Semi-Supervised Learning for Automatic Recognition of Human Head Gestures

Rijul Muhammed

Master of Science

School of Informatics

University of Edinburgh

2024

Abstract

This project explores semi-supervised learning techniques for automatic classification of human head gestures using motion capture data. It investigates self-training, ensemble methods, and mean teacher approaches to leverage unlabelled data and improve classification performance on the University of Edinburgh Speaker Personality and MoCap Dataset. The proposed novel ensemble self-training approach, combining Bidirectional LSTM and 1D CNN models, achieved a test F1-score of 0.56 using only 40% labelled data, comparable to fully supervised models with a test F1-score of 0.58 trained on 100% labelled data. Speaker-independent cross-validation demonstrated promising generalisation, with an average test accuracy of 66.4% and test F1-score of 0.53 across different speakers. In contrast, a fully supervised Bi-LSTM model trained on the same 40% labelled data subset achieved a significantly lower test F1-score of 0.33, highlighting the effectiveness of the semi-supervised approach.

This research advances head gesture recognition by incorporating semi-supervised techniques and demonstrating its effectiveness in reducing annotation efforts while maintaining high accuracy. The report provides insights into the trade-offs between labelling effort and model performance, and highlight some challenges in distinguishing subtle gestures. These findings have important implications for developing more efficient and accurate gesture recognition systems for human-computer interaction applications.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Rijul Muhammed*)

Acknowledgements

I would like to express my deepest gratitude to Prof. Hiroshi Shimodaira for his exceptional guidance, insightful feedback, and unwavering support throughout this project. His expertise and encouragement were instrumental in the completion of this research.

I am profoundly thankful to my family for their unconditional support and understanding during my studies. Their belief in me has been a constant source of motivation.

My sincere appreciation goes to my friends, whose moral support and stimulating discussions have been invaluable. Finally, I extend my thanks to the previous researchers whose work laid the foundation for this study.

Table of Contents

Chapter 1

Introduction

1.1 Motivation and Problem Statement

Head gestures are an integral component of non-verbal communication, facilitating the expression of acts such as agreement, disagreement, and attention. The recognition of these gestures has emerged as a significant area of research in the field of Human-Computer Interaction due to its potential applications across several domains, such as assistive technologies (Jiang et al., 2013), virtual reality (Rautaray and Agrawal, 2015), and human behaviour analysis (Kotsiantis et al., 2007). However, there are several challenges associated with head gesture recognition that need to be addressed.

The primary difficulty lies in the subtle nature of head gestures, which are often less pronounced than other body movements and exhibit high variability across individuals (Murphy-Chutorian and Trivedi, 2009). This subtlety, combined with the continuous nature of head movements, makes the task of gesture classification particularly complex. Traditional approaches to this problem such as Hidden Markov Models (Rabiner, 1986) and Dynamic Time Warping (Hachaj and Piekarczyk, 2019) have shown only limited success, primarily in constrained environments. These methods rely heavily on large annotated datasets, which are both time-consuming and expensive to create. Recent advancements in deep learning, particularly Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks, have enhanced gesture recognition capabilities by capturing temporal dependencies from the data (Neverova et al., 2014). While deep learning models demonstrate improved performance and better generalisation, they still face the significant bottleneck of the need for large amounts of annotated data.

Semi-supervised learning emerges as a potential solution to these challenges (Zhu

and Goldberg, 2009). By leveraging both labelled and unlabelled data, semi-supervised approaches offer the prospect of building more robust models without the prohibitive costs associated with extensive data annotation. Moreover, semi-supervised methods have the potential to continuously adapt and improve as new, unlabelled data becomes available. However, the application of semi-supervised learning to the specific domain of head gesture recognition remains largely unexplored. Hence, there is a clear need to investigate how these techniques can be effectively adapted to handle the unique challenges posed by head gesture data, such as the continuous nature of movements and the subtle distinctions between different gesture types. This research seeks to address this gap by investigating the effectiveness of semi-supervised learning techniques in the context of head gesture recognition by utilising the time series data obtained from The University of Edinburgh's speaker personality and MoCap dataset (Haag and Shimodaira, 2015).

1.2 Research Objectives and Hypothesis

The aim of this research is to investigate and develop semi-supervised learning techniques for automatic classification of human head gestures using motion capture data. This study seeks to address the challenges of limited labelled data in head gesture recognition while maintaining high classification accuracy. The specific objectives of this research are:

1. To develop and implement semi-supervised learning models, specifically Self-Training, Mean-Teacher and a novel ensemble Self-Training approach, for continuous head gesture classification using time series data obtained from the MoCap dataset.

2. To evaluate and compare the performance of the proposed semi-supervised learning techniques against each other and with supervised learning approaches.

3. To evaluate the performance of the semi-supervised models for varied proportions of labelled and unlabelled data.

4. To investigate the variation in model performance across different types of head gestures and identify any gesture-specific challenges or patterns.

5. To analyse the generalisability of the model by implementing a speaker-based cross validation strategy

The primary hypothesis of this research is that utilising unlabelled data through the proposed semi-supervised learning techniques will lead to better performance in head gesture classification compared to models that use only a limited amount of labelled data. Furthermore, it is hypothesised that the performance of semi-supervised models will approach that of fully supervised models trained on completely labelled datasets, while significantly reducing the required annotation effort.

1.3 Contributions

This research makes significant contributions to the field of head gesture recognition, focusing on the application of semi-supervised learning techniques. A key contribution is the development of semi-supervised learning pipelines specifically tailored for head gesture classification, including a unique ensemble semi-supervised learning approach. The research also uses a generative technique for data augmentation instead of traditional methods like adding noise that was used in previous works. Experiments involving different proportions of labelled data is done to provide insights related to model performance and annotation costs. Through detailed gesture-wise performance analysis, the study identifies and highlights gesture-specific challenges, contributing to a deeper understanding of head gesture recognition. Collectively, these contributions offer both theoretical insights and practical methodologies, paving the way for effective development of gesture recognition systems in the future.

1.4 Thesis Structure

This dissertation is structured as follows. Chapter 2 reviews relevant literature related to the research. Chapter 3 describes the MoCap dataset, including data representation, gesture types, and annotation process. Chapter 4 details the methodology, covering data analysis, preprocessing, augmentation techniques, baseline model development, and proposed semi-supervised learning approaches. Chapter 5 presents and discusses the results, comparing baseline and semi-supervised model performances, analysing different proportions of labelled data, analysing gesture-wise outcomes, and evaluating generalisation through speaker-independent cross-validation. Finally, Chapter 6 concludes the dissertation, summarising key findings and suggesting future research directions.

Chapter 2

Background and Related Work

2.1 Head Gestures in Human-Computer Interaction

Head gestures are a sophisticated form of non-verbal communication that convey attention, emotions, and cognitive processes, making them valuable in HCI, especially for users with limited mobility or in hands-free scenarios (Wagner et al., 2014). Unlike hand and body gestures, head gestures require minimal physical effort, benefiting individuals with severe motor impairments by enabling communication through assistive devices, thereby enhancing their independence (Terven et al., 2014).

The applications of head gesture recognition in HCI are diverse and impactful. In virtual and augmented reality, head gestures facilitate interaction with digital objects, such as using a head tilt to rotate a 3D model or a nod to confirm a selection (Zhao and Allison, 2017). In automotive safety, these systems monitor driver behaviour, detecting fatigue or distraction through head movements to trigger alerts (Choi and Kim, 2014). These applications show how head gesture recognition contributes to more intuitive, safer, and engaging interactive systems across various domains.

However, implementing robust head gesture recognition systems is challenging due to its subtle and complex nature. Moreover, there are variations in how individuals perform gestures in terms of duration and intensity. Nevertheless, advancements in machine learning has made the development of robust head gesture recognition systems more viable.

2.2 Machine Learning Approaches for Gesture Recognition

The evolution of machine learning techniques has advanced the field of gesture recognition, providing improved performance and generalisation in comparison to traditional methods. Early machine learning approaches often employed Hidden Markov Models (HMMs) (Rabiner, 1986), which were well-suited to modelling temporal sequences. However, their performance was limited by reliance on hand-crafted features and inability to model long-range dependencies effectively.

Gesture recognition tasks that use time series data such as motion capture data from sensors has been improved with deep learning. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have proven highly effective for capturing temporal dynamics in sequential data without the need for hand-crafted features (Hochreiter and Schmidhuber, 1997). LSTMs can learn long-range dependencies, making them well-suited to modelling complex temporal patterns in gesture sequences. One-dimensional Convolutional Neural Networks (1D CNNs) have also shown promise in time series gesture recognition. Unlike their 2D counterparts used in image recognition, 1D CNNs are designed to process sequential data directly. They excel at extracting local patterns and features from time series data, making them effective for gesture recognition tasks (Yang et al., 2019b).

In previous research related to gesture recognition, LSTMs have achieved an accuracy of 87% on signal data (Toro-Ossaba et al., 2022). Similarly, 1D CNNs have achieved an accuracy of 96% for hand gesture classification using EEG signals (Miah et al., 2022). When it comes to performance for head gesture classification using the university MoCap dataset, previous students have achieved accuracies of 52.8% and 60.4% for 1D CNN and Bi-LSTM models respectively (Chen, 2023). Despite these advancements, the need for large labelled datasets remains a bottleneck and researchers are exploring the viability of semi-supervised learning approaches to solve this challenge.

2.3 Semi-Supervised Learning in Time Series Classification

Semi-supervised learning (SSL) techniques that leverage unlabelled data to improve model performance are particularly relevant for time series classification tasks where

labelled data is scarce. Methods such as self-training, co-training, and mean-teacher models have proven effective in several domains.

Self-training has emerged as an effective semi-supervised learning technique for time series classification tasks like gesture recognition, where labelled data is often limited. For a cross-user gesture recognition task based on surface electromyography (sEMG), an iterative self-training method was used (Wang et al., 2023). This method iteratively trains on labelled data and assigns pseudo-labels to unlabelled data based on confidence level, updating the pseudo-labels after each iteration. To address class imbalance, they employed oversampling of minority classes. Experiments on multiple sEMG datasets demonstrated that self-training outperformed baseline and state-of-the-art methods, achieving over 25% improvement versus baselines and over 5% versus supervised domain adaptation approaches.

Co-training is another method that has shown promise for hand posture recognition tasks with limited labelled data. This method leverages two different feature representations to train separate classifiers that improve each other iteratively using unlabelled data (Fang et al., 2008). The method trains initial classifiers on a small labelled dataset, then has each classifier confidently label unlabelled examples to augment the other classifier's training set. Experiments on the Triesch hand posture dataset demonstrated the co-training approach improved accuracy by 5-9% for challenging postures compared to single classifiers, while using much less labelled data than previous approaches. The classifier achieved 90.1% average accuracy, outperforming prior methods.

Mean-teacher model is an interesting method that has shown promise for semi-supervised learning in hand gesture recognition tasks using radar data (Shi et al., 2024). It comprises of a student and teacher model. The student model is trained directly on labelled data, while the teacher model is updated using an exponential moving average (EMA) of the student model weights. This allows the teacher to produce more stable pseudo-labels for unlabelled data. The model enforces consistency between teacher and student predictions on augmented unlabelled samples, helping mitigate effects of individual differences and noise. Experiments on two public datasets demonstrated the effectiveness of this approach, achieving over 99% accuracy on both the Soli and Air-Writing datasets. The mean-teacher model outperformed fully-supervised baselines while leveraging unlabelled data.

The above-mentioned strategies used for gesture recognition provide a solid foundation for applying semi-supervised learning techniques specifically to the unexplored domain of head gestures that are subtle and complex in nature.

Chapter 3

Dataset

3.1 MoCap Dataset Overview

This study utilises the University of Edinburgh Speaker Personality and MoCap Dataset (Haag and Shimodaira, 2015), a comprehensive collection of motion capture data, video recordings, and audio files designed for research in human-computer interaction and gesture recognition. The dataset comprises recordings from 13 native English speakers (7 male, 6 female) engaged in conversational interactions.

A unique aspect of this dataset is its incorporation of personality variation. All 13 speakers initially scored high on extroversion in the Big Five personality tests. However, during the recordings, each speaker was instructed to exhibit three distinct personality types across different conversations: introverted, extroverted, and neutral. The dataset includes a total of 130 video recordings, with each conversation lasting approximately 5 minutes. The MoCap dataset contains two primary components of Motion capture data (ROV files) and Gesture annotations (ELAN files) essential for the research. In addition to these primary components, the dataset includes synchronised video and audio recordings. The video recordings are crucial for the manual annotation process, allowing annotators to visually identify and label head gestures. The audio recordings are useful in calculating the time lag between the videos and motion capture data by noting the time of specific ”beep” sounds from these files.

3.2 Motion Capture Data Representation

The motion capture data is recorded using the Natural Point OptiTrack system (Natural Point Inc., 2022), which captures head movements with high precision. This data

is stored in Rotation Vector (ROV) format, representing the head's orientation and movement in three-dimensional space. The ROV format encodes head motion using six degrees of freedom (6 DoF), which represent the independent ways an object can move in three-dimensional space. These six parameters are divided into the categories of rotation and translation.

The rotation parameters, denoted as RV_x, RV_y, and RV_z, represent a rotation vector in three-dimensional Euclidean space. This rotation vector combines both the axis of rotation and the angle of rotation into a single entity. The direction of the vector (RV_x, RV_y, RV_z) indicates the axis around which the rotation occurs, while the magnitude of the vector represents the angle of rotation in radians. This representation, known as the axis-angle representation or Euler vector, provides a concise way to describe 3D rotations (Wikipedia, 2024). The translation parameters, denoted as Tx, Ty, and Tz, represent movement along the x, y, and z axes respectively.

These six parameters provide a comprehensive description of head position and orientation at each time point. The data is sampled at a rate of 100 Hz using V100:R2 cameras, resulting in 100 frames per second, with each frame captured every 10ms. Table 3.1 shows a sample excerpt from a ROV file, which has been converted from the original .rov format to a .csv file by previous students (Yang, 2022). It is important to note that the rotational vector data was present in a normalised format to address the issue of different speakers being aligned differently during the recording sessions, enhancing the robustness and generalisability of the gesture recognition models. Furthermore, the rotational vector data has been time aligned with the video annotation data by previous students (Chen, 2023).

Frame	RVx	RVy	RVz	Tx	Ty	Tz
1	0.000476352	-0.000507661	0.001231	0.000685281	0.000496295	-8.88989e-05
2	0.000457434	-0.00038713	0.00128097	0.00107964	0.000750685	-0.000166619
3	0.000608655	-0.000576298	0.00230161	0.00156778	0.00102971	-0.000123638

Table 3.1: Sample Excerpt from a ROV File

3.3 Gesture Annotations Data Representation

The head gestures in the dataset are manually annotated using the ELAN (EUDICO Linguistic Annotator) software, which is a professional tool designed for the creation of complex annotations on video and audio resources. (Wittenburg et al., 2006). The

ELAN files contain the annotation information done by previous students who worked on this research (Wang, 2023). They have selected 8 out of the 12 speakers available and annotated 3 recordings for each of them. The speakers are Adam, Brian, Beve, Dani, Esmo, Ella, Paul, and Sophie. Table 3.2 presents a sample of the data contained in an ELAN annotation file. The type column mentions the gesture type. The speaking column represents whether the speaker was speaking(1) or not(0) during that period. Start and end time represent the starting and ending period of the gesture in milliseconds and the duration column is just the difference between the start and end times.

Type	Speaking	Start Time (ms)	End Time (ms)	Duration (ms)
nd	1	22040	23110	1070
ti	1	36040	37990	1950
nd	0	42070	43590	1520

Table 3.2: Sample Data From an ELAN Annotation File

The dataset includes annotations for seven distinct types of head gestures, each representing a common non-verbal communication cue. Table 3.3 provides an overview of these gesture types:

Label	Gesture Type	Description
nd	Nod	Vertical up-down movement
sh	Shake	Horizontal side-to-side movement
ti	Tilt	Inclining the head to either side
fu	Face Up	Upward movement of the face
fd	Face Down	Downward movement of the face
t	Turn	Rotating the head left or right
mnd	Multiple Nods	Series of rapid vertical movements

Table 3.3: Head Gesture Types and Their Descriptions

While the MoCap dataset provides a rich source of data for head gesture recognition, it poses some challenges. Gestures take up only a small proportion of each of the recordings and the addition of no gesture class to enable continuous classification will lead to a large imbalance in the dataset. Also, there will be inter-annotator variability present in the data as different annotators perceive gestures differently and label them accordingly. Inter-subject variability can also be present as different subjects perform gestures with different intensities and it can go unnoticed by the annotator. These shortcoming have to be taken into account when developing the classification system.

Chapter 4

Methodology

4.1 Data Analysis

4.1.1 Gesture Distribution

The analysis of gesture label distribution across the annotated dataset provides insights into the frequency and prevalence of different head gestures in natural conversations. Figure 4.1 illustrates the frequency of each gesture type across the 24 annotated ELAN data files.

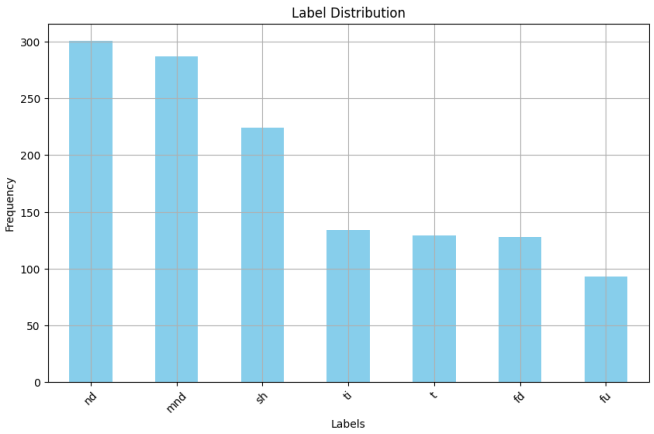


Figure 4.1: Gesture Distributions Across All Annotated Files

The label distribution reveals several key observations. Nodding (nd) and multiple nodding (mnd) gestures collectively represent the most frequent head movements in the dataset, aligning with their common use as non-verbal cues for agreement or acknowledgement in conversations. Head shaking (sh) also shows a substantial frequency, indicating its importance as a communicative gesture often used for disagreement or

negation. Gestures labelled as tilting (ti) and turning (t) appear with moderate frequency, which may be attributed to natural head movements during conversations that don't necessarily convey specific intentions. These could represent moments when a speaker is contemplating or briefly disengaged. Interestingly, face down (fd) and face up (fu) gestures are the least frequent in the dataset. This low frequency might be due to the nature of online communication, where participants tend to maintain eye contact, limiting vertical head movements. There is also the possibility of annotators classifying fu and fd gestures as nd as there is a significant overlap amongst these gestures.

The uneven distribution of gesture labels presents challenges for the classification task. The significant disparity between the most common (nd, mnd) and least common (fu, fd) gestures necessitates careful consideration in model development. To address this imbalance, techniques such as data augmentation or sampling techniques have to be implemented. These strategies will help prevent bias towards over-represented classes and ensure fair learning across all gesture types.

4.1.2 Feature Correlation Analysis

A correlation analysis was conducted on the six primary motion parameters: R_x, R_y, R_z (rotational vectors), and T_x, T_y, T_z (translational parameters). The analysis employed the Pearson correlation coefficient (Cohen et al., 2009), which measures the linear correlation between two variables. For variables X and Y (the features for which correlation is computed), the Pearson correlation coefficient r is computed as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4.1}$$

where $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively, and n is the number of observations. Figure 4.2 presents the averaged correlation matrix derived from all data files. The analysis reveals strong correlations between rotational and translational vectors:

- Strong positive correlation (0.86) between R_z and T_x
- High positive correlation (0.78) between R_z and T_y
- Strong negative correlation (-0.77) between R_x and T_z

These findings confirm the redundancy of translational vectors as mentioned in the previous students' papers (Lyu, 2023). Hence, the translational data can be omitted during model training without significant loss of information. This dimensionality

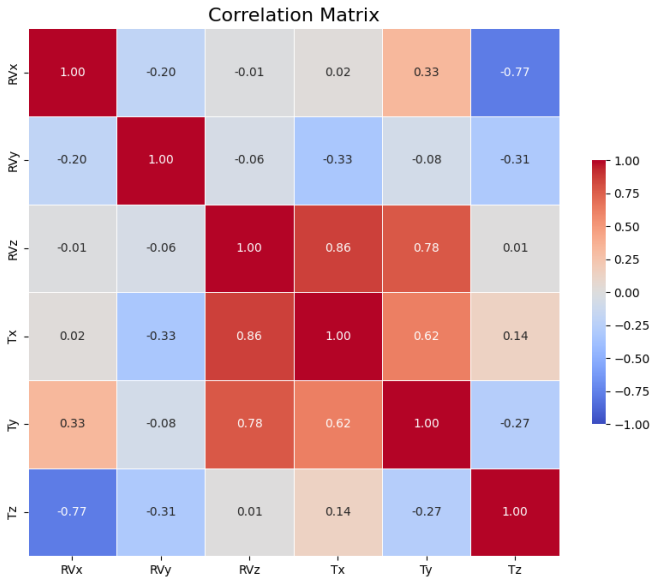


Figure 4.2: Correlation Matrix of Motion Features

reduction can play a significant role in reducing computational requirements during modelling.

4.1.3 Temporal Characteristic of Gestures

An analysis of the temporal characteristics of gestures reveals variations in duration across different annotations. Table 4.1 presents the average duration in frames for each gesture type, where each frame represents 10 milliseconds of motion capture data. The data indicates variation in the average duration of different gesture types:

Gesture Type	Average Duration (frames)
fd	197.53
fu	137.48
mnd	198.27
nd	95.62
sh	156.88
t	167.30
ti	116.94

Table 4.1: Gesture-Wise Average Duration

- Multiple nodding (mnd) and face down (fd) gestures exhibit the longest average

durations at 198.27 and 197.53 frames respectively.

- Single nods (nd) are typically the briefest, averaging 95.62 frames.
- Head shaking (sh) and turning (t) gestures show intermediate durations of 156.88 and 167.30 frames respectively.
- Tilting (ti) and face up (fu) gestures are relatively short, averaging 116.94 and 137.48 frames.

A key insight from this analysis is that the average gesture duration across the entire dataset is approximately 150 frames, or 1.5 seconds. This finding has important implications for the modelling process. Using 150 frames as the input size during modelling offers a balanced approach that can capture the majority of gesture patterns while remaining computationally practical.

4.1.4 Gesture-Specific Motion Patterns

Analysis of the rotation vector time series data revealed distinctive patterns associated with each gesture type, as illustrated in Figure 4.3. The flat lines present in the plots indicate padding which will be discussed in more detail in the data preprocessing section. Nodding and multiple nodding gestures have very similar patterns for the rotational vectors. Face up and face down gestures also displayed significant variations in RV_z, similar to the nodding gestures. The provided plots are just one sample and several such plots were analysed during the actual research to identify distinctive patterns. The patterns were most prevalent for multiple nodding, nodding, face down, and no gesture categories. The presence of distinctive patterns proved the potential for the rotational vectors to be used as discriminative features for classification.

4.2 Data Preprocessing

4.2.1 Gesture Label Consolidation

An initial analysis of the rotational vector patterns revealed significant similarities between the 'nodding' and 'multiple nodding' gestures. Moreover, these gestures convey similar non-verbal cues. Therefore, to simplify the classification task and improve class balance, the 'multiple nodding' gesture was reclassified as 'nodding'. This merging also helped to increase the total number of 'nodding' samples, partially

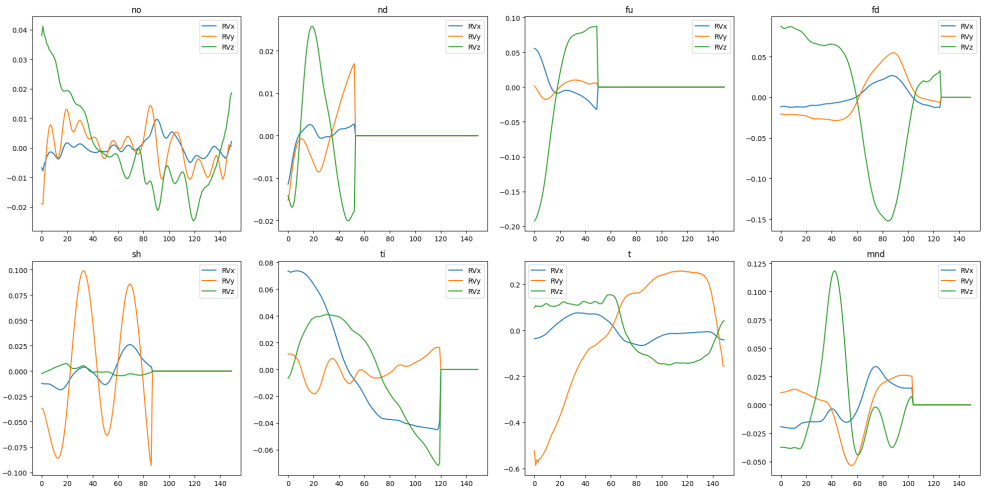


Figure 4.3: Time Series Plots of RV_x, RV_y, and RV_z for Different Gesture Types

addressing the class imbalance issue caused by the presence of a large number of non-gesture data.

4.2.2 Non-Gesture Annotation

The original ELAN annotation files only contained labels for observed gestures. However, for continuous classification, it is essential to identify periods of non-gesture. To accomplish this, all time periods in the ELAN files without specific gesture annotations were labelled as 'no' (no gesture).

4.2.3 Data Merging and Feature Selection

The rotational vector (ROV) data from the sensor files were merged with the annotated ELAN files based on frame numbers to create the final dataset required for classification. During this process, certain columns were excluded to optimise the dataset for modelling:

- The 'speaking' column, representing whether the speaker was talking or not, was removed due to the lack of speaking information for non-gesture data. Moreover, it does not seem to play a role in influencing the type of gesture as each gesture has a mix of annotations where the speaker is talking and silent.
- Translational vectors were excluded due to their high correlation with rotational vectors as seen earlier, reducing data redundancy.

4.2.4 Data Scaling

To ensure consistent feature ranges and improve model convergence, each data file was scaled to a range between -1 and 1. This transformation is defined by the following equation:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \cdot (max - min) + min \tag{4.2}$$

where:

- X is the original value
- X_{min} and X_{max} are the minimum and maximum values of the feature in the dataset
- min and max are the desired range of scaled values (-1 and 1 in our case)

For this project, the MinMaxScaler implementation from scikit-learn was utilized (Pedregosa et al., 2011a), which provides an efficient and reliable scaling method. Importantly, scaling was applied to entire CSV files rather than smaller chunks of data that will be passed as model input. This approach was chosen to prevent the amplification of noise values that could potentially affect classification accuracy.

4.2.5 Dataset Splitting

A speaker-independent split strategy was employed to maximize the model's generalizability. This approach ensures that the test set contains entirely new speakers, whose data the model has not encountered during training. This technique will help provide insights into how capable the model is in recognising gestures even when factors such as subject, duration and style of gestures vary. The split was as follows:

- Training set: Adam, Beve, Ella, Dani
- Validation set: Esmo, Brian
- Test set: Paul, Sophie

However, it is important to note that using only two speakers for the test set is not very representative of overall performance, as it may not capture the full variability across different speakers. To overcome this limitation, a cross-validation approach was implemented for the best performing model, which will be discussed in detail in the evaluation framework section.

For creating the unlabelled dataset, 21 files that have been annotated by a recent researcher who worked on the project will be used by removing the labels (Chen, 2023). By using annotated files as the unlabelled dataset, it will be possible to compare the performance of the semi-supervised learning approach with a fully supervised model.

4.2.6 Sliding Window Transformation

To prepare the data for sequence-based classification, a sliding window approach was implemented, with different strategies for training and testing/validation sets. The primary reason for splitting the data into chunks is that the context of an entire recording is not required to identify a gesture for a specific portion of the recording. Moreover, it would be computationally expensive to use the entire data as context each time for frame-level classification.

The window size was set to 150 frames, which corresponds to approximately 1.5 seconds of data. This duration was chosen based on the data analysis section, which revealed that the average duration of gestures in the dataset is around 1.5 seconds. By using this window size, it is ensured that most gestures can be captured entirely within a single window, allowing the model to learn complete gesture patterns.

For the training set, chunks of data belonging to the same gesture were extracted. If the number of frames spanning a gesture exceeded 150, it was further split into multiple chunks of size 150. Gestures with a duration less than 150 frames were padded with zeros at the end. This approach ensures that each training sample represents a complete gesture, improving the model's ability to learn gesture-specific patterns.

For testing and validation sets, a direct window slicing method was used with the chosen window size of 150 and a stride of 10. The reason for using sliding window instead of gesture-wise chunks is that in a real-life scenario there would be no information regarding the gestures and chunks cannot be created accordingly. A stride of 10 was used instead of shifting the window by 1 each time to ensure that the new dataset does not grow exponentially. This leads to the loss of only 0.1 seconds of information and is a reasonable adjustment considering the computational requirements of using a stride of 1. The label for each window was assigned based on the majority label within the 150 frames.

4.2.7 One-Hot Encoding

Deep learning models typically require numerical inputs, necessitating the conversion of categorical labels into a numerical format. For the head gesture classification task, one-hot encoding was employed, a common technique for representing categorical variables. One-hot encoding creates a binary vector for each category, where the length of the vector equals the number of unique categories (Pedregosa et al., 2011b). Each category is represented by a vector with a '1' in the position corresponding to that category and '0's elsewhere. This approach avoids introducing ordinal relationships between categories that don't inherently exist. For this dataset, labels were one-hot encoded based on the following order:

```
['no', 'nd', 'fu', 'fd', 'sh', 'ti', 't']
```

If the label of a data is 'nd', it will be encoded as [0,1,0,0,0,0,0] for modelling purposes. The main purpose of one-hot encoding the labels is to create models that can learn and output probabilities for each gesture category independently, which can be useful during the semi-supervised learning process for determining the confidence level of a model prediction.

4.3 Data Augmentation

Initial data preparation, which involved chunking continuous motion capture recordings into fixed-length segments, significantly increased the overall number of samples and amplified the existing class imbalance issue. The majority of chunks were labelled as 'no' (non-gesture), creating a heavily skewed distribution that posed challenges for model training and generalisation. Figure 4.4 illustrates the severe class imbalance in the dataset. To address this issue, several techniques commonly used for multivariate time-series data were initially explored. These included Gaussian noise injection (Wen et al., 2020), which adds random noise to the original motion data; time warping (Iwana and Uchida, 2021), which simulates variations in gesture speed; and Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), which creates synthetic examples of minority classes. While these methods showed promise, a generative augmentation technique based on a Variational Autoencoder (VAE) was finally adopted due to its ability to generate realistic and diverse samples by learning the patterns from the data (Iglesias et al., 2023).

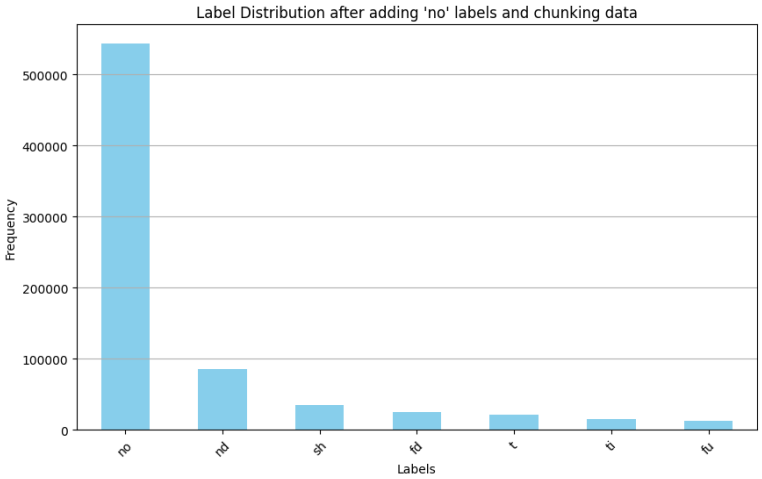


Figure 4.4: Distribution of Gesture Classes after Initial Data Preparation

4.3.1 Variational Autoencoder-Based Augmentation

Variational Autoencoders (VAEs) offer a powerful framework for data augmentation, particularly in the context of time series data like head gestures. VAEs are generative models that learn to encode input data into a lower-dimensional latent space and then reconstruct the data from this latent representation (Kingma, 2013). The key innovation of VAEs lies in their ability to learn a continuous, probabilistic latent space, which allows for the generation of new, diverse samples.

The underlying principle of VAEs is rooted in variational inference. The encoder network learns to map input data to a probability distribution (typically Gaussian) in the latent space, rather than to a fixed point. This allows the model to capture uncertainty and variability in the input data. It also enables smooth interpolation and sampling in the latent space, leading to the generation of new, plausible data points.

For time series data like head gestures, this probabilistic approach is particularly valuable. Head movements can vary in speed, amplitude, and duration, even for the same gesture type. The Gaussian latent space of a VAE can capture these variations, allowing for the generation of diverse yet realistic synthetic samples (Connor et al., 2021).

In the context of head gesture augmentation, a VAE can learn to encode the essential characteristics of each gesture type into the latent space. When sampling from this space to generate new data, the decoder can produce synthetic gesture sequences that maintain the core properties of the original gestures while introducing natural variations in timing and amplitude.

To implement this approach, The ”vae_conv5” model architecture from the TSGM (Time Series Generative Models) library was used for the head gesture classification task (Nikitin, 2022). This architecture employs convolutional layers in both the encoder and decoder, which are particularly effective at capturing local temporal patterns in time series data (Yao et al., 2019). Figure 4.5 shows a real data and synthetic sample generated by VAE of the nodding gesture. It is observed that even in the synthetic data the RV_z vector is the most pronounced and it has similar patterns to that of a real nodding gesture data.

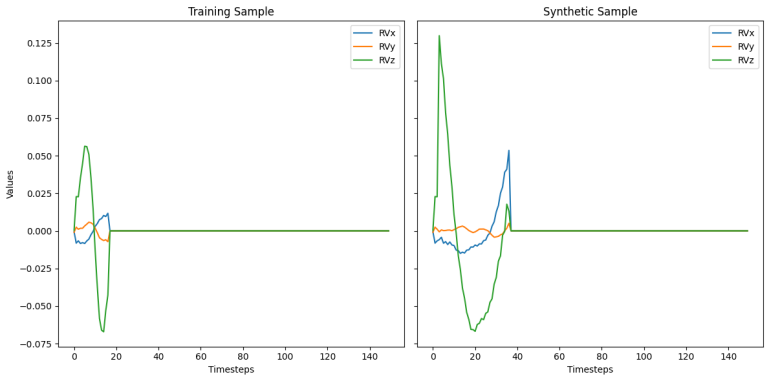


Figure 4.5: Original and Synthetic Data of Nodding Gesture

4.3.2 Balancing Strategy

The final balancing strategy proposed for the training data involved a combination of undersampling the majority class and oversampling the minority classes:

1. Undersampling: The ”no” gesture class was randomly undersampled to 30% of its original size.

2. Oversampling: All gesture classes were oversampled using the VAE to match the new size of the undersampled majority class.

This approach resulted in a completely balanced dataset, ensuring equal representation of all classes during model training. The use of VAE-generated samples for oversampling, as opposed to simpler techniques like adding Gaussian noise, was motivated by the VAE’s ability to capture and generate complex temporal patterns inherent in head gesture data (Cai et al., 2023).

4.4 Baseline Model Development

Establishing a robust baseline model is essential for evaluating the effectiveness of semi-supervised learning approaches in head gesture classification. This section details the development process of the baseline models, exploring both image-based techniques and other deep learning approaches suitable for sequential data.

4.4.1 Time Series to Image Conversion

Converting time series data to images is an interesting approach that has been adopted recently in several domains such as Speech Recognition and it has played a role in improving classification performance (Kaewrakmuk and Srinonchat, 2024). This approach offers several potential advantages:

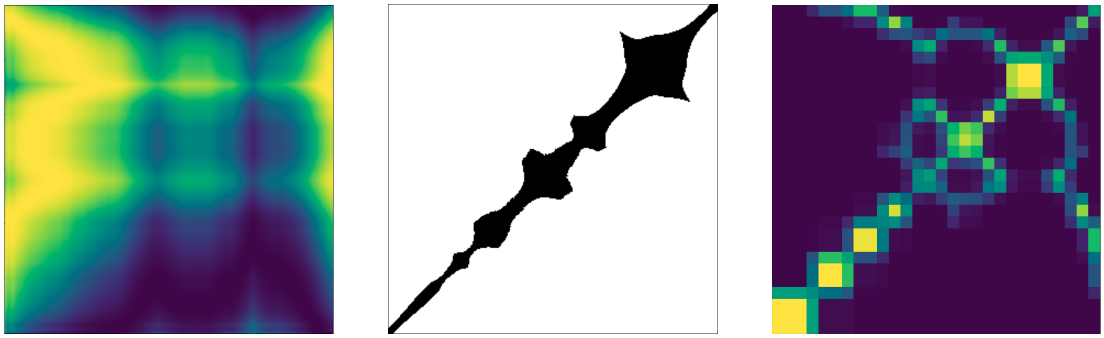
1. Variable-length inputs: Image conversion allows for flexibility in input sizes of the time series data that will be transformed, potentially capturing the optimal representation for each gesture type.

2. Utilisation of pre-trained models: State-of-the-art convolutional neural networks (CNNs) pre-trained on large image datasets can be leveraged through transfer learning.

Gramian Angular Difference Field (GADF) (Yang et al., 2019b), Recurrence Plots (RP) (Jiang et al., 2022), and Markov Transition Field (MTF) (Yang et al., 2019a) are the three main transformation techniques explored to convert the time series data to images. GADF represents temporal correlations as a polar coordinate image, RP highlights recurring patterns in the data, and MTF visualises the likelihood of transitions between different value ranges in the time series.

For each gesture segment, these transformations are applied to the rotational vector data (RV_x, RV_y, RV_z), resulting in square matrices that are treated as images. The transformations are applid to each of the features separately and the resulting images are stacked together as different channels. These images are then resized to 224x224 pixels to match the input requirements of pre-trained CNN architectures such as ResNet (He et al., 2015) and Vision Transformer (Dosovitskiy et al., 2021). The images after transformation are shown in Figure 4.6 for the nodding gesture.

Transfer learning was employed by fine-tuning these pre-trained models on the gesture dataset. Despite the success of this approach in other time series classification



(a) GADF (b) RP (c) MTF

Figure 4.6: Transformed Images for Nodding Gesture

tasks, significant challenges were encountered in achieving satisfactory performance for head gesture classification. The subtle nature of head movements and the complexity of the motion capture data likely contributed to difficulties in extracting meaningful features from the image representations. The variations amongst different images for different gestures was not substantial and the blurriness of the images could also have played a role in poor performance. It has also been observed in other researches that these techniques are not universally applicable to all domains and require modifications based on domain expertise and experimentation with different parameters (Li et al., 2024).

4.4.2 Neural Networks for Sequential Data

Given the sequential nature of the head gesture data, neural networks capable of handling time series data were explored as a more direct approach. Moreover, previous students who have worked on this project have achieved their best results using LSTMs and CNNs (Li, 2022).

4.4.2.1 1D Convolutional Neural Network (1D-CNN)

One-dimensional CNNs effectively capture local patterns and hierarchical features in time series data (Kiranyaz et al., 2021). A 1D-CNN model was developed to process the head gesture time series. The architecture includes: an input layer accepting 3D rotational vector data (150, 3); four 1D convolutional layers (filter sizes: 128, 128, 64, 128; kernel size: 4; ReLU activation); max pooling layers (pool size and stride: 2); a flatten layer; a dropout layer (rate: 0.5); and an output layer (dense, softmax activation). This design captures temporal patterns while mitigating overfitting through dropout

regularisation.

4.4.2.2 Bidirectional Long Short-Term Memory (Bi-LSTM)

LSTM networks excel at capturing long-term dependencies in sequential data (Hochreiter and Schmidhuber, 1997). A bidirectional LSTM (Bi-LSTM) was implemented to process time series data in both directions, capturing context from past and future time steps (Graves and Schmidhuber, 2005). The architecture consists of: an input layer for 3D rotational vector data (R_x, R_y, R_z); three stacked Bi-LSTM layers (300, 200, 100 units; recurrent dropout: 0.5); a dense layer (100 units, ReLU activation); a dropout layer (rate: 0.5); and an output layer (dense, softmax activation).

Both models were trained using categorical cross-entropy loss and the Adam optimiser. These baseline models provide a foundation for comparison with semi-supervised learning approaches explored later. The challenges in image-based approaches highlight the task's complexity and the importance of selecting appropriate architectures for time series data.

4.5 Proposed Semi-Supervised Learning Approaches

In this research, three semi-supervised learning approaches for head gesture classification are explored. These methods contribute to the utilisation of a large amount of unlabelled head motion data to enhance the generalisation and accuracy of the deep learning models. The reason for choosing the approaches mentioned below is the ability to use a baseline model that works well on head gesture data instead of using semi-supervised techniques based on architectures that might not work well with head gestures that are subtle in nature.

4.5.1 Self-Training Pipeline

Self-training is an iterative semi-supervised learning technique that progressively labels unlabelled data using a model's own predictions (Amini et al., 2024). The process begins with training a model on the available labelled data. This model is then used to make predictions on the unlabelled data. The most confident predictions, typically those exceeding a predefined threshold, are added to the labelled dataset as "pseudo-labels". The model is then retrained on this expanded dataset, and the process repeats for several iterations until the unlabelled dataset is exhausted or the model performance stops

improving. The best model in terms of F1-score is then used to make predictions on the test set.

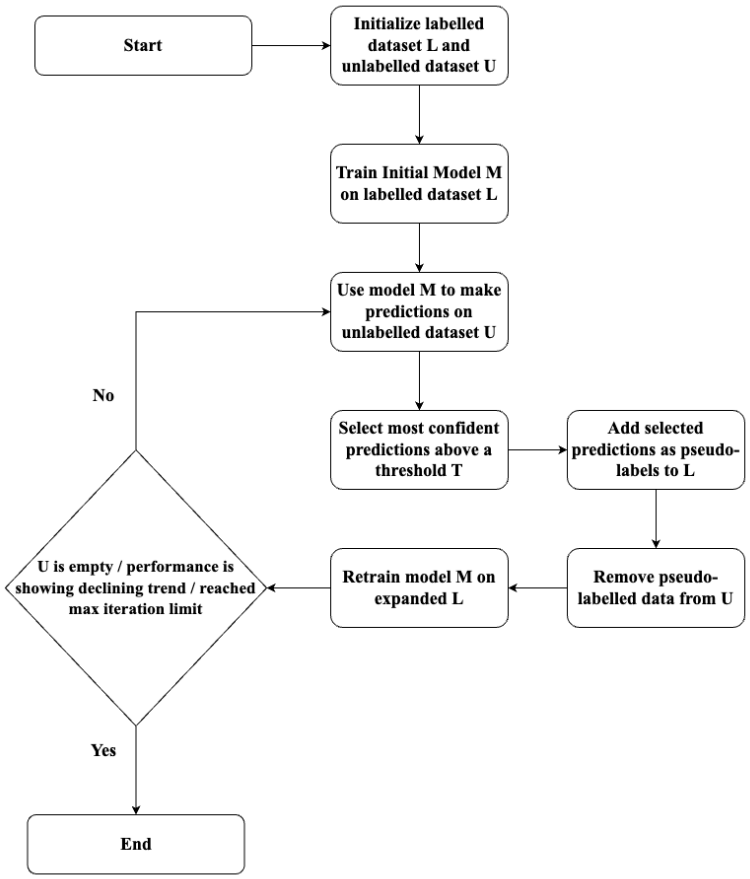


Figure 4.7: Self-Training Pipeline for Head Gesture Classification

The self-training pipeline implemented for head gesture classification is depicted in Figure 4.7. The Bidirectional LSTM (Bi-LSTM) and 1D CNN architectures are used as the base model due to better performance in comparison to the other architectures that were explored.

To thoroughly evaluate the self-training approach, several experiments have been designed:

- **Threshold Sensitivity**: The confidence threshold will be varied to understand its impact on pseudo-labelling accuracy and overall model performance. This experiment aims to find the optimal balance between incorporating more unlabelled data and maintaining high-quality pseudo-labels. It is expected that a medium threshold value that is not too flexible or restrictive will be the best option.

- **Labelled Data Proportion**: This research will investigate how the amount of initial labelled data affects the performance of self-training. This experiment will help understand the minimum amount of labelled data required for effective self-training in head gesture classification. It is expected that labelling at least half of the data and training it would give comparable performance to that of a fully supervised model.

4.5.2 Novel Ensemble Self-Training Pipeline

To further enhance the robustness and performance of the semi-supervised learning approach, a novel ensemble self-training pipeline was developed. This method combines the strengths of two different model architectures, a Bi-LSTM and a 1D Convolutional Neural Network (1D-CNN), inspired by the co-training semi-supervised learning technique (Blum and Mitchell, 1998). The distinctive feature in this ensemble approach

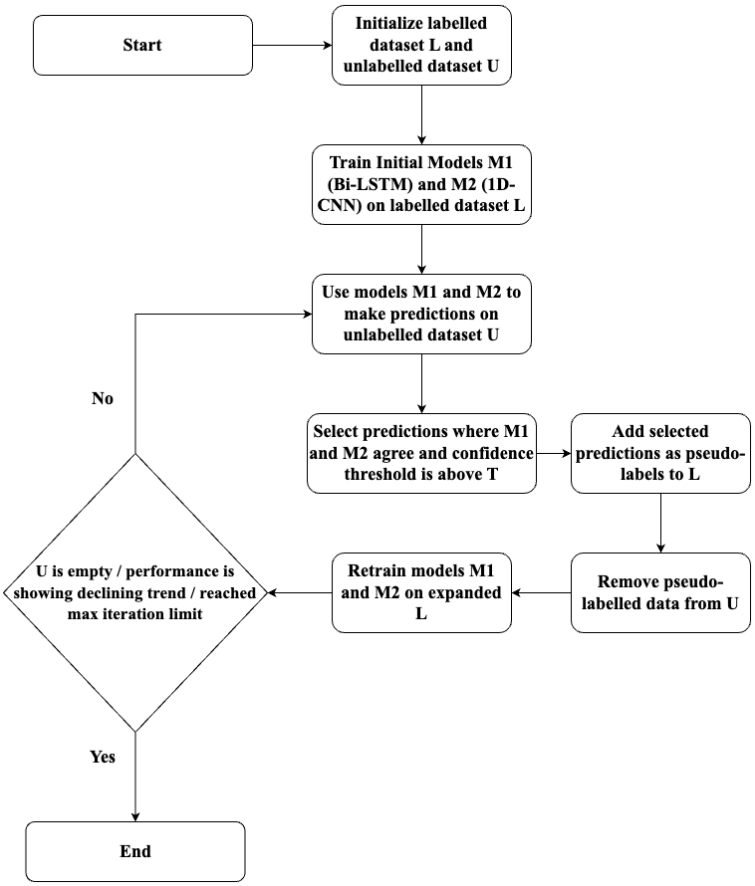


Figure 4.8: Ensemble Self-Training Pipeline for Head Gesture Classification

is the use of consensus between models to generate more reliable pseudo-labels. The

pipeline operates as shown in Figure 4.8. The 1D-CNN model excels at capturing local patterns and features in the time series data (Ige and Sibiya, 2024). In contrast, Bi-LSTMs process the entire sequence bidirectionally, capturing long-range dependencies and maintaining temporal context. By requiring agreement between these architecturally distinct models, the risk of propagating errors through pseudo-labelling is reduced.

For the ensemble self-training pipeline, the following experiments and analysis will be conducted:

- **Labelled Data Proportion**: Similar to the self-training pipeline, the model will be tested on different proportions of labelled data.
- **Architecture Comparison**: The performance of the proposed ensemble approach will be compared against each individual model (Bi-LSTM and 1D-CNN) to quantify the benefits of model combination in the context of semi-supervised learning for head gesture classification. It is expected that both the models will be distinct in terms of the type of predictions they make.

4.5.3 Mean Teacher Pipeline

The Mean Teacher method is another sophisticated approach to semi-supervised learning (Tarvainen and Valpola, 2018). This approach uses two models, a student model and a teacher model that are clones of each other in terms of the model architecture. It is the student model that is trained on all the labelled and unlabelled data and the teacher model is used for making predictions on the test set. The key idea is that the teacher model's weights are an exponential moving average of the student model's weights. This can be expressed mathematically as:

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t \tag{4.3}$$

where θ'_t represents the teacher model's weights at time step t, θ_t represents the student model's weights at time step t, and α is a smoothing coefficient that controls the update rate. It has been observed in studies that using the average model weights instead of the final weights leads to a better performing model (Polyak and Juditsky, 1992). The Mean Teacher pipeline for head gesture classification using Bi-LSTM model architecture is structured as shown in Figure 4.9. The student model is updated by minimising an overall cost function that combines the classification loss on labelled data and a consistency regularisation term. This overall cost function is defined as:

$$O(\theta) = \lambda C(\theta) + (1-\lambda)J(\theta) \tag{4.4}$$

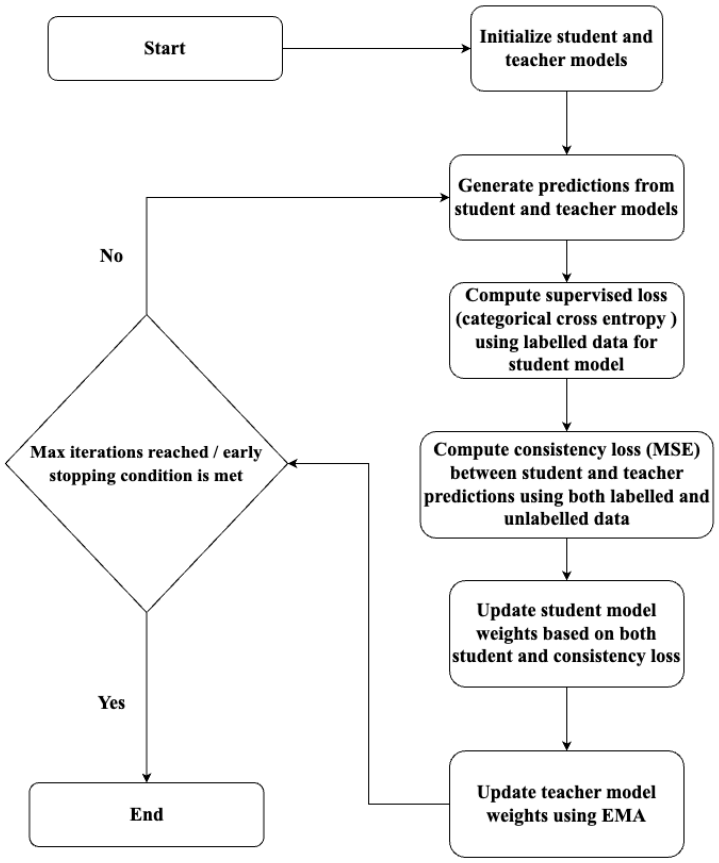


Figure 4.9: Mean Teacher Pipeline for Head Gesture Classification

where $O(\theta)$ is the overall cost function, $C(\theta)$ represents the classification loss on labelled data, $J(\theta)$ is the consistency regularisation term, and λ is a weighting parameter that balances the two terms.

The classification loss $C(\theta)$ ensures that the model performs well on the labelled examples, while the consistency regularisation term $J(\theta)$ encourages consistent predictions on unlabelled data, leveraging the teacher model's pseudo-labels. By minimising the cost function, the student model learns to make accurate predictions on labelled data while benefiting from the additional information provided by unlabelled examples, leading to the creation of a robust model. For the Mean Teacher approach, the model will be tested on different proportions of labelled data similar to the other pipelines.

4.6 Evaluation Framework

4.6.1 Performance Metrics

Two main metrics will be considered to evaluate the model performance:

- **Accuracy:** This metric provides an overall measure of the model's correctness across all classes. It is calculated as:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{4.5}$$

  While accuracy offers a quick overview of model performance, it can be misleading for imbalanced datasets (Powers, 2020). Therefore, it is complemented with more nuanced metrics.

- **F1-score:** The F1-score provides a balanced measure of precision and recall, making it particularly useful for multi-class classification tasks with potential class imbalances (Sasaki et al., 2007). It is calculated as the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.6}$$

  Where:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{4.7}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4.8}$$

  Macro F1-score is used to analyse the performance of the classifiers as it ensure that all gestures are treated equally without being biased based on the number of samples of each gesture (Opitz and Burst, 2021).

Apart from these above mentioned metrics, the best model will contain metrics such as Cohen's Kappa (McHugh, 2012) and Confusion Matrix (Heydarian et al., 2022) for a more in-depth analysis of performance.

4.6.2 Speaker-Independent Cross-Validation

To rigorously evaluate the models' ability to generalise across different speakers, a speaker-independent cross-validation strategy. This approach is crucial given the inherent variability in head gestures across individuals. An important point to note is that this approach is done only for the best semi-supervised model and a fully supervised model due to the large computational requirements.

The dataset comprises of recordings from 8 distinct speakers. For each fold of the cross-validation:

1. Data from one speaker was held out as the test set.
2. Data from the remaining 7 speakers was used for training and validation.
3. This process was repeated 8 times, with each speaker serving as the test set exactly once.

This strategy ensures that during testing, the model encounters head gestures from a speaker it has never seen during training, providing a stringent test of generalisation. Finally, the average performance across all 8 folds will be reported. It is expected that the average results will be comparable to the results observed for previous experiments.

Chapter 5

Results and Discussion

5.1 Baseline Model Performance

To establish a baseline for head gesture classification performance, several model architectures were evaluated. Table 5.1 presents the test accuracy and macro F1 scores for Bidirectional Long Short-Term Memory (Bi-LSTM), 1D Convolutional Neural Network (1D-CNN), and three image transformation techniques: Gramian Angular Difference Field (GADF), Markov Transition Field (MTF), and Recurrence Plot (RP), used with a ResNet model (He et al., 2015) by applying transfer learning.

Model	Accuracy (%)	Macro F1 Score
Bi-LSTM	58.02	0.39
1D-CNN	65.15	0.50
GADF	53.02	0.21
MTF	39.19	0.22
RP	62.00	0.26

Table 5.1: Performance Comparison of Baseline Models on Test Set

The results indicate that the 1D-CNN model achieved the highest performance, with an accuracy of 65.15% and a macro F1 score of 0.50 on the test set. This was followed by the Bi-LSTM model, which demonstrated moderate performance with an accuracy of 58.02% and a macro F1 score of 0.39. These results align with previous findings in time series classification tasks, where convolutional and recurrent neural networks have shown strong performance (Fawaz et al., 2019). Interestingly, while the RP technique achieved a relatively high accuracy of 62%, its low macro F1 score of 0.26 suggests

poor performance in classifying specific gestures. Upon closer examination of gesture-wise performance, it was observed that the image transformation techniques (GADF, MTF, and RP) primarily excelled at identifying no-gesture states and nodding gestures, while significantly misclassifying other gesture types. This imbalance in classification performance renders these models less suitable as robust baselines. As discussed earlier, the poor performance of image transformation techniques can be attributed to requirement of parameter tuning based on domain expertise and the complexities of the different head gesture. It was observed from the transformed images that the variations in the images for the different gestures apart from no gestures and nodding was subtle and the classifiers were not able to differentiate properly between them.

In conclusion, the 1D-CNN and Bi-LSTM models emerge as the most promising baselines for the head gesture classification task. These architectures demonstrate a better balance between overall accuracy and class-specific performance, making them more suitable for usage with semi-supervised learning techniques.

5.2 Proposed Semi-Supervised Learning Results

5.2.1 Self-Training Results

5.2.1.1 Comparison of Bi-LSTM and 1D CNN architectures

To evaluate the effectiveness of the proposed semi-supervised learning techniques for head gesture classification, a self-training approach using both 1D CNN and Bi-LSTM models as the base classifiers was implemented and tested. The self-training process utilised the recently annotated data from a previous researcher as unlabelled data, with a confidence threshold of 0.85 and a maximum of 50 iterations. Table 5.2 presents the test set performance of both models before and after applying the self-training pipeline.

Model	Initial Accuracy (%)	Initial F1-Score	Best Accuracy (%)	Best F1-Score	No. Iterations
1D CNN	65.35	0.52	66.93	0.56	14
Bi-LSTM	**58.02**	**0.39**	**68.32**	**0.63**	**32**

Table 5.2: Baseline and Best Performance on Test Set Using Self-Training

The results reveal several key insights. Both the 1D CNN and Bi-LSTM models show improvements in accuracy and F1-score after applying the self-training approach. This confirms the primary hypothesis that leveraging unlabelled data through semi-

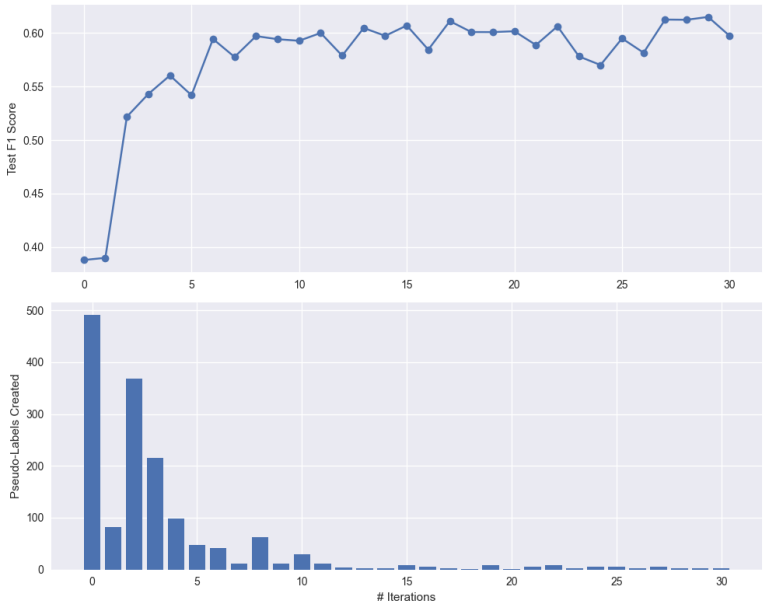


Figure 5.1: Pseudo-labels and Model Test Set Performance Across Different Iterations

supervised learning techniques can enhance classification performance. It also satisfies the objective of developing and implementing semi-supervised learning models.

The Bi-LSTM model demonstrates a more substantial improvement compared to the 1D CNN. While the 1D CNN shows a modest increase in accuracy (1.58 percentage points) and F1-score (0.04), the Bi-LSTM exhibits a remarkable jump in both metrics (10.3 percentage points in accuracy and 0.24 in F1-score). Interestingly, although the Bi-LSTM model started with lower initial performance (58.02% accuracy, 0.39 F1-score) compared to the 1D CNN (65.35% accuracy, 0.52 F1-score), it ultimately outperformed the 1D CNN after self-training. This suggests that the Bi-LSTM model was able to leverage the unlabelled data more effectively. The Bi-LSTM model was trained for more iterations (32) until the stopping condition was met compared to the 1D CNN (14). This indicates that the Bi-LSTM continued to learn and improve over a longer period in a stable manner.

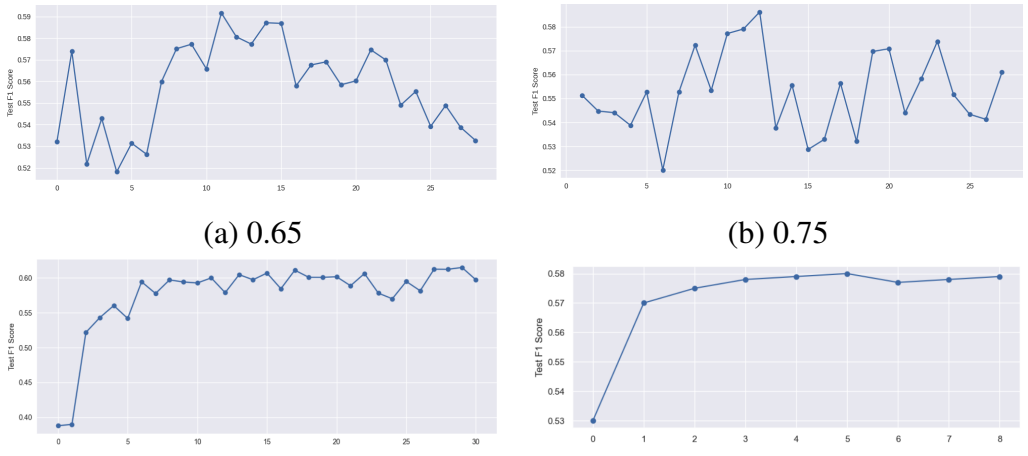
One of the reasons for the better performance of Bi-LSTMs overall is its ability to adaptively learn temporal scales. Head gestures occur at varying speeds and duration and Bi-LSTMs will be able to effectively capture both rapid movements and slower subtle gestures using its gating mechanisms (Hochreiter and Schmidhuber, 1997). Also, the ability of Bi-LSTMs to take both past and future context into account plays a major role.

Figure 5.1 provides an in-depth analysis of the Bi-LSTM model's performance on

the test set as it progresses through different iterations of the self-training mechanism. It shows the F1-score on the test set for different iterations as well as the number of pseudo-labels assigned during each iteration. It is observed that the model manages to assign several high confidence pseudo-labels in the initial iterations itself. This is reflected in the model performance as well since there is a spike in performance during the initial stages after which it has a gradual increasing pattern. Hence, the Bi-LSTM model combined with self-training is well-suited for the head gesture recognition task

5.2.1.2 Effect of Confidence Threshold on Self-Training Performance

To investigate the impact of the confidence threshold on the self-training process, experiments were conducted using the Bi-LSTM model with four different threshold values: 0.65, 0.75, 0.85, and 0.95. The choice of these values was to explore a range from moderately confident (0.65) to highly confident (0.95) predictions. Figure 5.2 illustrates the performance trajectories for each threshold on the test set. It is observed that the performance trajectory becomes more stable with increasing threshold values. For the lower values of 0.65 and 0.75, there is a lot of fluctuation and this can be attributed to misclassified pseudo-labels. For very high threshold values, the performance improvement is very minimal as the threshold is too restrictive. It stops in less than 10 iterations as well since it is unable to classify any more labels with high confidence. It is evident that a threshold of 0.85 is the optimal choice as it demonstrates a more stable and consistent performance improvement over iterations on the test set. Hence, 0.85 will be used as the optimal threshold values for future experiments.



(a) 0.65 (b) 0.75

(c) 0.85 (d) 0.95

Figure 5.2: Self-Training Performance Trajectories for Different Confidence Thresholds

5.2.2 Novel Ensemble Self-Training Results

5.2.2.1 Architecture Comparison

To understand the strengths of different neural network architectures used during the ensemble method and justify the superior performance of ensembling, their performance during the ensemble training pipeline is evaluated individually. Figure 5.3 illustrates the test F1 scores for each gesture type for both models. The results reveal distinct

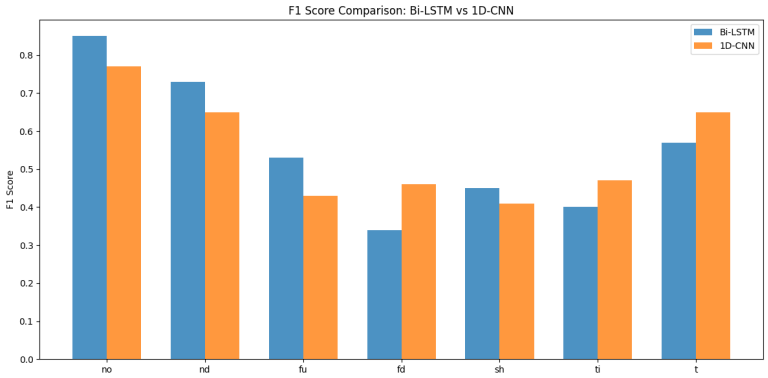


Figure 5.3: Test F1 Score Comparison of Bi-LSTM and 1D-CNN for Different Gestures

performance patterns. Bi-LSTM excels at recognising 'no gesture', 'nodding', 'face up', and 'shaking, while 1D-CNN shows superior performance for 'face down', 'tilting', and 'turning' gestures. The complementary nature of these architectures is evident where one model's weakness is often the other's strength. For instance, Bi-LSTM struggles with 'face down' gestures (F1 score of 0.34) while 1D-CNN excels (0.46), and Bi-LSTM's strong performance on 'no gesture' (0.85) compensates for 1D-CNN's relative weakness (0.77). This comparison justifies the ensemble approach, which leverages Bi-LSTM's temporal modelling and 1D-CNN's spatial feature detection to mitigate individual model weaknesses.

5.2.3 Impact of Varying Labelled Data Percentages Across Models

To investigate the effectiveness of the proposed semi-supervised learning techniques under different amounts of labelled data, experiments were conducted using varying proportions of labelled data for all three models: Self-Training, Novel Ensemble Self-Training, and Mean Teacher. The self-training approach uses a Bi-LSTM model with a 0.85 confidence threshold, the novel ensemble self-training approach uses a 1D-CNN and Bi-LSTM model, and the mean teacher model used a Bi-LSTM with a smoothing

coefficient (α) value of 0.99 and weighting parameter (λ) value of 0.6. These parameters were chosen as they led to the best performance. The label proportions were chosen based on previous literature related to semi-supervised learning (Xi et al., 2022). Table 5.3 summaries the results on the test set for all three models.

Model	**Labelled Data (%)**	**Initial**		**Final**	
		Accuracy (%)	**F1-Score**	**Accuracy (%)**	**F1-Score**
Self-Training	10	50.30	0.33	58.42	0.36
	20	57.03	0.28	62.97	0.44
	40	58.61	0.37	67.52	0.54
Ensemble Self-Training	10	50.30	0.33	50.50	0.37
	20	57.03	0.28	60.00	0.47
	40	**58.61**	**0.37**	**66.74**	**0.56**
Mean Teacher	10	50.30	0.33	51.09	0.40
	20	57.03	0.28	57.80	0.42
	40	58.61	0.37	59.00	0.45
Fully Supervised Bi-LSTM (100%)		-	-	71.09	0.58
Fully Supervised 1D CNN (100%)		-	-	68.00	0.54

Table 5.3: Test Performance Comparison with Varying Labelled Data

The results demonstrate that all three semi-supervised learning approaches consistently improve both accuracy and F1-score across different proportions of labelled data, confirming the effectiveness of leveraging unlabelled data to enhance performance. This addresses the project objectives of comparing semi-supervised and fully supervised models and analysing the trade-offs between model performance and annotation efforts.

A clear trend of improved final performance with an increase in the proportion of labelled data is observed across all models. This is expected, as more labelled data provides a stronger foundation for learning. Notably, when using 40% labelled data, all models show significant improvements, with the novel Ensemble Self-Training method achieving the highest F1-score of 0.56, closely followed by Self-Training at 0.54. These results are remarkably close to the fully supervised model's performance (F1-score of 0.58 for Bi-LSTM and 0.54 for 1D CNN), indicating that semi-supervised techniques can achieve comparable results with less than half of the labelled data.

The Self-Training approach demonstrates consistent improvement across all labelled data proportions. With 40% labelled data, it achieves an F1-score of 0.54, representing a 45.95% improvement from its initial score. This substantial gain aligns with standard performance improvements observed in literature related to semi-supervised learning techniques (Xi et al., 2022). The Self-Training method's performance using just 40%

labelled data is particularly noteworthy, as it comes very close to the fully supervised model's performance, with only a 0.04 difference in F1-score.

The Ensemble Self-Training method shows the most substantial gains, particularly with limited labelled data. Even with only 10% labelled data, the method shows improvement, increasing the F1 score from 0.33 to 0.37. The most significant gains are observed with 20% and 40% labelled data, where F1 scores increase by 0.19 in both cases. This indicates the method's effectiveness when there's a balance between labelled and unlabelled data. Notably, with 40% labelled data, the method achieves an F1 score of 0.56, which actually outperforms the fully supervised 1D CNN model (0.54) and nearly matches the fully supervised Bi-LSTM model (0.58).

The Mean Teacher approach, while showing improvements, underperforms compared to the other two methods. It demonstrates modest improvements in both accuracy and F1-score across all labelled data proportions. With 10% labelled data, the model achieves a 0.79 percentage point increase in accuracy and a 0.07 improvement in F1-score. Similar incremental gains are observed for 20% and 40% labelled data scenarios. However, there remains a significant gap between its performance with 40% labelled data and the fully supervised model.

The underperformance of the Mean Teacher approach can be attributed to several factors. One key issue is the lack of a thresholding system for assigning pseudo-labels. The Mean Teacher model assigns pseudo labels for all unlabelled data during each iteration, which can introduce noise during early stages of training when these labels might be unreliable. Additionally, while the Mean Teacher model employs consistency regularisation, the specific characteristics of head gesture data may require more sophisticated regularisation techniques. Methods like MixMatch (Berthelot et al., 2019) that combine multiple regularisation strategies could potentially yield better results.

These findings clearly indicate that for scenarios where labelling data is costly or time-consuming, it would be possible to annotate less than 50% of the data and still achieve comparable performance using semi-supervised learning techniques, particularly with the Novel Ensemble Self-Training or Self-Training methods. This insight directly addresses the hypothesis that semi-supervised learning can reduce the reliance on large annotated datasets while maintaining high performance.

The superior performance of Self-Training and Ensemble Self-Training methods in comparison to the Mean Teacher model highlights the importance of method selection in semi-supervised learning. These results underscore the need for careful consider-

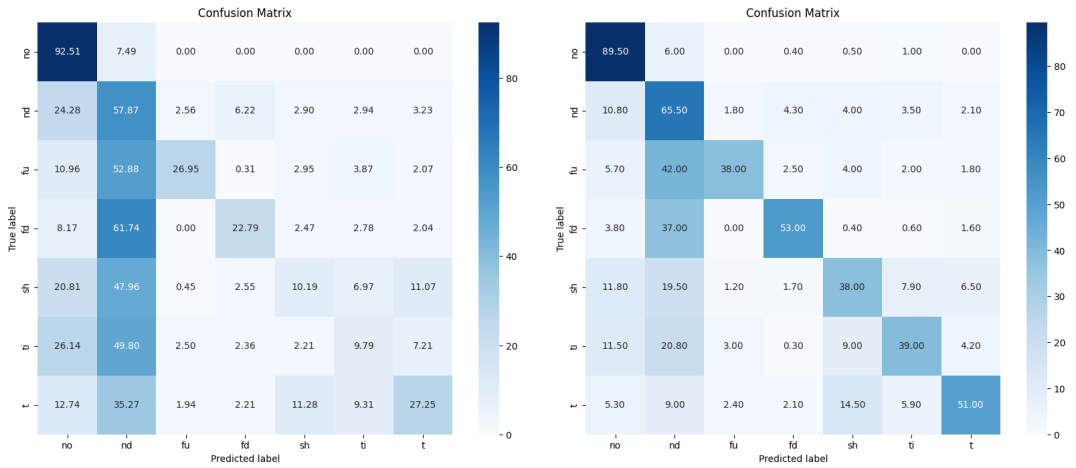
ation of the underlying data characteristics and model architectures when applying semi-supervised techniques to specialised domains like head gesture recognition. In conclusion, these methods not only significantly reduce the need for labelled data but also achieve performance levels comparable to fully supervised models, thus offering a valuable approach for scenarios where data annotation is resource-intensive.

5.3 Speaker-Independent Cross-Validation for Best Performing Model and Fully Supervised Model

To evaluate the generalisation capability of the best performing model and assess the impact of semi-supervised learning, a speaker-independent cross-validation was conducted. The ensemble self-training method with 40% labelled data was compared against a fully supervised Bi-LSTM model trained on the same 40% labelled data subset. This comparison directly addresses the research objective of analysing the model's generalisability across different speakers and quantifies the effectiveness of leveraging unlabelled data through semi-supervised learning.

The cross-validation was performed across 8 folds, with each fold using a different speaker as the test set. This speaker-independent approach yielded an average test accuracy of 66.4%, indicating the model correctly classified about two-thirds of all gestures across different speakers. The Cohen's Kappa value of 0.52 suggests moderate agreement between the model's predictions and true labels, accounting for chance agreement. The macro F1-score of 0.53, which balances precision and recall across all classes, indicates reasonable performance across different gesture types. These test results are promising in comparison to the previous works that had access to nearly 50% more annotations (Chen, 2023) In contrast, the fully supervised model achieved an accuracy of 54.77%, a Cohen's Kappa of 0.37, and a macro F1-score of 0.33. These results demonstrate a substantial improvement in performance when utilising semi-supervised learning, with increases of 11.63 percentage points in accuracy, 0.15 in Cohen's Kappa, and 0.20 in macro F1-score.

Figure 5.4 presents the average confusion matrices across all 8 folds for both models. The semi-supervised model demonstrates notably improved performance across most gesture categories. It shows particular strength in identifying 'nodding' (65.5% vs 57.87% accuracy) gesture. The performance for 'no gesture' is very similar for both models (89.5% vs 92.51%). The semi-supervised approach also significantly



(a) Fully Supervised Model (b) Semi-supervised Model

Figure 5.4: Confusion Matrices for Speaker-Independent Cross-Validation on Test Set

outperforms the fully supervised model in recognising less common gestures such as 'tilting' (39% vs 9.79% accuracy) and 'turning' (51% vs 27.25% accuracy).

Both models struggle with distinguishing between 'face up' (fu) and 'face down' (fd) gestures from the 'nodding' (nd) gesture, likely due to the overlap in motion patterns between these gestures, highlighting the challenge of distinguishing subtle differences in head movements. However, the semi-supervised model shows improved performance, particularly for 'face down' gestures (53% vs 22.79% accuracy). The semi-supervised model struggles most with 'face up' (fu) gestures, correctly identifying only 38% of instances, possibly due to the relative scarcity of these gestures in the dataset, as noted in the initial data analysis.

These results highlight several key points. The semi-supervised learning approach significantly enhances the model's ability to generalise across different speakers, as evidenced by the consistent improvement across all metrics. Leveraging unlabelled data through self-training helps the model learn more robust features, particularly benefiting the recognition of less common gestures. While challenges remain in distinguishing similar gestures, the semi-supervised approach shows promise in mitigating these difficulties. The substantial performance gap between the two models showcases the value of incorporating unlabelled data in the training process, especially in scenarios where annotated data is limited.

Despite these improvements, both models face challenges related to gesture similarity, data imbalance, and inter-speaker variability. However, the semi-supervised model's superior performance demonstrates its potential to better address these issues

by effectively utilising the additional information present in unlabelled data.

In conclusion, this comparison provides strong evidence supporting the hypothesis that semi-supervised learning can significantly enhance head gesture recognition performance in a speaker-independent context. The achieved improvements in accuracy, Cohen's Kappa, and F1-score represent a substantial step forward in the field, highlighting the potential of semi-supervised techniques in head gesture recognition.

Chapter 6

Conclusions

6.1 Summary

This research has helped improve automatic head gesture recognition through the application of semi-supervised learning techniques. The study addressed the challenge of limited labelled data in head gesture classification by leveraging unlabelled data to enhance model performance. The findings demonstrate the effectiveness of the proposed semi-supervised learning techniques, particularly self-training and ensemble methods, and support the primary hypothesis of the ability of semi-supervised learning to improve performance.

A key contribution to this study is the development of a novel ensemble self-training approach, which combines the strengths of Bi-LSTM and 1D-CNN architectures. This approach outperformed individual models and traditional self-training, achieving a test F1-score of 0.56 with only 40% labelled data, comparable to fully supervised models trained on 100% labelled data that has a test F1-score of 0.58. This result provides valuable insights into the trade-off between the amount of labelled data and model performance, suggesting that semi-supervised models can achieve performance close to fully supervised models with significantly reduced annotation efforts.

The research also revealed complementary strengths of Bi-LSTM and 1D-CNN architectures in recognising different types of head gestures, justifying the ensemble approach and providing insights for future model design in this domain. Furthermore, speaker-independent cross-validation demonstrated the model's ability to generalise across different individuals, achieving an average accuracy of 66.4% and an average F1-score of 0.53 on the test dataset. This represents a significant improvement over the fully supervised model, which achieved an accuracy of 54.77% and an F1-score

of 0.33 when trained on the same 40% labelled data subset. Moreover, the semi-supervised learning technique outperformed the previous works in speaker-independent head gesture recognition (Chen, 2023).

Despite these achievements, the research highlighted specific challenges in head gesture recognition, including the difficulty in distinguishing subtle differences between certain gestures (nodding vs. face up/down) and the impact of data imbalance on less common gestures. These findings not only advance the understanding of semi-supervised learning in the context of head gesture recognition but also provide practical insights for developing more accurate gesture recognition systems. The demonstrated ability to achieve high performance with reduced labelled data has significant implications for reducing the time and cost associated with data annotation in real-world applications.

6.2 Future Work

While this research has made substantial progress in head gesture recognition using semi-supervised learning, several avenues for future work have been identified. Future research should focus on enhancing the generalisability of the system by including samples from diverse cultural backgrounds. Head gestures can vary significantly across cultures, and a more diverse dataset would allow for the development of more robust and universally applicable models (Kita, 2009). Additionally, a multimodal approach incorporating audio and transcript data alongside motion capture data, could provide additional context related to the gestures (Baltrušaitis et al., 2019). This approach could help differentiate between similar gestures by considering verbal cues and conversation context.

Combining semi-supervised learning with self-supervised learning is an interesting approach to explore as it could possibly provide better performance with an even smaller amount of labelled data (Xi et al., 2022). Active learning strategies can also be implemented as it would improve the quality of the pseudo-labels due to human intervention (Settles, 2009).

These future directions aim to address the current limitations of the research and further advance the field of head gesture recognition. By pursuing these, researchers can work towards developing more robust, generalise, and practical head gesture recognition systems useful for human-computer interaction.

Bibliography

Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., and Maximov, Y. (2024). Self-Training: A Survey.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Cai, B., Yang, S., Gao, L., and Xiang, Y. (2023). Hybrid variational autoencoder for time series forecasting. *Knowledge-Based Systems*, 281:111079.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, X. (2023). Automatic classification of human head gestures using motion capture data. Master's thesis, University of Edinburgh, Edinburgh, UK. Master's thesis.

Choi, I.-H. and Kim, Y.-G. (2014). Head pose and gaze direction tracking for detecting a drowsy driver. In *2014 International Conference on Big Data and Smart Computing (BIGCOMP)*, pages 241–244.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Connor, M., Canal, G., and Rozell, C. (2021). Variational autoencoder with learned latent structure. In *International conference on artificial intelligence and statistics*, pages 2359–2367. PMLR.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Fang, Y., Cheng, J., Wang, J., Wang, K., Liu, J., and Lu, H. (2008). Hand posture recognition with co-training. In *2008 19th International Conference on Pattern Recognition*, pages 1–4.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE.

Haag, K. and Shimodaira, H. (2015). The University of Edinburgh speaker personality and MoCap dataset. In *Proceedings of the Facial Analysis and Animation*, pages 1–2.

Hachaj, T. and Piekarczyk, M. (2019). Evaluation of pattern recognition methods for head gesture-based interface of a virtual reality helmet equipped with a single IMU sensor. *Sensors*, 19(24):5408.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition.

Heydarian, M., Doyle, T. E., and Samavi, R. (2022). MLCM: Multi-label confusion matrix. *IEEE Access*, 10:19083–19095.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ige, A. O. and Sibiya, M. (2024). State-of-the-art in 1D Convolutional Neural Networks: A Survey. *IEEE Access*.

Iglesias, G., Talavera, E., González-Prieto, Á., Mozo, A., and Gómez-Canaval, S. (2023). Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*, 35(14):10123–10145.

Iwana, B. K. and Uchida, S. (2021). An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7):e0254841.

Jiang, H., Duerstock, B. S., and Wachs, J. P. (2013). A machine vision-based gestural interface for people with upper extremity physical impairments. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2870–2875. IEEE.

Jiang, W., Zhang, D., Ling, L., and Lin, R. (2022). Time series classification based on image transformation using feature fusion strategy. *Neural Processing Letters*, 54(5):3727–3748.

Kaewrakmuk, T. and Srinonchat, J. (2024). Multi-Sensor Data Fusion and Time Series to Image Encoding for Hardness Recognition. *IEEE Sensors Journal*.

Kingma, D. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398.

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes*, 24(2):145–167.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24.

Li, J. (2022). Automatic clustering of human head gestures using motion capture data. Master's thesis, University of Edinburgh, Edinburgh, UK. Master of Science Dissertation, School of Informatics.

Li, Z., Li, S., and Yan, X. (2024). Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36.

Lyu, Z. (2023). Automatic classification of human head gestures with neural networks. Master's thesis, University of Edinburgh, Edinburgh, UK. Master of Science Dissertation, School of Informatics.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Miah, A. S. M., Hadiuzzaman, M., Ali, M. S., and Mahdee, T. M. (2022). EEG-Based Hand Gesture Classification Using Machine Learning Approach. *BAUST JOURNAL*, page 19.

Murphy-Chutorian, E. and Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626.

Natural Point Inc. (2022). OptiTrack - Motion Capture Systems. https://optitrack.com/. Accessed: 2024-07-19.

Neverova, N., Wolf, C., Taylor, G. W., and Nebout, F. (2014). Multi-scale deep learning for gesture detection and localization. *IEEE transactions on cybernetics*, 45(3):502–514.

Nikitin, A. (2022). *Introduction — tsgm 0.0.7 documentation*. https://tsgm.readthedocs.io/en/latest/guides/introduction.html.

Opitz, J. and Burst, S. (2021). Macro F1 and Macro F1.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011a). Scikit-learn: Machine Learning in Python.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, (2011b). *sklearn.preprocessing.OneHotEncoder*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Rabiner, L. R. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16.

Rautaray, S. S. and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, 43(1):1–54.

Sasaki, Y. et al. (2007). The truth of the F-measure. Teach tutor mater, 1 (5), 1–5.

Settles, B. (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Shi, Y., Qiao, L., Shu, Y., Li, B., Xiao, B., Li, W., and Gao, X. (2024). Semi-Supervised FMCW Radar Hand Gesture Recognition via Pseudo-Label Consistency Learning. *Remote Sensing*, 16(13):2267.

Tarvainen, A. and Valpola, H. (2018). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Terven, J. R., Salas, J., and Raducanu, B. (2014). Robust head gestures recognition for assistive technology. In *International Conference on Pattern Recognition*, pages 359–364. Springer.

Toro-Ossaba, A., Jaramillo-Tigreros, J., Tejada, J. C., Peña, A., López-González, A., and Castanho, R. A. (2022). LSTM recurrent neural network for hand gesture recognition using EMG signals. *Applied Sciences*, 12(19):9700.

Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232.

Wang, K., Chen, Y., Zhang, Y., Yang, X., and Hu, C. (2023). Iterative Self-Training Based Domain Adaptation for Cross-User sEMG Gesture Recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2974–2987.

Wang, Y. (2023). Automatic classification of human head gestures with neural networks. Master's thesis, University of Edinburgh, Edinburgh, UK. Master of Science Dissertation, School of Informatics.

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., and Xu, H. (2020). Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*.

Wikipedia (2024). Axis–angle representation — Wikipedia, The Free Encyclopedia. Accessed: 2024-08-16.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Xi, L., Yun, Z., Liu, H., Wang, R., Huang, X., and Fan, H. (2022). Semi-supervised time series classification model with self-supervised learning. *Engineering Applications of Artificial Intelligence*, 116:105331.

Yang, C.-L., Chen, Z.-X., and Yang, C.-Y. (2019a). Sensor classification using convolutional neural network by encoding multivariate time series as two-dimensional colored images. *Sensors*, 20(1):168.

Yang, C.-L., Yang, C.-Y., Chen, Z.-X., and Lo, N.-W. (2019b). Multivariate time series data transformation for convolutional neural network. In *2019 IEEE/SICE International Symposium on System Integration (SII)*, pages 188–192. IEEE.

Yang, Z. (2022). Automatic Classification and Clustering of Human Head Gestures. Master's thesis, University of Edinburgh, Edinburgh, UK. 4th Year Project Report, Artificial Intelligence, School of Informatics.

Yao, G., Lei, T., and Zhong, J. (2019). A review of convolutional-neural-network-based action recognition. *Pattern Recognition Letters*, 118:14–22.

Zhao, J. and Allison, R. S. (2017). Real-time head gesture recognition on head-mounted displays using cascaded hidden Markov models. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2361–2366.

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.