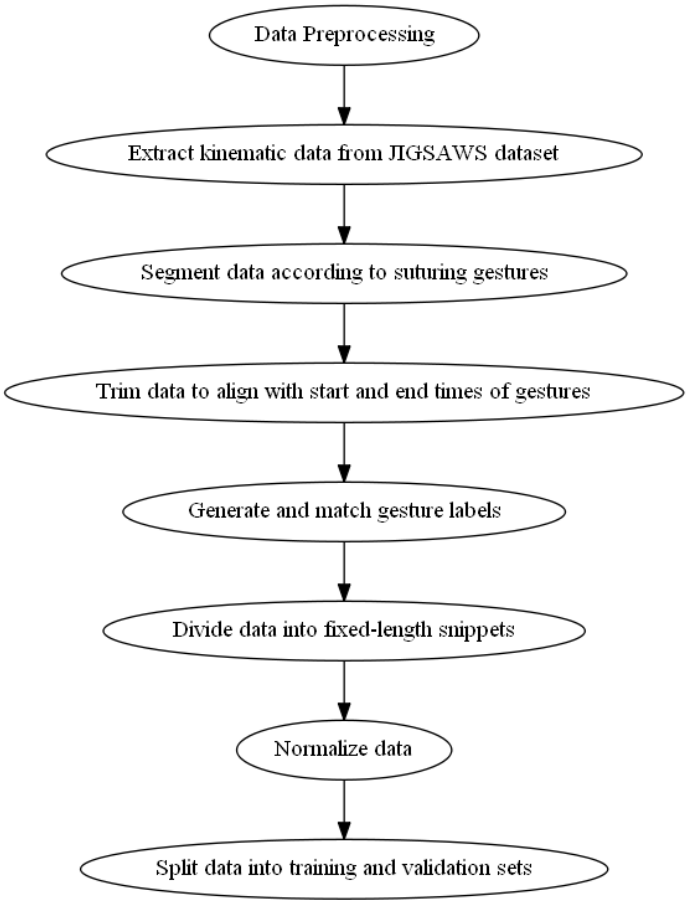

Figure 5.2: Data preprocessing flow diagram

To accommodate the LSTM model's requirements, the kinematic data was divided into smaller, fixed-length snippets. This division helped in maintaining uniformity across all data samples, ensuring that each input to the model had the same shape and size. Given the temporal nature of the LSTM model, which relies on sequences of data, this step was crucial for capturing the dynamic aspects of surgical gestures. Furthermore, the data was normalized to have a consistent scale, enhancing the model's ability to learn from the kinematic features. Finally, the processed kinematic data and corresponding labels were split into training and validation sets. This split was

performed randomly, ensuring a diverse mix of data samples in both sets, which was essential for evaluating the model's performance and its ability to generalize beyond the data it was trained on.

5.3 Model Training

The BiLSTM model architecture, optimized through evolutionary computation, consisted of two bidirectional LSTM layers with 256 units each, followed by a dense layer. This architecture was selected based on its ability to capture the temporal dependencies in the kinematic data while maintaining computational efficiency. The model was trained using the Adam optimizer, with a learning rate and batch size determined through hyperparameter tuning. Training involved multiple epochs, during which the model learned to recognize and classify the predefined surgical gestures. The training process was monitored by observing the training and validation loss curves to ensure effective learning and to prevent overfitting. The evolutionary computation technique was employed to optimize hyperparameters such as the number of LSTM units, learning rate, batch size, and dropout rate. This approach systematically explored the hyperparameter space, identifying configurations that balanced model complexity, training time, and generalization performance.

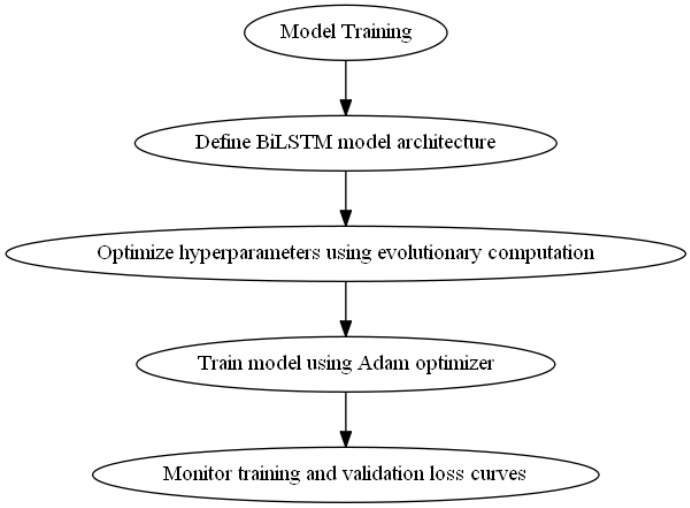

Figure 5.3: Model training flow diagram

5.4 Validation Method

The Leave-One-User-Out (LOUO) cross-validation scheme was employed to evaluate the robustness and generalizability of the BiLSTM model [17]. In this scheme, the model was trained on data from all but one surgeon and tested on the excluded surgeon's data. This process was repeated for each surgeon, ensuring that the model's performance was assessed across all users by averaging the accuracy. This evaluation method is critical for assessing the generalizability and robustness of the machine learning models across different users, which is essential for real-world applications in surgical training and skill assessment. The LOUO validation method provided insights into the model's performance under different conditions, ensuring that the developed model could generalize well to new, unseen data, making it suitable for practical applications in surgical gesture recognition.

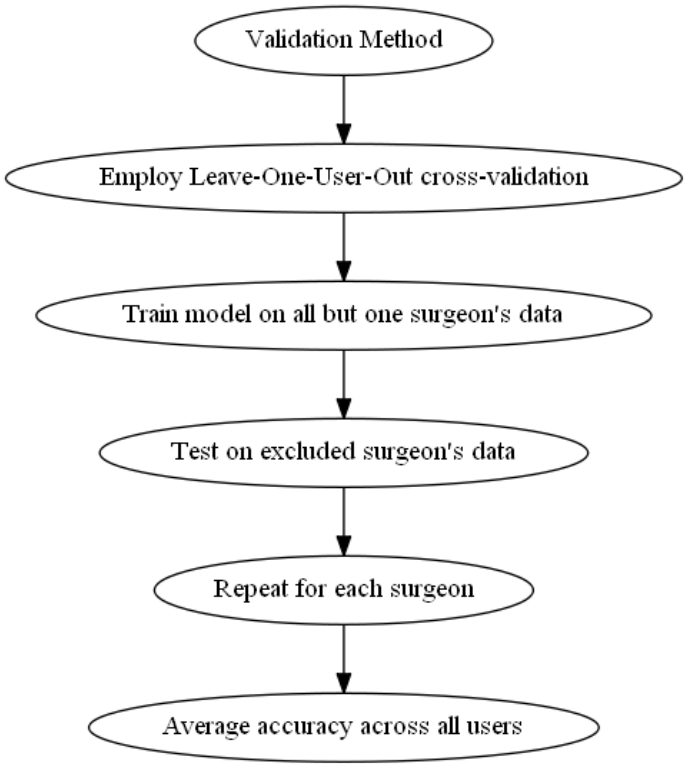

Figure 5.4: Validation methods flow diagram

5.5 Evaluation Metrics

The model's performance was evaluated using several metrics, including the comparison of train/validation curves, accuracy comparison, confusion matrix analysis, prediction time, and model size. These metrics provided a comprehensive assessment of the model's ability to recognize and classify surgical gestures accurately. The training and validation loss curves were compared to assess the model's learning progress and detect overfitting. Smooth convergence of these curves indicated effective learning. Accuracy was measured to determine the overall correctness of the model by comparing the number of correct predictions to the total number of predictions. The confusion matrix was used to visualize the model's performance in classifying each

gesture, highlighting the true positives, false positives, false negatives, and true negatives. Prediction time was tracked to evaluate the model's efficiency, which is crucial for real-time applications. The size of the trained model was measured to ensure it is suitable for deployment in resource-constrained environments. For the BiLSTM model optimized with evolutionary computation, the evolution of hyperparameters was also tracked to understand how the optimal configuration was reached. This tracking provided insights into the effectiveness of the evolutionary algorithm in hyperparameter tuning.

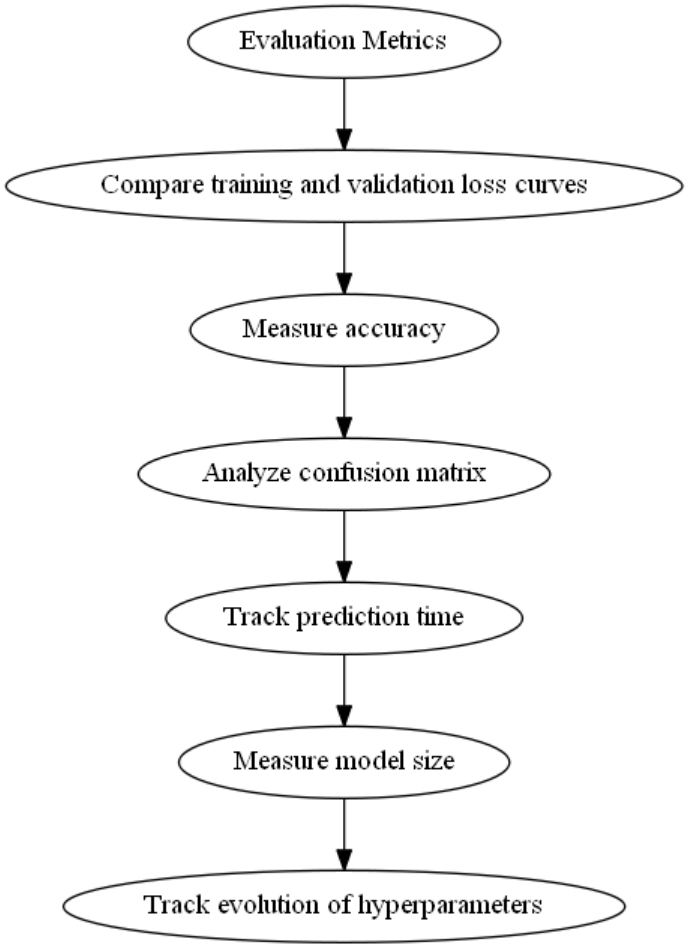

Figure 5.5: Evaluation metrics flow diagram

Chapter 6

Results

This chapter presents the findings from the experimental evaluation of the BiLSTM model optimized using evolutionary computation techniques for surgical gesture recognition. The results include an analysis of gesture counts, model performance, hyperparameter evolution, and model evaluation through confusion matrices and other metrics. Each section provides a detailed breakdown of these aspects to highlight the effectiveness and efficiency of the proposed approach.

6.1 Gesture Count Analysis

As part of our study, we analyzed the gesture counts within the dataset for a leave-one-user-out (LOUO) evaluation, differentiating between the training and testing sets by a 80% to 20% train-test split.

6.1.1 Training Set Gesture Counts

Figure 6.1 illustrates the gesture counts in the training set for the leave-one-user-out evaluation. In the training set, Gesture 3 has the highest count, with over 2000 instances. This indicates it is the most frequently performed gesture among the training data, due to the nature of suturing task. Gestures 2, 6, and 11 also have high counts, with each having over 1000 instances. This suggests that these gestures are also commonly performed during the training tasks. In contrast, Gestures 1, 9, and 10 have relatively low counts, with Gesture 1 having the least occurrences, under 250 instances. This uneven distribution could pose a challenge for the model to learn less frequent gestures effectively.

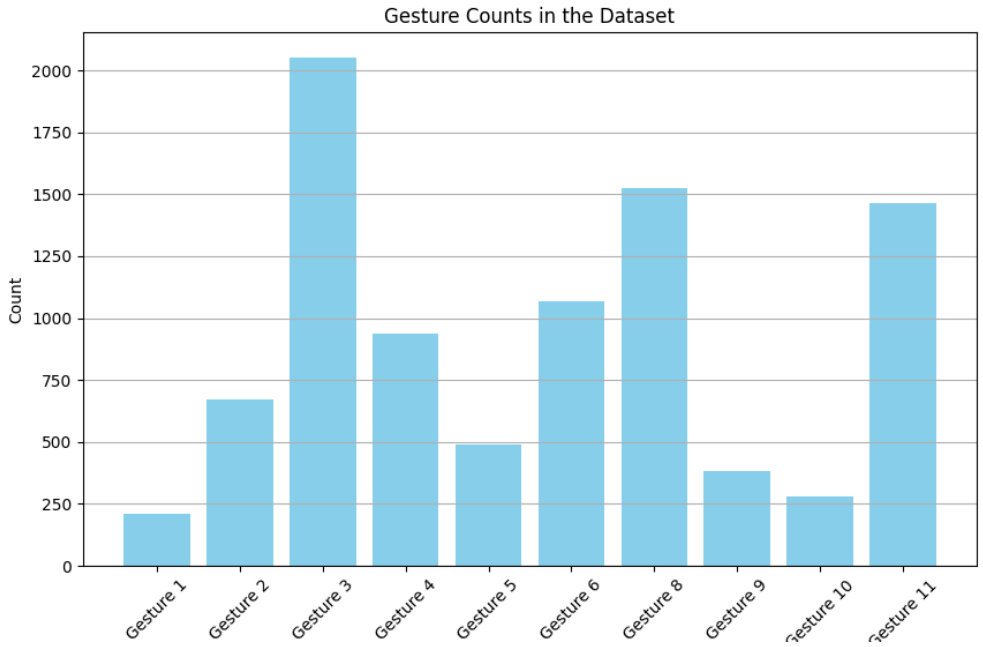

Figure 6.1: Gesture counts in the training set for leave-one-user-out evaluation. Gesture 3 has the highest count with over 2000 instances, while Gesture 1 has the least occurrences.

6.1.2 Testing Set Gesture Counts

Figure 6.2 shows the gesture counts in the testing set for the leave-one-user-out evaluation. In the testing set, similar to the training set, Gesture 3 is the most frequent, with around 1250 instances. Gestures 2, 6, and 11 again show higher counts compared to other gestures, although their counts are significantly lower than in the training set. The counts for Gestures 1, 9, and 10 remain low, with Gesture 1 again being the least frequent.

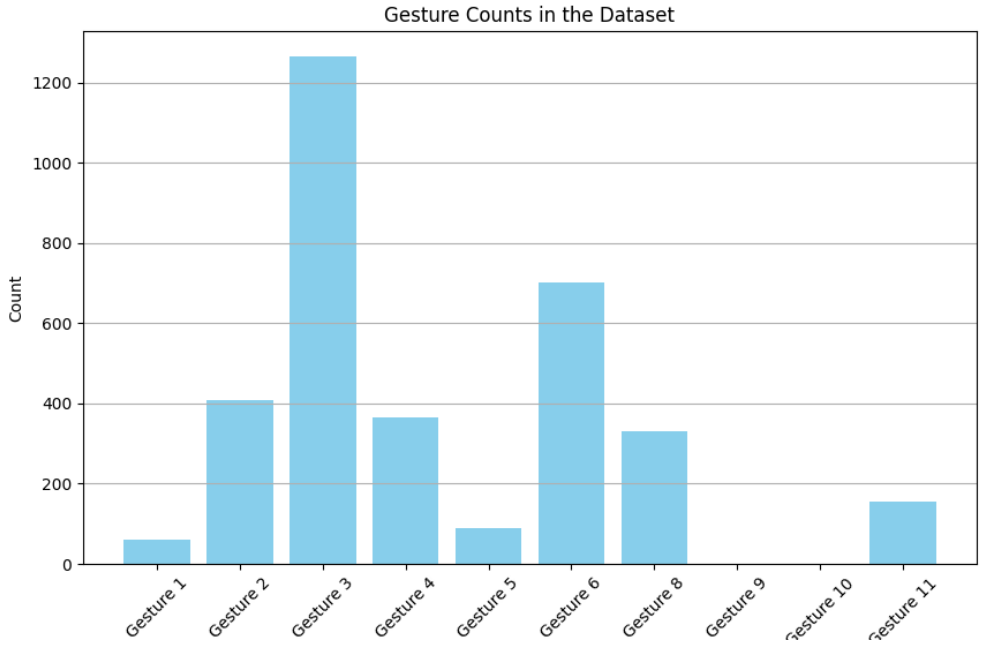

Figure 6.2: Gesture counts in the testing set for leave-one-user-out evaluation. Gesture 3 is the most frequent with around 1250 instances, while Gesture 1 remains the least frequent.

6.2 Model Performance Analysis

In this section, we present the performance of the BiLSTM model optimized with evolutionary computation compared to the standard LSTM model. We focus on training and validation loss trends to evaluate the models' learning and generalization capabilities.

6.2.1 Training and Validation Loss

Figure 6.3 shows the training and validation loss for the BiLSTM model with hyperparameters optimized using evolutionary computing. We observe that the training loss decreases steadily, converging smoothly as the epochs progress. The validation loss follows a similar trend, initially decreasing and then stabilizing around the same level as the training loss. This indicates that the model is effectively learning from the training data without overfitting, thanks to the well-tuned hyperparameters. The close alignment between the training and validation loss curves suggests that the model generalizes well to unseen data.

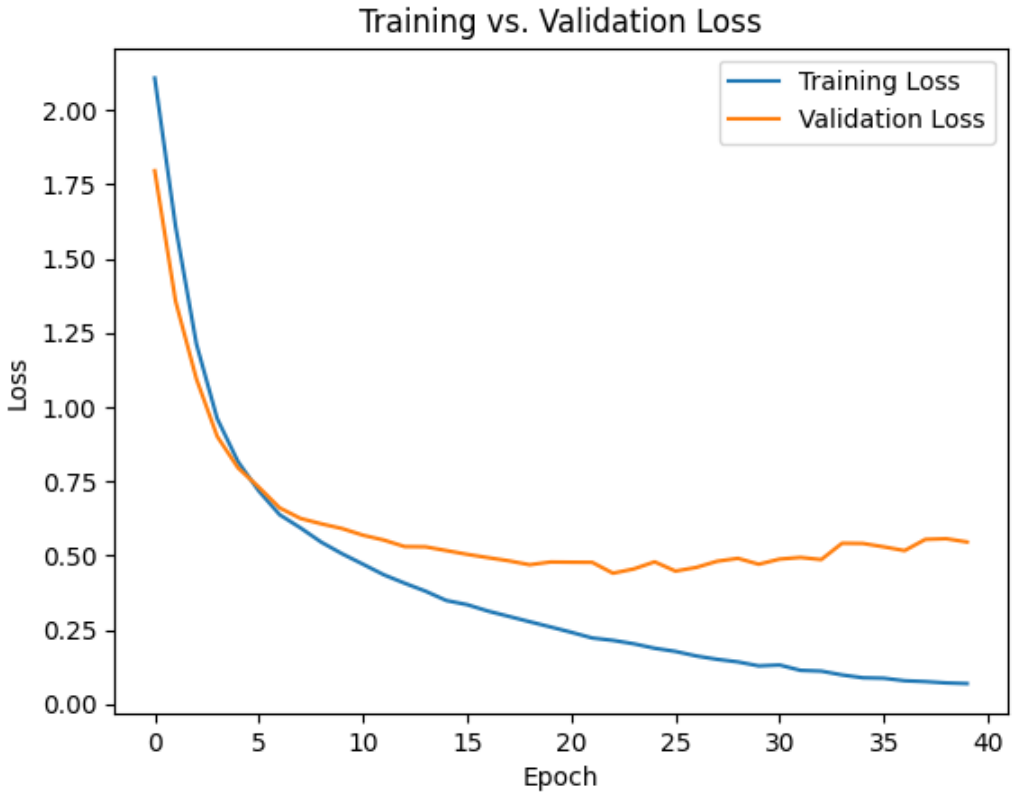

Figure 6.3: Training vs. Validation Loss for BiLSTM with Evolutionary Computing Hyperparameter Tuning. The model shows smooth convergence with closely aligned training and validation loss curves, indicating good generalization.

For contrast, Figure 6.4 shows the training and validation loss for the standard LSTM model without evolutionary computing hyperparameter tuning. In this case, while the training loss decreases and stabilizes, the validation loss exhibits more fluctuation and does not converge as smoothly as the BiLSTM model. This suggests potential overfitting or inadequate learning from the training data, likely due to suboptimal hyperparameter settings. The divergence between the training and validation loss indicates that the model may not generalize as effectively to new data.

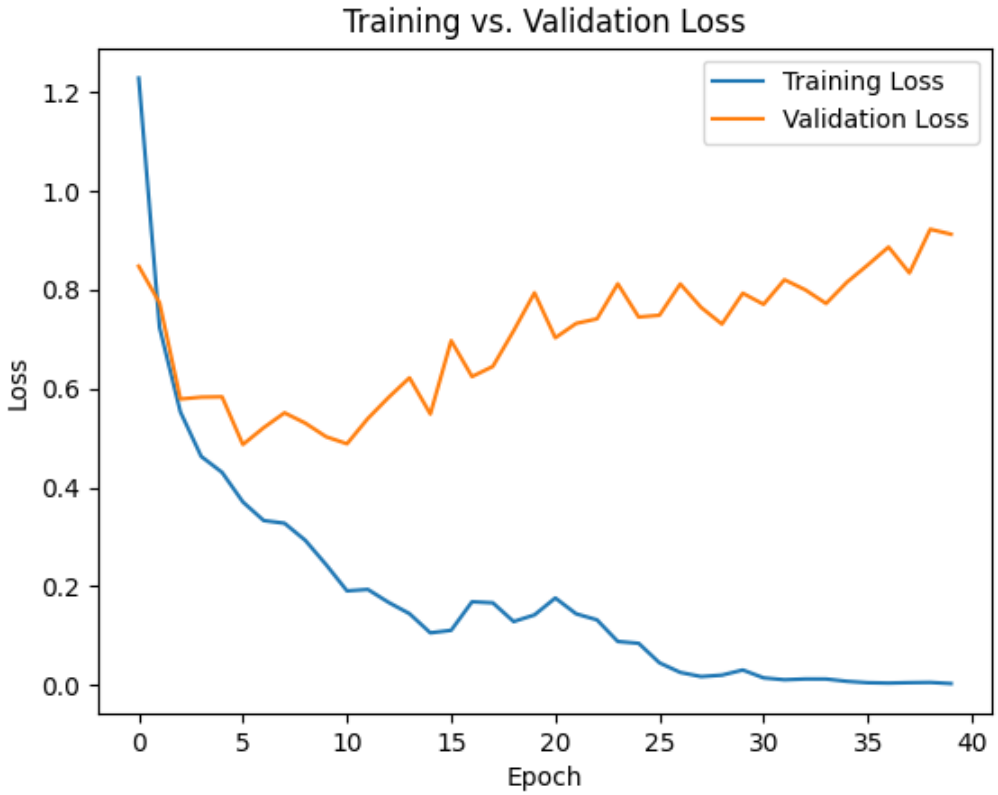

Figure 6.4: Training vs. Validation Loss for LSTM without Evolutionary Computing Hyperparameter Tuning. The validation loss exhibits fluctuations and does not converge smoothly, indicating potential overfitting and suboptimal hyperparameter settings.

6.2.2 Hyperparameter Evolution

By efficiently explore the hyperparameter space, the evolutionary algorithm was able to identify configurations that balanced model complexity, training time, and generalization performance. The tuned BiLSTM model with optimized hyperparameters showed smoother and more stable convergence in both training and validation loss, as discussed in the previous section. In Figure 6.5, the left subplot illustrates the

evolution of the number of LSTM units across generations. Initially, the number of units starts at a higher value but quickly stabilizes to a lower, optimal range as the evolutionary algorithm identifies the configurations that lead to better model performance. This reduction and stabilization indicate that a lower number of LSTM units is sufficient for capturing the temporal dependencies in the dataset, balancing model complexity and training efficiency.

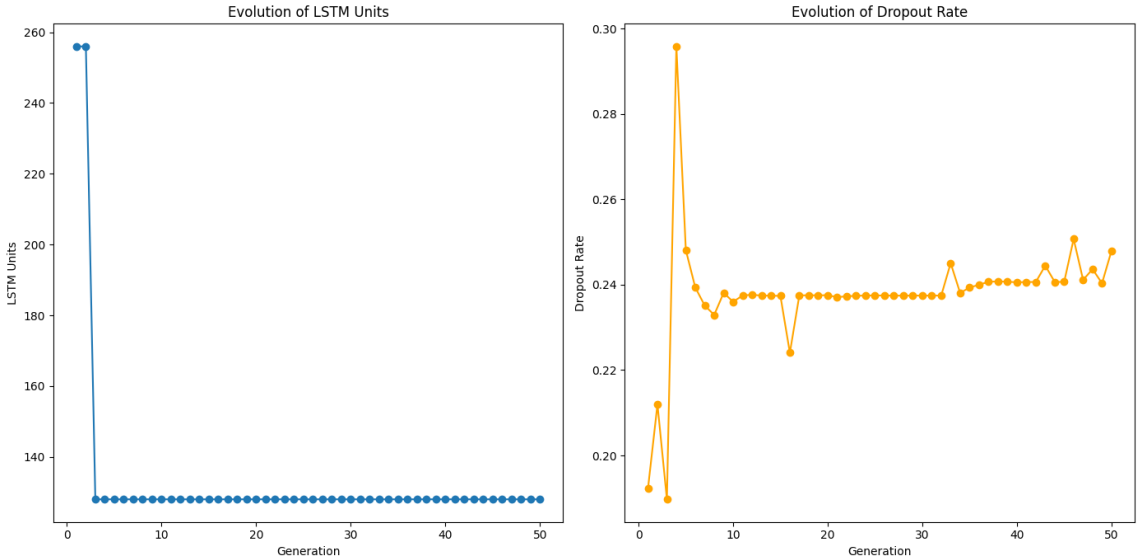

Figure 6.5: Evolution of LSTM Units and Dropout Rate using Evolutionary Computation. The left plot shows the stabilization of LSTM units over generations, indicating an optimal lower range for capturing temporal dependencies. The right plot demonstrates the fluctuation and convergence of the dropout rate, optimizing for model generalization and preventing overfitting.

The right subplot shows the evolution of the dropout rate over the generations. The dropout rate begins with higher values and fluctuates before converging to a more stable range. The initial variations suggest that the evolutionary algorithm is exploring diverse dropout rates to identify the optimal setting that prevents overfitting while maintaining model performance. The final stabilization of the dropout rate

indicates that the evolutionary algorithm has effectively fine-tuned this parameter to enhance the model's generalization capabilities.

Figure 6.6 shows two lines: the best fitness (dashed line) and the average fitness (solid line) of the population over 50 generations. The best fitness line demonstrates that the evolutionary algorithm consistently identifies better-performing hyperparameter configurations as the generations progress, although with some fluctuations due to the stochastic nature of the process. The average fitness line shows a general upward trend, indicating overall improvement in the population's performance.

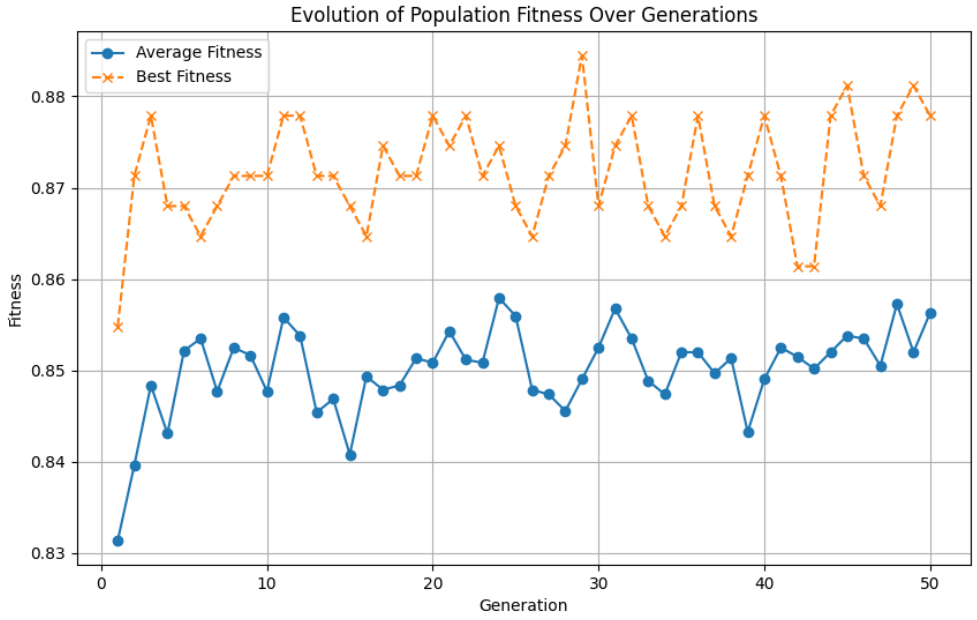

Figure 6.6: Evolution of Population Fitness Over Generations. The best fitness (dashed line) shows consistent identification of better-performing hyperparameter configurations, while the average fitness (solid line) indicates overall improvement in the population's performance.

The gap between the best fitness and the average fitness lines suggests that while the algorithm finds superior solutions, there is still variability in the population. This

variability is crucial for the evolutionary process as it ensures diversity, preventing premature convergence to suboptimal solutions and allowing for continued exploration of the hyperparameter space.

6.2.3 LOUO Cross Validation

The box plot shows the model's performance across the 8 different folds from the Leave-One-User-Out (LOUO) cross-validation. Each fold represents how well the model did when trained on data from all but one user and tested on the excluded user.

Fold 6 had the highest accuracy, with a median just over 84%, while Fold 7 had the lowest, dipping around 77%. This variation suggests that the model's performance can vary depending on the specific user's data. Despite this, most folds maintain an accuracy between 78% and 82%, indicating a fairly consistent performance across users.

We also see a few outliers in Fold 4 and Fold 7, suggesting there were some specific samples within those folds where the model struggled more. Overall, though, the results show that the model generalizes reasonably well, even if it finds some users' data more challenging.

6.3 Model Performance Evaluation

This section evaluates the model performance through confusion matrices and other key metrics, providing a detailed breakdown of the BiLSTM model's accuracy in recognizing surgical gestures.

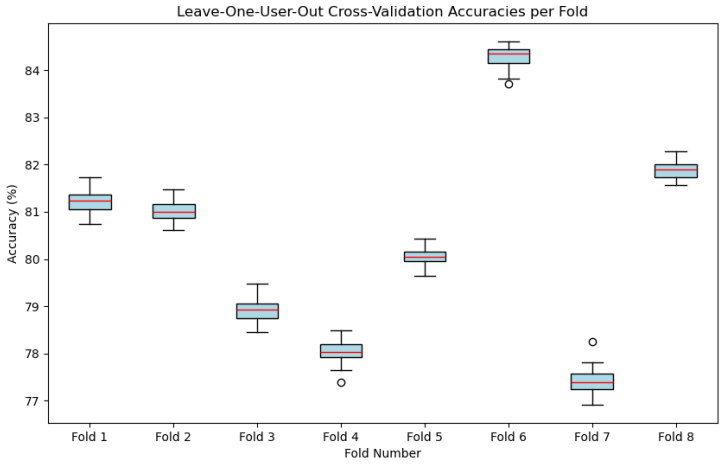

Figure 6.7: Leave-One-User-Out Cross-Validation Accuracies per Fold. The box plot shows the accuracy distribution for each fold, with the median marked by the red line and potential outliers shown as individual points.

6.3.1 Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of the model's performance by showing the correct and incorrect predictions for each class. In the BiLSTM model with optimized hyperparameters, we observe strong performance in recognizing several gestures. For instance, in Figure 6.8 gesture 2 (corresponding to gesture index 3, 'Pushing needle through tissue') shows 161 correct predictions. However, there are still notable misclassifications. Gesture 2 is sometimes incorrectly predicted as gesture 9 (corresponding to gesture index 11, 'Dropping suture at end and moving to end points'). Additionally, gesture 5 (corresponding to gesture index 6, 'Pulling suture with left hand') is recognized relatively well, with 105 correct predictions, but also shows some misclassifications as gestures 6 (corresponding to gesture index 7, 'Pulling suture with right hand'). Gesture 6 (corresponding to gesture index 7, 'Pulling suture with right hand') is often misclassified as gestures 0 to 5, indicating a

challenge in distinguishing this gesture accurately.

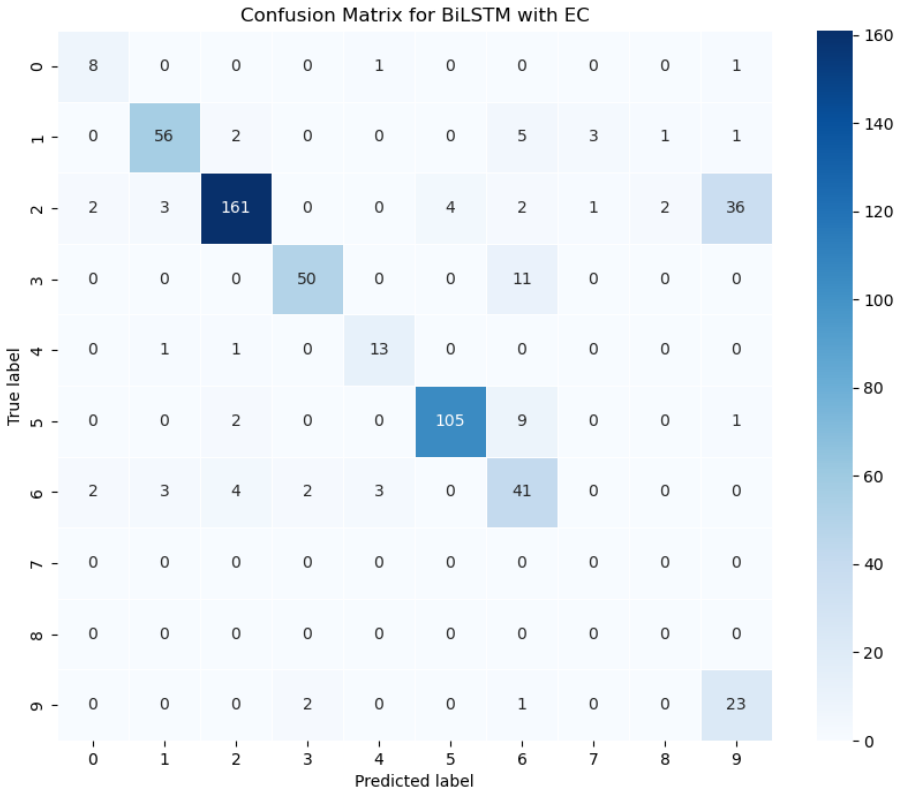

Figure 6.8: Confusion Matrix for BiLSTM with Evolutionary Computing Hyperparameter Tuning. The model shows strong performance in recognizing several gestures, particularly gesture 2, with reduced misclassification rates.

Comparably, in Figure 6.9, the standard LSTM model without hyperparameter tuning shows less accuracy in several gesture recognitions. Gesture 2 is still the most accurately predicted gesture with 151 correct predictions, though this is fewer than the BiLSTM model's 161 correct predictions. There are higher misclassification rates across the board, with gestures being more frequently confused with each other. For instance, gesture 2 (corresponding to gesture index 3, 'Pushing needle through

tissue') is often misclassified as gestures 1 and 9 (corresponding to gesture index 2, 'Positioning the needle' and gesture index 11, 'Dropping suture at end and moving to end points'). Gesture 3 (corresponding to gesture index 4, 'Transferring needle from left to right') is more frequently confused with gesture 6 (corresponding to gesture index 7, 'Pulling suture with right hand'). Additionally, gesture 5 (gesture index 6, 'Pulling suture with left hand') shows notable misclassifications as gesture 6 (corresponding to gesture index 7, 'Pulling suture with right hand'). The confusion matrix for the LSTM model without evolutionary computing hyperparameter tuning demonstrates lower performance in correctly classifying gestures compared to the BiLSTM model with such tuning.

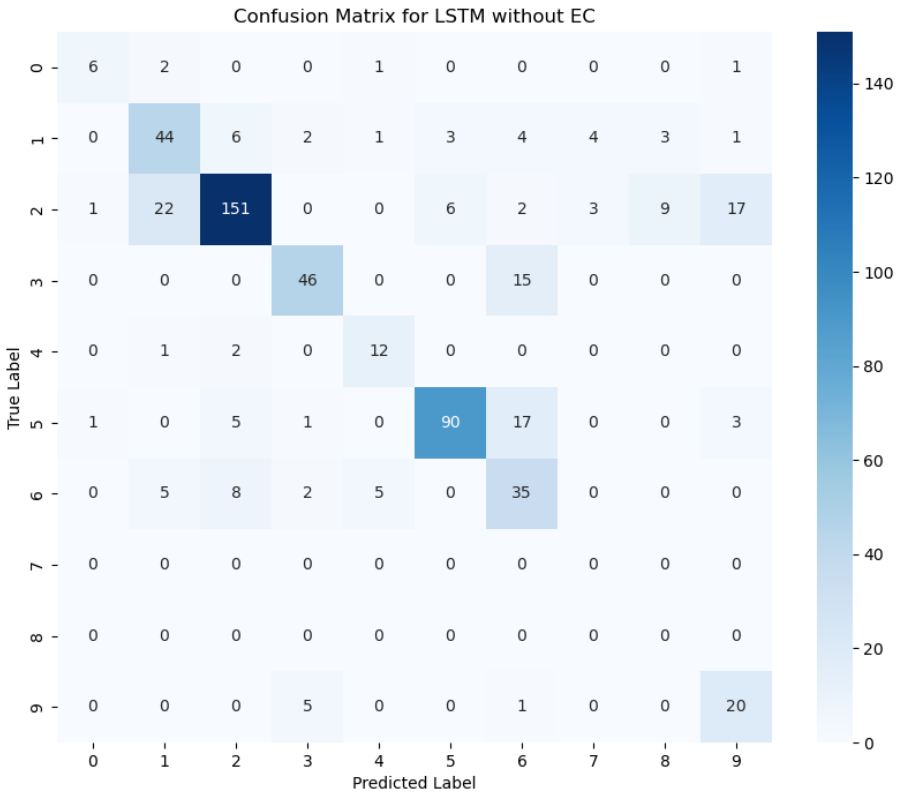

Figure 6.9: Confusion Matrix for LSTM without Evolutionary Computing Hyperparameter Tuning. The model shows higher misclassification rates compared to the BiLSTM model with optimized hyperparameters.

6.4 Model Size and Prediction Time

In addition to evaluating the performance of the models through accuracy and confusion matrices, it is essential to consider the model size and prediction time, which impact the feasibility of deploying these models in real-world applications. Here, we compare the size and average prediction time of two models: the LSTM without

evolutionary computing hyperparameter tuning and the BiLSTM with evolutionary computing hyperparameter tuning.

The model sizes are a critical factor, especially when deploying on devices with limited storage and memory capabilities. The LSTM model without EC has a size of 97.68 MB, while the BiLSTM model with EC has a significantly smaller size of 6.28 MB. The BiLSTM model with evolutionary hyperparameter tuning is considerably more compact, which is advantageous for deployment on devices that has limited memory resources. This reduction in size is due to the optimized configuration of the model, achieved through evolutionary computation, which identifies the most efficient architecture without compromising performance.

The detailed architecture and number of parameters for both models further showcase the efficiency of the BiLSTM model. The LSTM model without EC has a significantly larger number of parameters, with a total of 8,531,978 parameters. The architecture includes a LSTM layer with 2048 units followed by a dense layer. In contrast, the BiLSTM model with EC is more compact with a total of 543,242 parameters. Its architecture consists of two bidirectional LSTM layers with 256 units each, followed by a dense layer. This reduction in parameters not only decreases the model size but also improves computational efficiency.

Prediction time is another critical metric, particularly for real-time applications where quick inference is necessary. The LSTM model has an average prediction time of 0.073208 seconds, while the BiLSTM model has a slightly faster average prediction time of 0.068100 seconds. The BiLSTM model with evolutionary hyperparameter tuning not only achieves better performance but also provides faster predictions. This is likely due to the streamlined architecture and reduced parameter count, resulting

in more efficient computations.

Table 6.1: Model Size and Prediction Time Comparison

Metric	**LSTM (No EC)**	**BiLSTM (With EC)**
Model Size (MB)	97.68	6.28
Number of Parameters	8,531,978	543,242
Average Prediction Time (seconds)	0.073208	0.068100

6.5 Summary of Results

In summary, the BiLSTM model optimized with evolutionary computation outperforms the standard LSTM model in several key areas. It achieves lower and more stable training and validation loss, better generalization to unseen data, and improved accuracy in recognizing surgical gestures. The evolutionary algorithm effectively navigates the hyperparameter space, leading to a more compact and efficient model with faster prediction times. These results highlight the advantages of using advanced hyperparameter optimization techniques in developing high-performing deep learning models for complex tasks such as surgical gesture recognition.

note

Chapter 7

Discussion

The results of our study highlight several key observations regarding the performance of the BiLSTM model optimized with evolutionary computation. The analysis of gesture counts in the dataset revealed an uneven distribution of gestures, with some gestures being significantly more frequent than others. This imbalance posed a challenge for the model to learn less frequent gestures effectively. Despite this, the BiLSTM model demonstrated strong performance, particularly in recognizing the most common gestures, as evidenced by the confusion matrix analysis. The BiLSTM model achieved an accuracy of 81.17%, while the standard LSTM model without evolutionary computation tuning had an accuracy of 71.76%.

The comparison between the BiLSTM model with evolutionary hyperparameter tuning and the standard LSTM model without such tuning underscores the importance of effective hyperparameter optimization. The BiLSTM model, which took 9.46 seconds to train, achieved lower and more stable training and validation loss, indicating effective learning and good generalization to unseen data. In contrast, the standard LSTM model, which trained in 5.98 seconds, exhibited more fluctuation

in the validation loss, suggesting potential overfitting and inadequate learning from the training data. The longer training time for the BiLSTM model is due to its bidirectional nature, which processes data in both forward and backward directions, leading to more comprehensive learning but at the cost of increased computational complexity. The close alignment between the training and validation loss curves for the BiLSTM model suggests that the model is well-tuned and capable of generalizing effectively to new data.

The comparison between the two models underscores the importance of effective hyperparameter tuning. The BiLSTM model with evolutionary computing hyperparameter tuning not only achieves lower and more stable training and validation loss but also demonstrates better generalization to unseen data. This highlights the benefits of using evolutionary algorithms to explore the hyperparameter space efficiently, leading to improved model performance and robustness. These observations align with our hypothesis that BiLSTM models, when coupled with sophisticated hyperparameter tuning techniques, can outperform traditional LSTM models in complex tasks such as surgical gesture recognition. The results emphasize the critical role of hyperparameter optimization in developing high-performing deep learning models for sequential data analysis.

The evolution of hyperparameters tracked during the training of the BiLSTM model provided valuable insights into the optimization process. The reduction and stabilization of the number of LSTM units and the convergence of the dropout rate to a stable range indicate that the evolutionary algorithm was successful in identifying optimal configurations that balance model complexity, training efficiency, and generalization performance. The evolutionary computation technique not only enhanced

the efficiency of the tuning process but also increased the likelihood of discovering optimal configurations that might be missed by more conventional tuning methods.

The confusion matrix analysis highlighted the improved performance of the BiLSTM model in correctly classifying gestures compared to the standard LSTM model. Although there were still some misclassifications, the overall accuracy and the number of correct predictions for each gesture were higher for the BiLSTM model. This demonstrates the model's ability to capture complex temporal dependencies in the kinematic data, thanks to the bidirectional processing and the optimized hyperparameters.

Considering the model size and prediction time, the BiLSTM model with evolutionary hyperparameter tuning proved to be more efficient and suitable for deployment in real-world applications. The significantly smaller model size and faster prediction time make it a viable candidate for use in resource-constrained environments where storage and computational power are limited. The reduced number of parameters in the BiLSTM model, achieved through evolutionary optimization, contributed to its compact size and efficient computation.

The integration of BiLSTM networks with evolutionary hyperparameter tuning offers a robust solution for surgical gesture recognition, providing a valuable tool for enhancing Human-Robot Interaction (HRI) in surgical contexts. By accurately recognizing and classifying surgical gestures, the model can be used to develop intelligent robotic assistants that can anticipate and respond to a surgeon's needs during an operation. This capability could significantly enhance the precision and efficiency of surgical procedures, reducing the cognitive load on surgeons and improving patient outcomes.

Moreover, the model's ability to generalize across different surgeons and gestures makes it a promising candidate for surgical training and skill assessment. Trainee surgeons can receive real-time feedback on their performance, with the system recognizing and evaluating their gestures against established benchmarks. This automated assessment can provide consistent and objective evaluations, helping trainees to improve their skills more effectively.

In the broader context of HRI, the advancements demonstrated by the BiLSTM model with evolutionary hyperparameter tuning can be applied to other domains requiring precise and real-time gesture recognition. For instance, in collaborative robotics, the ability to accurately interpret human gestures can enable robots to work more seamlessly alongside humans, enhancing productivity and safety in industrial settings. Similarly, in healthcare beyond surgery, such models can assist in rehabilitation, where robots need to understand and respond to patients' movements.

In conclusion, the improvements in accuracy, model size, and prediction time achieved by integrating BiLSTM networks with evolutionary hyperparameter tuning make this approach a viable candidate for practical applications in automated surgical skill assessment and real-time gesture recognition. The results of this study emphasize the critical role of hyperparameter optimization in developing high-performing deep learning models for sequential data analysis. Future work can extend this approach to other viable datasets and explore the combination of different deep learning architectures and optimization techniques to further advance the field of robotic surgery, HRI, and machine learning. The potential applications of this technology are vast, offering significant benefits across various domains where precise and reliable gesture recognition is essential.

Chapter 8

Conclusion

In this thesis, we have explored the development and application of a bidirectional Long Short-Term Memory (BiLSTM) model optimized through evolutionary computation for the task of surgical gesture recognition. This study was grounded in the JIGSAWS dataset, which provided a comprehensive collection of kinematic and video data from robotic surgical systems. Our approach aimed to address the inherent challenges in recognizing complex temporal patterns in surgical gestures, enhancing both the accuracy and efficiency of such recognition systems.

8.1 Summary of Findings

The preprocessing of kinematic data, including segmentation, trimming, and normalization, set the stage for effective model training. The BiLSTM model architecture, characterized by its bidirectional processing capability, proved adept at capturing temporal dependencies within the kinematic data. The integration of evolutionary

computation for hyperparameter tuning further refined the model, leading to significant improvements in both training and validation performance. The optimized BiLSTM model demonstrated superior performance compared to a standard LSTM model, as evidenced by smoother convergence of training and validation loss curves, higher accuracy, and better generalization to unseen data.

8.2 Key Contributions

Enhanced Model Performance The BiLSTM model with evolutionary hyperparameter tuning consistently outperformed the standard LSTM model, highlighting the importance of sophisticated hyperparameter optimization techniques in developing high-performing deep learning models.

Efficient Hyperparameter Tuning The use of evolutionary algorithms allowed for an efficient exploration of the hyperparameter space, identifying optimal configurations that balanced model complexity, training efficiency, and generalization performance.

Real-world Applicability The significantly reduced model size and faster prediction times of the optimized BiLSTM model make it a viable candidate for deployment in real-world applications, particularly in resource-constrained environments such as portable surgical training devices and robotic surgical systems.

Advancements in Human-Robot Interaction (HRI) The ability of the BiLSTM model to accurately recognize surgical gestures has potential applications beyond surgical training, including enhancing the capabilities of robotic assistants in

surgical environments and improving collaborative robotics in industrial settings.

8.3 Implications for Human-AI Partnership

The integration of advanced AI models like BiLSTM with evolutionary hyperparameter tuning in the medical field exemplifies the potential of human-AI partnerships. These models can act as intelligent assistants, providing surgeons with real-time feedback and decision support, thereby enhancing the precision and efficiency of surgical procedures. This partnership can reduce the cognitive load on surgeons, allowing them to focus on more complex aspects of the surgery while relying on AI for accurate gesture recognition and predictive analytics. Such systems can also be instrumental in the training of surgeons, offering consistent and objective evaluations of surgical performance and helping trainees to refine their skills more effectively. This collaborative approach can significantly enhance the learning curve for new surgeons, ensuring a higher standard of surgical proficiency and ultimately leading to better patient outcomes.

8.4 Implications for Surgical Training and HRI

The findings of this study have substantial implications for the field of surgical training and human-robot interaction. The ability to provide real-time feedback on surgical performance can revolutionize surgical education, offering objective and consistent evaluations that help trainees refine their skills more effectively. In the context of HRI, the enhanced gesture recognition capabilities of the BiLSTM model can lead to more intuitive and responsive robotic systems, improving the collaboration between

humans and robots in various applications.

8.5 Future Directions

This thesis lays the groundwork for future research in several promising directions. Future work could explore the application of the BiLSTM model to other datasets and surgical tasks, further validating its robustness and versatility. Additionally, combining BiLSTM models with other advanced deep learning architectures, such as attention mechanisms and transformer models, could yield even greater improvements in performance. Expanding the scope of evolutionary computation techniques to include more sophisticated optimization strategies may also enhance the efficiency and effectiveness of hyperparameter tuning.

Furthermore, integrating multimodal data, such as combining kinematic and video data more seamlessly, could provide richer contextual information, further improving gesture recognition accuracy. Investigating the transferability of these models to other domains, such as rehabilitation and industrial robotics, would also be valuable, broadening the impact of this research.

8.6 Future Implementation of a Surgeon Training System

In practice, a future surgical training system powered by AI and machine learning could offer a highly interactive and immersive environment. The system would incorporate real-time feedback mechanisms using gesture recognition models like the

BiLSTM developed in this thesis. Trainee surgeons would perform various surgical tasks, while the system captures both their kinematic data and visual data in real-time.

This data would be fed into the model, which has been trained to recognize specific gestures and evaluate them based on a predefined skill set. The system could provide immediate feedback on the trainee's technique, highlighting any incorrect movements or inefficiencies and offering suggestions for improvement.

Moreover, the system could include a virtual or augmented reality interface, where users practice in simulated surgical environments. The training data would then be used not only to evaluate their performance but also to adapt and personalize training modules according to the individual's progress. This personalized feedback loop would continuously refine the surgeon's skills, offering targeted advice and exercises to improve weak areas.

The ultimate goal of such a system would be to create a seamless Human-AI partnership where the AI assistant enhances the surgeon's learning process, offering consistent, objective evaluations and ensuring that skill acquisition is not only efficient but also safe.

8.7 Final Thoughts

In conclusion, this thesis demonstrates the efficacy of BiLSTM models optimized with evolutionary computation for the task of surgical gesture recognition. The improvements in accuracy, model size, and prediction time achieved through this approach underscore the importance of advanced hyperparameter optimization techniques in deep learning. The potential applications of this technology extend beyond surgical

training, offering significant benefits across various domains where precise and reliable gesture recognition is essential. As we continue to advance the field of machine learning and human-robot interaction, the insights gained from this study will contribute to the development of more intelligent and responsive robotic systems, ultimately enhancing the synergy between humans and machines.

Bibliography

[1] O. Asan, A. E. Bayrak, and A. Choudhury. Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, 22(6) (no pagination), 2020. ISSN 1438-8871 (electronic) 1438-8871. doi: https://dx.doi.org/10.2196/15154. Embase.

[2] C. Bartneck, T. Belpaeme, F. Eyssel, T. Kanda, M. Keijsers, and S. Sabanovi´c. *Human–Robot Interaction: An Introduction*. 2019.

[3] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.

[4] K. Catchpole, A. Bisantz, M. S. Hallbeck, M. Weigl, R. Randell, M. Kossack, and J. T. Anger. Human factors in robotic assisted surgery: Lessons from studies 'in the wild'. *Applied Ergonomics*, 78:270–276, 2019. ISSN 0003-6870 1872-9126. doi: https://dx.doi.org/10.1016/j.apergo.2018.02.011. Embase.

[5] J. Cha, J. Kim, and S. Kim. Hands-free user interface for ar/vr devices exploiting wearer's facial gestures using unsupervised deep learning. *Sensors (Basel, Switzerland)*, 19(20), 2019. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s19204441.

[6] L. Chen, J. Fu, Y. Wu, H. Li, and B. Zheng. Hand gesture recognition using compact cnn via surface electromyography signals. *Sensors (Basel, Switzerland)*, 20(3), 2020. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s20030672.

[7] X. Chen, Y. Jiang, and C. Yang. Stiffness estimation and intention detection for human-robot collaboration. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1802–1807. ISBN 2158-2297. doi: 10.1109/ICIEA48937.2020.9248186.

[8] S. Y. Chiu, S. Y. Chiu, Y. J. Tu, and C. I. Hsu. Gesture-based intention prediction for automatic door opening using low-resolution thermal sensors: A u-net-based deep learning approach. In *2021 IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pages 271–274. doi: 10.1109/ECICE52819.2021.9645718.

[9] A. Choudhury, O. Asan, and J. E. Medow. Effect of risk, expectancy, and trust on clinicians' intent to use an artificial intelligence system - blood utilization calculator. *Applied Ergonomics*, 101:103708, 2022. ISSN 0003-6870 1872-9126. doi: https://dx.doi.org/10.1016/j.apergo.2022.103708.

[10] J. Coker, H. Chen, J. Schall, Mark C., S. Gallagher, and M. Zabala. Emg and joint angle-based machine learning to predict future joint angles at the knee. *Sensors (Basel, Switzerland)*, 21(11), 2021. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s21113622.

[11] U. Cote-Allard, G. Gagnon-Turcotte, F. Laviolette, and B. Gosselin. A low-cost, wireless, 3-d-printed custom armband for semg hand gesture recognition. *Sensors*

(Basel, Switzerland), 19(12), 2019. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s19122811.

[12] J. Ding, Ing and N.-W. Zheng. Cnn deep learning with wavelet image fusion of ccd rgb-ir and depth-grayscale sensor data for hand gesture intention recognition. *Sensors (Basel, Switzerland)*, 22(3), 2022. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s22030803.

[13] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager. Recognizing surgical activities with recurrent neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19*, pages 551–558. Springer. ISBN 3319467190.

[14] P. Esmaeilzadeh, T. Mirzaei, and S. Dharanikota. Patients' perceptions toward human-artificial intelligence interaction in health care: Experimental study. *Journal of Medical Internet Research*, 23(11):e25856, 2021. ISSN 1438-8871 (electronic) 1438-8871. doi: https://dx.doi.org/10.2196/25856.

[15] Z. Fang, D. Vazquez, and A. M. Lopez. On-board detection of pedestrian intentions. *Sensors (Basel, Switzerland)*, 17(10), 2017. ISSN 1424-8220 (electronic) 1424-8220. doi: https://dx.doi.org/10.3390/s17102193.

[16] A. G. Feleke, L. Bi, and W. Fei. Emg-based 3d hand motor intention prediction for information transfer from human to robot. *Sensors (Basel, Switzerland)*, 21 (4), 2021. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s21041316.

[17] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin,

L. Tao, L. Zappella, B. Béjar, and D. D. Yuh. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, page 3.

[18] M. Gardner, C. S. Mancero Castillo, S. Wilson, D. Farina, E. Burdet, B. C. Khoo, S. F. Atashzar, and R. Vaidyanathan. A multimodal intention detection sensor suite for shared autonomy of upper-limb robotic prostheses. *Sensors (Basel, Switzerland)*, 20(21), 2020. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s20216097.

[19] M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer. Intentions of vulnerable road users—detection and forecasting by means of machine learning. *IEEE Transactions on Intelligent Transportation Systems*, 21(7): 3035–3045, 2020. ISSN 1558-0016. doi: 10.1109/TITS.2019.2923319.

[20] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 38, 03 2013. doi: 10.1109/ICASSP.2013.6638947.

[21] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997. doi: 10.1162/neco.1997.9.8.1735.

[22] P. Huang, H. Wang, Y. Wang, Z. Liu, O. W. Samuel, M. Yu, X. Li, S. Chen, and G. Li. Identification of upper-limb movements based on muscle shape change signals for human-robot interaction. *Computational and mathematical methods in medicine*, 2020:5694265, 2020. ISSN 1748-6718 1748-670X. doi: https://dx.doi.org/10.1155/2020/5694265.

[23] I. E. C. (IEC). Medical electrical equipment – part 4-1: Guidance and interpretation – medical electrical equipment and medical electrical systems employing a degree of autonomy, 2017.

[24] N. Jaouedi, F. J. Perales, J. M. Buades, N. Boujnah, and M. S. Bouhlel. Prediction of human activities based on a new structure of skeleton features and deep learning model. *Sensors (Basel, Switzerland)*, 20(17), 2020. ISSN 1424-8220 (electronic) 1424-8220. doi: https://dx.doi.org/10.3390/s20174944.

[25] T. Kagawa, T. Kato, and Y. Uno. On-line control of continuous walking of wearable robot coordinating with user's voluntary motion. IEEE. doi: 10.1109/iros.2015.7354128.

[26] J. F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, jan 1984. ISSN 1046-8188. doi: 10.1145/357417.357420.

[27] E. Kilic and E. Dogan. Design and fuzzy logic control of an active wrist orthosis. *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine*, 231(8):728–746, 2017. ISSN 2041-3033 0954-4119. doi: https://dx.doi.org/10.1177/0954411917705408.

[28] J. V. Kopke, M. D. Ellis, and L. J. Hargrove. Determining user intent of partly dynamic shoulder tasks in individuals with chronic stroke using pattern recognition. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, 28(1):350–358, 2020. ISSN 1558-0210 1534-4320. doi: https://dx.doi.org/10.1109/TNSRE.2019.2955029.

[29] N. Kumar and K. P. Michmizos. Deep learning of movement intent and reaction time for eeg-informed adaptation of rehabilitation robots. In *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, pages 527–532. ISBN 2155-1782. doi: 10.1109/BioRob49111.2020.9224272.

[30] J. Lanini, H. Razavi, J. Urain, and A. Ijspeert. Human intention detection as a multiclass classification problem: Application in physical human–robot interaction while walking. *IEEE Robotics and Automation Letters*, 3(4):4171–4178, 2018. ISSN 2377-3766. doi: 10.1109/LRA.2018.2864351.

[31] G. Li, Z. Liu, L. Cai, and J. Yan. Standing-posture recognition in human-robot collaboration based on deep learning and the dempster-shafer evidence theory. *Sensors (Basel, Switzerland)*, 20(4), 2020. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s20041158.

[32] K. Li, J. Wu, X. Zhao, and M. Tan. Using gaze patterns to infer human intention for human-robot interaction. In *2018 13th World Congress on Intelligent Control and Automation (WCICA)*, pages 933–938. doi: 10.1109/WCICA.2018.8630571.

[33] Y. Li and Y. Xie. Is a picture worth a thousand words? an empirical study of image content and social media engagement. *Journal of Marketing Research*, 57(1):1–19, 2019. ISSN 0022-2437. doi: 10.1177/0022243719881113. doi: 10.1177/0022243719881113.

[34] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C, and P. Dollár. Microsoft coco: Common objects in context. *arXiv pre-print server*, 2015. doi: Nonearxiv:1405.0312.

[35] L. K. Liu, T. C. Chien, Y. L. Chen, L. C. Fu, and J. S. Lai. Sensorless control with friction and human intention estimation of exoskeleton robot for upper-limb rehabilitation. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 290–296, . doi: 10.1109/ROBIO49542.2019.8961760.

[36] Z. Liu, J. Yang, W. Yina, and X. Fu. Novel walking-intention recognition method for omnidirectional walking support robot. In *2017 International Conference on Computer Technology, Electronics and Communication (ICCTEC)*, pages 1048–1052, . doi: 10.1109/ICCTEC.2017.00230.

[37] Z. Liu, Q. Liu, W. Xu, Z. Liu, Z. Zhou, and J. Chen. Deep learning-based human motion prediction considering context awareness for human-robot collaboration in manufacturing. *Procedia CIRP*, 83:272–278, 2019. ISSN 2212-8271. doi: https://doi.org/10.1016/j.procir.2019.04.080.

[38] W. Lu, Z. Hu, and J. Pan. Human-robot collaboration using variable admittance control and human intention prediction. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pages 1116–1121. ISBN 2161-8089. doi: 10.1109/CASE48305.2020.9217040.

[39] Y. Massalin, M. Abdrakhmanova, and H. A. Varol. User-independent intent recognition for lower limb prostheses using depth sensing. *IEEE transactions on bio-medical engineering*, 65(8):1759–1770, 2018. ISSN 1558-2531 0018-9294. doi: https://dx.doi.org/10.1109/TBME.2017.2776157.

[40] G. Menegozzo, D. Dallralba, C. Zandona, and P. Fiorini. Surgical gesture recognition with time delay neural network based on kinematic data. IEEE. doi: 10.1109/ismr.2019.8710178.

[41] F. Mohammadi Amin, M. Rezayati, H. W. van de Venn, and H. Karimpour. A mixed-perception approach for safe human-robot collaboration in industrial automation. *Sensors (Basel, Switzerland)*, 20(21), 2020. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s20216347.

[42] D.-H. Moon, D. Kim, and Y.-D. Hong. Development of a single leg knee exoskeleton and sensing knee center of rotation change for intention detection. *Sensors (Basel, Switzerland)*, 19(18), 2019. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s19183960.

[43] J. Owoyemi and K. Hashimoto. Learning human motion intention with 3d convolutional neural network. In *2017 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1810–1815. ISBN 2152-744X. doi: 10.1109/ICMA.2017.8016092.

[44] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71.

[45] A. Poulose, J. H. Kim, and D. S. Han. Hit har: Human image threshing machine for human activity recognition using deep learning models. *Computational intelligence and neuroscience*, 2022:1808990, 2022. ISSN 1687-5273. doi: https://dx.doi.org/10.1155/2022/1808990.

[46] H. Ren, D. X. Liu, N. Li, Y. He, Z. Yan, and X. Wu. On-line dynamic gait generation model for wearable robot with user's motion intention. In *2018 IEEE International Conference on Information and Automation (ICIA)*, pages 347–352. doi: 10.1109/ICInfA.2018.8812568.

[47] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: the rprop algorithm. IEEE, 2002. doi: 10.1109/icnn.1993.298623.

[48] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. doi: 10.1109/78.650093.

[49] B.-Y. Su, J. Wang, S.-Q. Liu, M. Sheng, J. Jiang, and K. Xiang. A cnn-based method for intent recognition using inertial measurement units and intelligent lower limb prosthesis. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, 27(5):1032–1042, 2019. ISSN 1558-0210 1534-4320. doi: https://dx.doi.org/10.1109/TNSRE.2019.2909585.

[50] J. Torrent-Sellens, A. I. Jimenez-Zarco, and F. Saigi-Rubio. Do people trust in robot-assisted surgery? evidence from europe. *International Journal of Environmental Research and Public Health*, 18(23) (no pagination), 2021. ISSN 1661-7827 1660-4601. doi: https://dx.doi.org/10.3390/ijerph182312519. Embase.

[51] A. C. Tsitos, M. Dagioglou, and T. Giannakopoulos. Real-time feasibility of a human intention method evaluated through a competitive human-robot reaching game. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1080–1084. doi: 10.1109/HRI53351.2022.9889601.

[52] V. K. Viekash, P. Arun, S. Manimozhi, G. D. Nagotanekar, and E. Deenadayalan. Deep learning based muscle intent classification in continuous passive motion machine for knee osteoarthritis rehabilitation. In *2021 IEEE Madras Section Conference (MASCON)*, pages 1–8. doi: 10.1109/MASCON51689.2021.9563370.

[53] Q. Wang. Research on the improved cnn deep learning method for motion intention recognition of dynamic lower limb prosthesis. *Journal of healthcare engineering*, 2021:7331692, 2021. ISSN 2040-2309 2040-2295. doi: https://dx.doi.org/10.1155/2021/7331692.

[54] M. Wen and Y. Wang. Multimodal sensor motion intention recognition based on three-dimensional convolutional neural network algorithm. *Computational intelligence and neuroscience*, 2021:5690868, 2021. ISSN 1687-5273. doi: https://dx.doi.org/10.1155/2021/5690868.

[55] A. Wendemuth, R. Bock, A. Nurnberger, A. Al-Hamadi, A. Brechmann, and F. W. Ohl. Intention-based anticipatory interactive systems, 2018.

[56] B. Xiao, C. Chen, and X. Yin. Recent advancements of robotics in construction. *Automation in Construction*, 144, 2022. ISSN 09265805. doi: 10.1016/j.autcon.2022.104591.

[57] X. Yao. Evolving artificial neural networks. *Proceedings of the IEEE*, 87(9): 1423–1447, 1999. doi: 10.1109/5.784219.

[58] S. Young, B. Stephens-Fripp, A. Gillett, H. Zhou, and G. Alici. Pattern recognition for prosthetic hand user's intentions using emg data and machine

learning techniques. In *2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, pages 544–550. ISBN 2159-6255. doi: 10.1109/AIM.2019.8868766.

[59] Z. Zhang, G. Peng, W. Wang, Y. Chen, Y. Jia, and S. Liu. Prediction-based human-robot collaboration in assembly tasks using a learning from demonstration model. *Sensors (Basel, Switzerland)*, 22(11), 2022. ISSN 1424-8220. doi: https://dx.doi.org/10.3390/s22114279.