Alsofyani, Huda (2024) *Attachment recognition in school-age children through multimodal analysis of verbal and non-verbal behaviour.* PhD thesis

http://theses.gla.ac.uk/84392/

Attachment Recognition in School-age Children Through Multimodal Analysis of Verbal and Non-verbal Behaviour

Huda Alsofyani

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

June 2024

Abstract

Attachment is a psychological construct concerning the affectional and emotional bonds between children and their caregivers. Attachment styles can be one of two main types: Secure attachment which describes a state where the child's emotional needs are fulfilled, and Insecure attachment when these needs are unfulfilled. The formation of these attachment styles during childhood shapes the internal representation of close relationships which in turn impacts the quality of adulthood relationships and life. Moreover, past studies found that insecure attachment is linked to major issues such as heart diseases and antisocial behaviours. Early psychological interventions can mitigate the potential negative consequences of insecure attachment. To this end, it is essential to identify insecure children as early as possible. Various attachment assessment tests have been devised for infants and children which rely on observing their behaviours when exposed to distress situations. However, these tests suffer from a major drawback that impedes their applicability for population large-scale screenings, which is the need for trained professionals. One promising solution to overcome this challenge is by automating the assessment process and therefore increasing the applicability of the assessment tests.

In this thesis, automatic attachment recognition approaches based on the Manchester Child Attachment Story Task (MCAST) assessment test are proposed based on three main behavioural channels: facial expressions, paralanguage, and language. In addition, this thesis explores the benefit of combining different modalities to enhance the recognition rate. The results show that the attachment styles can be detected automatically with an accuracy of up to 75% by combining paralanguage and language modalities, and an F1-score of up to 68% by combining face and language modalities. Additionally, age and gender based performance analyses are conducted revealing that older children are more likely to express their condition through facial expressions while younger children are more likely to express it through paralanguage and language. Gender based effects are observed too and it is found that female children are more likely to express their condition through facial expressions, while male children tend to express it through language. These findings can enhance the recognition rate because it might be useful to adopt different modalities for different age or gender groups. It is also shown in this thesis that incorporating a confidence measure can increase the applicability of the approaches, for instance, by setting an acceptance level corresponding to 80% accuracy, which can be considered as human-level performance, 77% of the predictions can be accepted using an approach based on combining all of the three behavioural channels.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my sincere gratitude to Professor Alessandro Vinciarelli, my supervisor, for all of his assistance, support, patience, and kindness during this invaluable journey that started with a master's degree course, went on to include the master's project, and concluded with the completion of this PhD research. Prof. Vinciarelli was really eager to support me in pursuing my objectives and endeavours over this lengthy journey.

I would like to extend my gratitude to my annual review panel members, Prof. Iadh Ounis and Dr. Michele Sevegnani who were very supportive and their constructive suggestions of the materials I submitted for evaluation were extremely helpful and directed me toward improving my research.

I would like to thank my sponsor, Taif University, for offering me this scholarship opportunity to pursue my dreams and for their continuous support throughout my studies.

Furthermore, I would like to thank everyone who shared this journey with me and offered words of support, encouragement, and inspiration along the way.

Lastly, my deepest and warmest appreciations go to my mother, my father (who is in heaven), and my siblings for their unwavering support and belief in me since I was a young child and for always trying their hardest to assist me to reach this point in my life.

THANK YOU SO MUCH

Declaration

With the exception of chapters 1, 2, and 3, which contain introductory material, the background survey, and the dataset description, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Introduction

1.1 Attachment

Attachment theory provides a framework to understand the interactions and affectional bonds between children and their caregivers. The foundations of this theory were first developed by John Bowlby during the 1930s and later formulated in his famous trilogy *Attachment and Loss* [1–3] in which he described the maternal bonds and the effects of its disruptions through loss and separation. The theory was further developed by Mary Ainsworth who advanced the understanding of infant-mother attachment patterns and devised the Strange Situation Procedure (SSP) [4] which became the standard assessment test for attachment patterns in infants. In general, the theory states that the interactions with the caregivers shape the internal mental representations (Internal Working Models, IWMs) of children about close relationships and hence will impact their relationships throughout their adult lives.

According to the theory, there are two main attachment styles in children: *Secure attachment*, which describes the psychological state in which the affectional needs of children are met, and *Insecure attachment* which describes the state in which these needs are not met. Children develop a particular attachment style depending on the quality of their interactions with their caregivers. Insecure attachment has major effects on one's quality of life as it shapes their internal representation of e.g., friendship and professional relationships. Moreover, past studies found that insecure attachment style is linked to various issues such as heart pathologies [5, 6] and antisocial behaviours [7, 8]. Another study suggested that the differences in the way adults experience romantic love are rooted in the differences in their IWMs that were shaped by the attachment style during childhood [9].

Early psychological interventions can alter the IWMs of children which therefore will be reflected in their adulthood life quality. For this to be achieved, it is essential to identify children who develop insecure attachment as early as possible. Several psychometric tools have been devised which all share the core idea of observing the behaviours of children when exposed to distress situations. Among these, the Strange Situation Procedure (SSP) was developed to identify attachment in infants, the Attachment Doll-play Interview (ADI) was developed for pre-schoolers [10], and the Manchester Child Attachment Story Task (MCAST) was devised for school-age children [11].

These tools have been used intensively for research purposes and have been proven effective in identifying attachment-related issues, however, they suffer from a major drawback that limits their applicability in clinical practices [12]. Administering these tools takes around 20 minutes per child and the assessment process can take between one to two hours depending on the complexity of the materials [11, 12]. Moreover, in order for a practitioner to be eligible to conduct the assessment test, a rigorous training procedure has to be followed to meet a certain level of agreement with experts, which can take up to two weeks of training [12]. This, as one may expect, reduces the applicability of these tools as it limits the number of professionals who are eligible to conduct these assessment sessions and the amount of assessment load per person is rather limited.

Another important issue regarding the attachment condition is how prevalent it is in the general population. Past studies showed that the percentage of school-age children with insecure attachment style in the general population can be as high as 40%- 50% [13, 14]. This emphasises the need for a large-scale screening during early childhood which is not feasible under the existing conditions.

Given the aforementioned limitations of the existing assessment procedures, one promising solution is to automate the assessment process in which the behaviours of children are analysed automatically to predict the attachment styles. Such a solution is possible in principle, due to the advancements in signal processing and deep learning over the past decades. In fact, these technologies were used for similar problems such as depression detection, autism detection, and emotion recognition.

This thesis proposes automating the Manchester Child Attachment Story Task (MCAST) [11]. This is attempted by analysing the verbal and non-verbal behaviours of children during the MCAST sessions in which they were introduced to several distressing scenarios and the role is to detect patterns of variations that may be associated with attachment styles. The behavioural channels that are explored in this research comprise the three main behavioural modalities that humans use to manifest their inner feelings and mental states: facial expressions, paralanguage[1], and language. Moreover, motivated by the fact that these channels often complement each other and due to the success of combining these channels to predict other mental state conditions, multimodal approaches are proposed in which all possible combinations of these channels are attempted.

The MCAST is a story-completion doll-playing method that aims to elicit attachment behaviours in school-age children. It is based on introducing children to five stressful scenarios in the forms of stories designed to be age-appropriate. After listening to these scenarios, the children are asked to complete these stories (using dolls representing the child and the primary caregiver) and to describe the feelings of the dolls. These stories are expected to stimulate attachment-related behaviours in the way children complete the stories and describe the dolls' feelings.

The proposed approaches are validated on a corpus of 104 children of age 5-9 years and the overall results confirm that the variations in attachment styles leave machine-detectable behavioural cues in the aforementioned channels and it is shown that the attachment can be detected with an accuracy of 75.6% and F1-score of 68.8%.

1.2 Thesis Statement

This thesis investigates whether the variations in attachment styles leave traces in children's behaviours that can be detected automatically by means of machine learning and artificial intelligence approaches. This can be exploited to build an automatic assessment tool for attachment styles in children. Specifically, this Ph.D. research investigates the feasibility of automatically detecting the attachment styles in school-age children through unimodal and multimodal analy-

[1] Non-verbal aspects of speech.

sis based on three behavioural channels: facial expressions, paralanguage, and language.

1.3 Contributions and Publications

The main contributions of this thesis can be summarised in the following points:

1. To the best of our knowledge, this is the first attempt to detect the attachment styles in school-age children by analysing: facial expressions, paralanguage, and language using well-established pattern recognition methodologies. The results show that the attachment styles can be detected from these channels with an accuracy between 64.3% for the facial expressions-based approach and 72.1% for the language-based approach and F1-score between 54.8% and 63.9%.

2. First analysis of the effect of age and gender on attachment recognition performance.

3. First analysis of the effects of the variations in the dataset including recording length and the amount of speech which reveals important aspects of the approaches.

4. Definition of a confidence measure in order to increase the applicability of the approaches.

Several parts of this thesis (mainly from chapters 4 and 5) are based on papers that were presented at international conferences:

1. H. Alsofyani and A. Vinciarelli, “Stacked recurrent neural networks for speech-based inference of attachment condition in school age children.,” in Proceedings of Interspeech, pp. 2491–2495, 2021.

2. H. Alsofyani and A. Vinciarelli, “Attachment recognition in school age children based on automatic analysis of facial expressions and nonverbal vocal behaviour,” in Proceedings of the International Conference on Multimodal Interaction, pp. 221–228, 2021.

3. H. Alsofyani and A. Vinciarelli, “Attachment recognition in school-age children: A multimodal approach based on language and paralanguage analysis,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8172–8176, 2022.

4. A. Buker, H. Alsofyani, and A. Vinciarelli, “Multiple Instance Learning for Inference of Child Attachment From Paralinguistic Aspects of Speech,” in Proceedings of Interspeech, pp. 1045–1049, 2023.

1.4 Thesis Structure

This thesis is structured as follows, chapter 2 presents the literature review and background information about methodologies that were adopted in the recognition approaches. Next, chapter 3 describes the MCAST assessment tool, the SAM administration system, and the dataset that was used to evaluate the approaches along with various related statistical information. Then, chapter 4 describes the three unimodal approaches and presents the effect of age and gender on their performance. The chapter also provides a comparative analysis of the performance of the three approaches. The multimodal approaches are described and the results are presented in chapter 5 along with an analysis of the effect of age and gender. In chapter 6, analysis studies are presented which were conducted to inspect the effects of various factors of variation in the dataset, including recording length and amount of speech. This chapter also presents the performance after incorporating a confidence measure that was developed to increase the applicability of the approaches. Lastly, chapter 7 draws conclusions, describes the challenges that were faced in conducting this research, and suggests directions for future work.

Chapter 2

Literature Review and Methodological Background

This chapter presents the literature review and the relevant methodological background for the later chapters. The literature review is presented in three parts to cover the research fields outlined in Figure 2.1: The first part reviews previous works that aim to automatically detect mental and developmental disorders such as depression and autism, the second reviews the works concerning attachment condition in particular, in which computing and technology play a key role in conducting these studies, for example, several works aim to foster the attachment bonds between children and their caregivers while other works aim to enhance these attachment bonds with social robots, the third reviews the works that lie in the intersection of the previous two, i.e., those that aim at detecting the attachment style automatically. This PhD research falls in the third part as it aims to perform such a task in school-age children.

The relevant background is covered in several sections which present the methodologies that were adopted in the recognition approaches. In particular, the descriptions of the RNN, LSTM, and Deep RNN are presented in sections 2.2.1-2.2.3, whereas CNNs and their applications to NLP problems are discussed in section 2.2.4. Logistic regression is briefly discussed in section 2.2.5. Moreover, section 2.2.6 explains the most common ways to create classifier ensembles.

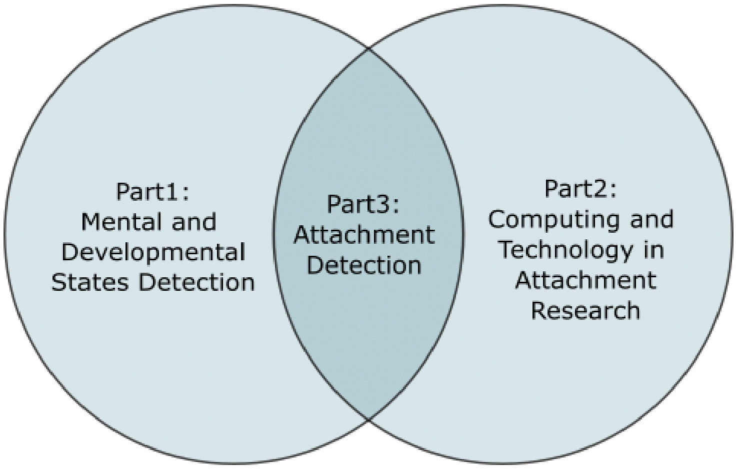


Figure 2.1: **Literature Survey Outline**. This figure shows the different categories of the literature survey. Part 1 is presented in section 2.1.1, Part 2 is presented in section 2.1.2, and finally, Part 3 is presented in section 2.1.3, in which this PhD research is situated.

2.1 Survey of Previous Work

2.1.1 Mental States Automatic Detection

The task of automating the detection of mental and developmental conditions has drawn increasing attention in the past years. The main reason is that diagnosis processes tend to be expensive and time consuming because they require trained professionals to conduct these assessments. The automatic detection of conditions such as depression and autism have been explored widely in the literature, whereas other conditions such as schizophrenia and attachment are still in their infancy. These attempts have explored several modalities e.g., facial expressions, movement, vocality, language, brain activity, etc. In the following sections, a detailed survey of the main works is presented. Two sections will be dedicated to depression and autism related works, while the remaining section reviews the main works that tackle other conditions.

Depression

In the case of depression, the main explored channels are visual and vocal, while linguistics and brain activity have been investigated in a limited number of works. Several datasets have been developed for depression detection as part of the Audio-Visual Emotion Challenge (AVEC) such as AVEC2013 and AVEC2014 in which the number of participants is 82 and 84, respectively. Another widely adopted dataset is the DAIC-WOZ which is the benchmark dataset for the later editions of the AVEC challenge (the number of participants in this dataset is 189).

In these datasets, the depression severity labels are provided by self assessment questionnaires such as the Beck Depression Inventory (BDI-II) and the Patient Health Questionnaire (PHQ). The majority of the works listed below have adopted these datasets.

The authors in [15] used the facial spatiotemporal features and head movements, and proposed a deep learning approach to predict the BDI-II score. In their approach, a 3D Convolutional Neural Network (CNN) is used to learn deep features from face and head regions. These features are then fed to a Recurrent Neural Network (RNN) to estimate the final score. They evaluated their approach on AVEC2013 and AVEC2014 and were able to achieve 9.28 and 9.20 RMSE (Root Mean Square Error), respectively.

Facial spatial features are also adopted in [16], where a deep CNN is built to map pixel-level features into depression scores. In this work, the authors built a global CNN and multi-regional CNNs where the full face images are fed to the global CNN while sub-regions of the face images (top, centre, and bottom) are fed to the corresponding CNNs. They also evaluated the performance on AVEC2013 and AVEC2014 and yielded 8.28 and 8.04 RMSE, respectively. One main finding of this work is that the top-central region of the face is more discriminating than the lower region for the purpose of depression detection.

Similarly, Facial appearance and dynamic features are adopted in [17, 18]. In [17], two separate deep CNN models were constructed. The first one is the appearance model in which the raw video frames are mapped to a depression score, whereas the second one is the dynamic model in which optical flow images are used to capture the face dynamics. This approach led to 9.82 and 9.55 RMSE when evaluated in AVEC2013 and AVEC2014, respectively. Additionally, they concluded that the dynamics of the face are as important in detecting the depression level as the facial appearance.

A similar two-stream framework is applied in [18]. In this work, spatial features were extracted by the Inception-Resnet network either by using the full frame image (holistic features) or mini patches of frames (local features) while dynamics were captured using Volume Local Directional Number (VLDN). Both spatial and dynamic features were aggregated and fed to a separate bidirectional long short-term memory (Bi-LSTM), and the results were averaged to obtain the final prediction. This framework was also evaluated in AVEC2013 and AVEC2014, yielding 8.93 and 8.78 RMSE, respectively. One important finding is that the inclusion of the mini patches' local features has improved the performance compared to when only the global holistic features were used.

Temporal facial features were explored in [19] in which the time series of three facial features were extracted: eye gaze, facial landmarks, and action units. Several deep learning models were explored in this study (CNN, LSTM, and Recurrent CNN). In addition, the authors explored adding a self-attention layer to each one of these models. The best performance was achieved by the CNN model coupled with the attention layer using the eye gaze features yielding an F1-score of 0.81 when evaluated on a subset of the DIAC-WOZ dataset. Additionally, they found that the attention layer has improved the performance for both CNN and LSTM models but not for the RCNN.

In another work [20], the authors argued that the depression score should be regarded as an ordinal variable in which there is no guaranteed regular relationship between features and depression score, an assumption that is widely adopted in conventional regression solutions. To handle this ordinal nature, they proposed a 2-stage Rank Regression framework, where first, the depression scores range is partitioned, and a ranking function is defined per partition. These ranking functions generate a score that measures the association between the features and the corresponding partition. In the second stage, the ranking results are fed to a regression model to predict the depression score. They evaluated their framework on AVEC2013 using different speech features, and they were able to achieve 8.50 RMSE.

Other work proposed using an unsupervised pre-training strategy with a hierarchical attention mechanism to tackle the problem of the limited amount of labeled materials in depression datasets which may hinder the applicability of deep learning techniques [21]. In this work, an autoencoder was pre-trained with a rich speech recognition dataset to learn the hierarchical attention maps[1] and then a Bi-LSTM model was fine-tuned with the depression dataset and utilising the attention weights which was learnt over the larger dataset. They evaluated their approach in DAIC-WOZ achieving 5.51 RMSE.

Psycholinguistics and mood were explored in [22], where the authors analysed 10000 posts written in online communities. They extracted several psycholinguistic features such as affective ratings, mood tags, and topics and were able to discriminate between depressed and non-depressed

[1] Two levels of attention were learnt (Frame-Level and Sentence-Level).

people with an accuracy up to 93% using only topics features and Lasso logistic regression.

Other works fused different modalities for detecting depression levels. For instance, visual and audio channels were fused in [23–27] while in [28], these channels were further augmented with a text descriptor. In [23], Mel Frequency Cepstral Coefficients (MFCC) were extracted and combined with GMM-UBM (Gaussian Mixture Models-Universal Background Mode) paradigm to construct the audio feature vectors whereas Space-Time Interest Points (STIP) and Pyramid of Histogram of Gradients (PHOG) were adopted to construct the visual features. These features were fused and fed to SVR[2] for prediction. The performance was also evaluated on AVEC2013 achieving 10.62 RMSE. However, both audio and visual features performed better individually than the fused ones achieving 10.17 and 10.45 RMSE, respectively.

Different types of visual and audio descriptors were chosen in [24], where Local Binary Patterns (LBP), Edge Orientation Histogram (EOH), and Local Phase Quantisation (LPQ) were used as visual descriptors. Using these features, Motion History Histogram (MHH) was constructed to capture the temporal variations. Low-level descriptors (LLD) and MFCC were extracted to form the audio features. By fusing these descriptors at the decision level and using Partial Least Squares (PLS) regression, a performance of 10.26 RMSE was obtained. Similar work was done in [26], where the hand-crafted facial features were instead substituted with deep ones. These deep features were extracted using a pre-trained CNN on a larger non-depression dataset (VGGFace and AlexNet). Next, the temporal movement was captured from these features using Feature Dynamic History Histogram (FDHH). Better performance was achieved (7.43 RMSE) by fusing these features with the same audio features as in [24] (10.26 RMSE), which suggests that deep features outperform the hand-crafted ones.

A two-stage framework was suggested in [25], where first a sample is classified into one of the depression classes then, within that class, the depression BDI score is predicted using a regression method. Several visual descriptors were extracted and integrated using the Fisher Vector and then fused with MFCC vocal features. This framework was evaluated on AVEC2013, and the performance was improved from 9.51 RMSE, when only a regression method is implemented, to 8.29 RMSE. It is worth noting that, in this work, the authors argued that the depression detection problem should be regarded as a cost-sensitive problem in which the penalty of

[2] Support Vector Regression.

misclassifying an individual to a class that is far from the actual one should be higher than misclassifying to a nearer class. They were able to achieve this by constraining the classifier cost function to a weight matrix that takes into account the distance between the actual and predicted classes.

A multimodal spatiotemporal attention framework was proposed in [27]. In this work, a spatiotemporal representation was constructed from segments of audio and video input data in which attention was used to relate both representations and highlight the frames that contributed to the detection. Next, the authors proposed using Eigen Evolution Pooling (EEP) to summarise the changes in the segment-level representations and to aggregate these summaries from both modalities. Subsequently, a multimodal attention fusion was used to better capture the complementary information conveyed by these different modalities and lastly, this multimodal representation was mapped to the depression score using an SVR model. This framework was evaluated on AVEC2013 and AVEC2014 yielding 8.16 and 7.03 RMSE respectively.

Along with the visual and audio channels, a text descriptor was constructed in [28] to estimate the PHQ-8 score. In this work, several visual and audio features were fed to a DCNN to be mapped to a higher representation, which then was fed to a deep network DNN to estimate the depression score achieving 6.34 RMSE on AVEC2016 dataset. The authors also demonstrated that the speaking segments of the interviews yielded better performance than the non-speaking segments. Additionally, a paragraph vector was extracted from the interviews' transcripts and classified using SVM. The results were fused through a Random Forest with the aforementioned DNN model to classify the dataset into Depressed/NonDepressed groups yielding an F1-score of 0.746.

In [29], the duration and n-gram counts of six speech landmarks were investigated and validated on two different datasets. The first dataset (DAIC-WOZ) is laboratory controlled with 28 depressed and 114 healthy participants, whereas the second one (SH2-FS) was collected in an uncontrolled environment with 97 depressed and 470 healthy participants. Using SVM, an accuracy of 94.3% and 72.7% was achieved in DAIC-WOZ and SH2-FS, respectively.

All previous works are based on self-reported questionnaires such as (BDI-II or PHQ-8) to measure the depression level. One downside of using questionnaires-based labels is that these labels can be subjective and do not always reflect the actual condition. For this reason, other works have adopted depression datasets with clinical diagnosis [30–34]. In [30], the authors regarded the depression detection as a 3-classes classification problem (Remission, Mild, and Moderate to Severe). They adopted facial dynamics (time series of coordinates of 49 facial points), head movements (time series of 3 degrees of freedom: pitch, yaw, and roll), and vocalisation (switching pause and fundamental frequency) and generated deep representations of these features using Stacked Denoising Autoencoders (SDAE). This representation was further encoded to produce fixed-length descriptors across input videos using Improved Fisher Vector coding (IFV) and Compact Dynamic Feature Set (DFS) which was later classified using logistic regression achieving 78.67% accuracy on a clinical dataset of 49 participants.

Similar feature modalities were adopted in [31], where speech signals, eye activity, and head movements were explored, and a clinically validated dataset with 60 subjects (30 depressed and 30 healthy control) was utilised. The best performance (88.3 Average Recall AR) was yielded by fusing speech and head modalities at the feature level and classified using SVM.

Vocal prosody was explored and validated on a clinical dataset of 57 participants with Major Depressive Disorder (MDD) in [32] through Switching Pause (SP) and Fundamental Frequency F0 features of both participants and interviewers. Using logistic regression, the depression level was detected with an accuracy of 69%. One of the main findings is that naive listeners can differentiate between participants with low and moderate-to-severe scores from vocal prosody with an accuracy up to 73%; however, moderate-to-severe depression scores are less predictable than lower scores.

Finally, EEG[3] signals were explored in [33, 34], where in [33], signals from a 3-electrode EEG device were used to build a Case-based Reasoning (CBR) model. In this work, several classifiers were constructed first, and the best-performing one was adopted as a feature selection method. After that, these selected features were used to create a case database from previous depression cases. Then a similarity measure based on the Euclidean Distance was employed to retrieve similar cases. This CBR model was evaluated in a dataset of 160 subjects and achieved an accuracy of 91.25%. In [34], EEG signals were used to detect the depression level in which an improved feature extraction method based on Singular Value decomposition (SVD) was proposed, which led to more accurate feature extractions. These features were later classified using

[3] Electrical signals produced by the brain (Electroencephalogram signals).

an SVM model and evaluated on 4 different depression datasets with clinical diagnosis achieving accuracies between 81.9% and 88.1%.

Autism

After depression, Autism Spectrum Disorder (ASD) is the most investigated condition in terms of the automation of the diagnosis process. Several cues have been investigated to help identifying autistic individuals such as facial expressions, gaze, repetitive behaviours, and brain signals. In addition to the detection, several related works are dedicated to evaluating the progress of ASD training and therapy sessions automatically.

Facial expressions were explored for detection purposes in [35], to identify the facial behaviours atypicality of High Functioning Autism (HFA) children in [36], while in [37], it was adopted as part of therapy sessions. In [35], several facial features (facial expressions, action units, arousal, and valance) were learnt by a CNN model. These learnt features were then projected to a lower dimensional representation and fed to a binary classifier for prediction. The CNN model was pre-trained with a large facial dataset and fine-tuned using an ASD dataset. This method was evaluated on a clinically labeled dataset of 88 participants (49 ASD and 39 non-ASD), achieving an F1-score of 0.76.

A different method was adopted in [36], where the authors investigated the differences of facial dynamics between High Functioning Autism (HFA) and Typically Developed (TD) children. Particularly, they measured the complexity of facial dynamics from three different aspects: the amount of facial movements, the repetition of movements, and the correlation between facial regions. They used the Multiple Scale Entropy and Granger Causality model to measure complexity and correlation, respectively. They evaluated their method on a clinically labeled dataset of 39 subjects, 20 are HFA while the remaining 19 are TD. Their most interesting finding is that HFA children tend to show reduced facial complexity, especially when expressing joy, sadness, and disgust. Moreover, they found that the highest complexity region is around the cheeks for HFA, whereas it is around the eyes for TD.

In another study [37], a robot-child interaction system was introduced to evaluate the effectiveness of ASD therapy by measuring the ability of ASD children to recognise and imitate certain facial expressions (happiness, sadness, anger, and fear). First, a facial recognition approach was constructed based on the HOG[4] descriptor and an SVM classifier, which was used by the robot system to recognise children's imitations. Then, certain expressions were performed by the robot and the children were asked to imitate them. The performance of this system was evaluated on three autistic children performing 60 robot interactions, out of which, 31 interactions were imitated and recognised by the system successfully while in 3 interactions, the system failed at recognising the imitations.

Eye gaze was explored in [38] motivated by the fact that gaze behaviours can reveal attention differences between ASD and TD groups. In this study, the gaze data were collected during two web tasks: browsing and searching, in which each participant was shown six web pages with varying visual complexity and asked to perform these tasks. The dataset consists of 15 ASD and 15 TD adults and the best accuracy was 75% for the searching task and 71% for the browsing task.

A video-based detection approach was proposed in [39], in which the authors proposed using spatial attentional bilinear pooling to capture the fine-grained information of video frames which is regarded as an essential element in ASD detection. Specifically, the spatiotemporal feature maps were extracted by a 3D-CNN model which then were pooled using Bilinear Pooling (BP) method coupled with the attention mechanism to preserve the spatial information that is normally lost by the traditional BP method. Finally, the resulting sequences of features were fed to an LSTM model for prediction. This approach achieved 87.2% detection accuracy on an ASD dataset of 40 participants [40].

Other studies investigated detecting ASD behavioural risk markers such as attention in [41], engagement and name call response in [42], gaze and viewing patterns in [43], and repetitive behaviours in [44,45]. In [41], a method was proposed to estimate the fraction of time when ASD and non-ASD groups are paying attention to social and non-social screen stimuli by tracking the direction of their gaze and head position. The main finding was that there is a significant difference between groups in the overall attention, but no difference between attention toward social vs. non-social stimuli was detected i.e., both groups seem to engage with both stimuli in a similar pattern.

Another method was introduced in [42], where the authors were able to detect engagement,

[4] Histogram of Gradients.

name call response, and emotion (Happy vs. other) that matches expert human raters with an ICC>0.80 (Intra-class Correlation Coefficient). They evaluated their method on a dataset of 33 participants (15 ASD and 18 TD), and in their subsequent study [46], they found that ASD children tend to respond fewer times to name calls and take longer time to respond than TD children.

A virtual-reality environment combined with an eye tracker was designed in [43] to detect anxiety in ASD individuals through their gaze and viewing behaviours. In this work, participants were required to listen and interact with several social tasks that are chosen depending on one's performance on previous tasks. By evaluating this method on 16 subjects (8 ASD and 8 TD), they found that the pupil diameter in ASD individuals is greater when they see a non-happy face compared to when they see a happy face. Moreover, ASD individuals are found to blink more when exposed to an angry expression compared to when exposed to a happy one while the opposite is correct for TD individuals.

Repetitive behaviours -one of the main characteristics of ASD behaviours- were explored in [44], where a 3D-CNN was adopted to extract spatial and temporal features from input videos and then a pyramid architecture was built to extract temporal information at different time scales and simultaneously predict the ASD actions and the presence of repetitive behaviours. The dataset consists of 40 hours long of 30 videos and contains the five ASD most common behaviours (hand flipping, head banging, spinning in a circle, toe walking, and moving fingers in front of the eyes). This approach was able to detect the repetitive behaviours with an accuracy of 95%.

Another study [45] proposed regarding the ASD detection problem as a video anomaly detection where a detection model can be trained on unlabelled videos containing normal behaviours which allows for detecting the abnormal behaviours during the inference phase. In this work, a two-stream framework was constructed where the first stream predicts the repetitive behaviours and the second one predicts the pose, and anomaly scores were aggregated from both streams. The dataset contains 75 videos with three ASD behaviours (arm flapping, head banging, and spinning) where 55 videos are used for training excluding any video clips that contain an ASD behaviour. This approach achieved 73.4% AUROC[5] and it is found that the repetitive behaviours detection module plays a significant role in the anomaly detection.

[5] Area Under the ROC Curve.

A cost-effective method using EEG signals was introduced in [47], where different combinations of signal channels and classification methods were explored. Using Correlation-based Feature Selection and Random Forest classifier, they obtained an accuracy of 93% on a dataset of 15 participants (10 ASD and 5 non-ASD). This performance was achieved using only 5 EEG channels, which makes this method easier and more affordable to implement compared to similar previous works. Another study investigated the use of fMRI[6] time series with a deep learning model [48], where the set of features were learnt by an MLP model and fed to an SVM classifier. This model was evaluated on four different datasets with the best accuracy is 80% for 26 participants (12 ASD and 14 Control).

The similarity between brain regions to detect ASD in adults was investigated in [49]. In this work, 7 morphological features of 358 regions of brain images were used to construct a Multi-Feature Network (MFN) for each subject where nodes denote brain regions, and edges denote the similarity between these regions. The properties of these MFNs were then classified using SVM to discriminate between ASD and Control individuals, achieving an accuracy of 78.63% on a dataset of 132 subjects (66 ASD and 66 Control). Similarly, the brain network was explored in [50], where a network of 264 brain regions was constructed where nodes denote regions and edges were defined by the correlation of time-series measurements of these regions. From the network connectivity matrix, 264 eigenvalues of the Laplacian matrix were extracted, and along with other three network features, were then fed to an LDA[7] classifier. This approach was evaluated on a dataset of 871 subjects (403 ASD and 468 Control), yielding an accuracy of 77.7%.

Other Conditions

Several attempts were done to detect other psychiatric conditions such as anxiety [51–53], Post-Traumatic Stress Disorder (PTSD) [54], and Schizophrenia [55]. For anxiety, the authors in [51] investigated the possibility of using audio features to detect anxiety and depression. They proposed a weakly supervised model in which they segmented the audio samples into multiple instances without labelling each segment. More specifically, they constructed audio codewords using GMM, which then are mapped to a dense representation using an NN model. The output vector sequences were then fed to a Bidirectional LSTM for prediction. The model was eval-

[6] Functional Magnetic Resonance Imaging.

[7] The Linear Discriminant Analysis.

uated on a dataset of 105 subjects for anxiety and 142 for depression yielding an F1-score of 90.1% and 85.44%, respectively.

In another study [52], the influence of anxiety on heartbeats was investigated. In this study, 12 heartbeat features were collected through ECG electrodes and then classified by SVM. This approach was evaluated in two anxiety-inducing events (public speaking with 59 participants and thesis defence with 9 participants) where the ground truth labels were obtained from trained audience achieving an F1-score of 0.9. Additionally, the results suggest that the complexity of heartbeats is reduced significantly by higher levels of anxiety. In a similar anxiety-inducing environment, another study [53] explored the relationship between EEG signals and multiple levels of anxiety. They extracted a wide range of EEG features and retained only the ones that contributed to accuracy improvements using the Pearson Correlation method. These retained features were then classified using different methods where SVM gave the best results. This approach was evaluated in a set of only 12 participants with a self-reported questionnaire as labels and achieved 62.5% accuracy.

The influence of Depression and PTSD on vowel production was investigated in [54]. In this work, the authors tracked the first and second formants F1 and F2 of the vowels /i/, /a/, and /u/ using the COVAREP toolbox. Next, they constructed a vowel area for each subject in F1 and F2 space and compared the subjects with respect to these areas. They found that there are significant statistical differences between Depressed and Non-depressed individuals and between PTSD and non-PTSD in the vowel area, using a dataset of 253 subjects along with self-reported labels.

In other work [56], an approach to detect Unipolar and Bipolar Depressive Disorders was proposed using a clinical dataset with six emotion-eliciting videos. Due to the difficulty of collecting and labelling a diverse clinical dataset, a method to generate emotion profiles was adopted. To generate these profiles, the clinical dataset is adapted to another labeled dataset, and then bottleneck features were extracted using a Denoising AutoEncoder (DAE) which was afterward fed to an LSTM model to generate emotion profiles. These profiles were then clustered to codewords using K-means, and then a Latent Affective Structure Model (LASM) was used to map these codewords to a specific disorder class. They achieved an accuracy of 73.33% in a dataset of 45 clinically diagnosed subjects.

Schizophrenia was investigated in [55] where the brain connectivity network was obtained during the MMN[8] process (one of the promising biomarkers of schizophrenia) and fed to a Graph Neural Network (GNN) model for classification. They were able to differentiate between first-episode schizophrenia (40 subjects), chronic schizophrenia (40), and healthy control (40) with an accuracy of 84.17%.

Conclusion

This section reviews the major works that aimed to detect various mental and developmental state conditions. Various behavioural channels were explored with different channels found more discriminant for different conditions, for instance, major advancements were achieved in depression detection by analysing facial and vocal behaviours whereas, for autism detection, risk markers such as repetitive behaviours seem to be the most useful channel. Moreover, the feasibility of fusing different channels was particularly evident in depression detection. Promising performance was achieved in the works that exploited the attention mechanism [19,21,39] to help capture the relevant information to the task at hand, indicating a useful direction for future work in the area.

2.1.2 Computing and Technology in Attachment Research

The literature shows several groups of attachment works where computing and technology play a key role. The majority of these researches were directed towards investigating whether an attachment bond can be formed between users and technological devices and this is desirable for a variety of reasons such as developing responsive social robots [57, 58], advancing the attachment research in psychology [59], and enhancing product design and sustainability [60–62]. Another group of works used technology to enhance the attachment bonds between children and their caregivers [63–65]. A group of research that was introduced in [66, 67] is focusing on modelling the attachment types and behavioural patterns with the strong attractors of the Hopfield Neural Networks. Much earlier works explored whether there is an association between attachment styles and various machine-detectable bio-markers such as fingertips waves [68], ear waves [69], and heart rate variability [70].

Within the first group of works, the authors in [57] explored whether an attachment bond can be formed between users and a social robot depending on the degree of interaction and respon-

[8] Mismatch negativity.

siveness of that robot. They found that there is no increase in the attachment bond regardless of the variations of the robot's capabilities. In [59], the authors investigated how adults interact with a baby robot that exhibits infant-like behaviours with varying degrees of neediness and whether a robot can be developed to elicit an emotional response from human adults. Their robot succeeded at eliciting positive emotions regardless of their degree of neediness, however, the interaction with the less needy robot was found to be less enjoyable. Similarly, the authors in [58] argued that the robots have to be realistic in appearance in order to induce caregivers' attachment, for this reason, they developed a child robot "*affetto*" that has capabilities to show different kinds of facial expressions and affect gestures. A recent study [62] investigated the negative impacts of forming an attachment bond with a social robot and proposed several design guidelines to help mitigate these issues.

The attachment towards mobile phones was investigated in [60] in which a definition of mobile attachment and its design implications were presented. Several design factors were identified in [61] to increase the attachment bond between users and devices and thus increase devices' sustainability. In the same vein, the authors in [71] found that the novelty of the design leads to enhanced sustainability behaviours by increasing the attachment towards technological devices.

The influence of attachment styles in Child-Computer Interaction was explored in [72] concluding that secure children are more consistent during interacting tasks compared to their insecure counterparts which may impact the ability of the latters to operate technology devices which in turn could pose a violation of the design principle of inclusion.

The second group of research focused on employing technology to enhance the attachment bond between children and their caregivers. For instance, in [63], a communication interface was developed that aims to help children invent their own interfaces to make communicating with their distant loved ones more meaningful and entertaining. Similarly, in [64] an application was developed that connects children in divorced households with their parents which helps mitigate the inherent communication challenges in divorced households such as lack of proper engagement and spontaneous contact. A storytelling application was designed in [65] to help foster a connection and a mutual understanding between deaf children and their hearing parents while improving their sign language. This application was specifically directed to deaf children as they face a unique challenge of language development due to the lack of communication with their hearing parents.

Other research that was introduced in [66] proposed modelling the attachment types with the strong attractors of the Hopfield neural networks. The rationale behind their decision is that attachment types as defined by psychiatrists result from repeated patterns of interactions during childhood, therefore, the strong attractors which represent the patterns that are stored in the network multiple times can be used to model the attachment types formation. They demonstrated with simulations that the old strong attractors can be altered when the network is exposed to a new strong attractor. In their subsequent study [67], they introduced the concept of the temperature of the stochastic Hopfield networks thus the model is more likely to resemble the functions of the biological brain. Their findings generally confirmed the results obtained in the previous study despite the increased temperature in the model. Based on these findings, they introduced the self-attachment therapeutic technique [73] where a set of protocols were defined to construct new internal affectional bonds that help regulate emotions and foster a secure attachment with the "inner child".

About three decades earlier, a group of research explored whether there is an association between attachment types and several machine-detectable bio-markers. One of the earliest works in this area [70] found that there is an association between heart rate variability and attachment. In this study, the heart rate in infants (<13 months) was recorded using a recording electrode and it was concluded that heart rate variability is significantly higher in insecure children. A similar finding was reported in [74] where heart rate (HR) was recorded during the Strange Situation Procedure SSP [75] and it was found that disorganised infants have higher HR elevations compared to the secure and avoidant infants. Another study explores the relationship between attachment in adults and skin conductance [76] where they measure the skin conductance score of 50 college students during the Adult Attachment Interview AAI [77] and found an increased level of skin conductance of subjects who employ deactivating attachment strategies when they are asked to recall attachment-related memories such as separation and rejection.

More recent works in this area reported similar findings. For instance, in [68] the fingertip waves were measured using a Photoplethysmography (PPG) sensor, and the Lyapunov exponent was calculated in children aged 0-5 years old and it was found that children who have strong attachment with their mothers have high scores in Lyapunov exponent while those who have moderate to low attachment have lower scores. In a similar manner, in [69] the ear pulse waves were measured in two settings: first, when a child is standing alone (single situation) and second when he is facing another child (pair situation). They found that children with low attachment levels recorded high Lyapunov exponents in pair situation while secure ones recorded lower scores.

This section shows the tremendous role that computing has played recently in advancing attachment research which was tackled from many different angles. Moreover, a substantial number of works reveal that attachment styles leave detectable biomarkers implying the feasibility of exploring ways for the automatic detection of attachment.

2.1.3 Attachment Recognition

This section reviews the studies that were directed toward the problem of attachment recognition which has been addressed only recently in a limited number of works. In particular, attachment recognition was explored for infants [78, 79], for school-age children [80], and for adults [81]. The authors in [78] explored the possibility of recognising attachment in infants (5-8 months old) through visual and audio features. In this study, 64 infants-mother participants were recorded during the Still-Face Paradigm (SFP)[9] when they were 5-8 months old then the attachment labels were obtained through SSP when they were around two years old. Both the visual and audio components of the recordings were fed to classification models to predict the attachment labels. For the visual component, they employed the VGG19 model with soft attention to extract frame-based features which then were fed to an LSTM model, whereas for the audio component, they extracted various emotion-related features such as pitch and MFCC and fed them to an SVM model for classification. The approach was evaluated on a testing set of 18 participants (9 Secure, 9 Insecure) achieving an accuracy of 78%, and 65% for visual and audio components respectively.

In another work concerning attachment in infants [79], the authors explored quantifying the movement behaviours of mothers and infants during the SSP sessions and whether it can be used to predict the expert ratings. To this end, they employed the Kinect sensors which were attached to both mother and infant to measure several behaviours such as contact duration, contact

[9] A paradigm used to examine the infant behaviours under stress conditions.

initiation, and infant velocity toward the mother. They found that there are significant correlations between several attachment behaviours (proximity-seeking, contact-maintenance, resistance, and avoidance) and the measured movement features. Moreover, they applied a stepwise regression model to predict the ratings from these movement features achieving a mean R^2 of 0.56.

The attachment in school-age children was addressed in [80] where 105 school-age children underwent the MCAST test in which they were asked to listen to and complete five attachment-related stories using two dolls to represent the child and his/her mother. The test sessions were automatically administered and recorded through the SAM system [82] (see section 3.3). Subsequently, the authors attempted to build a model that detects the attachment automatically through various hand movement features i.e., the way children move the dolls while interacting with the system. These features include hand positions, the distance between hands, hand speed, acceleration, 1D trajectories, and hand presence. Next, these features were mapped to the attachment labels using a stacked LSTM model achieving an accuracy of 82.8%.

The adult attachment in terms of its two components (anxiety and avoidance) was explored in [81] with respect to 5 different feature modalities (facial expressions, head pose, heart rate variability, linguistics, and paralinguistics). In this work, the participants were exposed to 14 attachment-related stimuli and then asked to express their feelings during which the aforementioned features were extracted. Then, a 3-step approach was implemented where in the first step, a score per modality per stimuli was obtained using linear regression. Subsequently, in the second and third steps, these scores were fused to produce one score per stimuli and per participant, respectively, using a shallow NN. This approach was trained using a dataset of 57 French-speaking participants and validated on a dataset of 19 English-speaking participants, and the results were found to be correlating with the ground-truth labels obtained from AAQ (Adult Attachment Questionnaire) with the correlation coefficient R= 0.79 for attachment anxiety and 0.74 for avoidance. Moreover, they found that facial expressions is the most important modality in detecting attachment anxiety, whereas head pose and paralinguistic are crucial in detecting attachment avoidance.

This section shows that there are few studies that aim at detecting attachment automatically. While certain modalities such as facial expression and paralanguage have been explored for infants and adults leading to reasonable performances, these modalities have not been explored for school-age children. Therefore, it is worth exploring the usefulness of building an automatic approach for this age group based on these modalities.

2.1.4 Discussion

This review shows that there have been major efforts towards building automatic detection approaches for mental and developmental state conditions in the past decade yielding a significant advancement in some cases. Depression, in particular, achieves a remarkable performance by only analysing the non-verbal behaviours which leads to detection approaches that are both non-invasive and more affordable to deploy. On the other hand, studies in attachment recognition (for various age groups) suggest the feasibility of exploiting various behavioural channels to detect attachment. This PhD research aims at detecting the attachment condition automatically in school-age children (5-9 years) based on the MCAST assessment tool. This is attempted by analysing and fusing three behavioural modalities namely, facial expressions, paralanguage, and language.

While this seems like a promising direction to enhance attachment recognition, several challenges may be faced which are unique to the age group of the participants in this study. For instance, various tools that were built for feature extractions are mostly trained on a set of only adult participants and therefore it is unknown how well these tools can perform on children datasets. This may require intensified efforts to tackle this major limitation in the existing literature.

2.2 Background

This section provides background information about the methodologies of our approaches. In particular, Recurrent Neural Network (RNN) and its variant LSTM are described in sections 2.2.1-2.2.2. Next, Deep RNN architecture is illustrated in section 2.2.3. Then, the Convolutional Neural Network (CNN) and its application to Natural Language Processing NLP is described in 2.2.4. Furthermore, the Logistic Regression binary classifier (LR) is described in section 2.2.5. In the final section 2.2.6, several methods are introduced that are often used to create classifier ensembles such as Sum-Rule and Majority Voting.

2.2.1 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are types of Artificial Neural Networks (ANNs) that were built to detect patterns in sequential data i.e., they map a sequence (timesteps) of input vectors X to an output vector or a sequence of output vectors Y where the output can depend on inputs from previous (or following) time steps as opposed to ANNs where each output can depend only on the current input. This is achieved through the recurrence connections (self-connections) of the hidden units which feed the activations from the previous timesteps as input to the current timestep. This augments the RNN network with the ability to learn patterns from past events in the input sequence. Figure 2.2 shows the architecture of a simple RNN with a single hidden layer. Each unit in the hidden layer at time step t computes a representation S_t that summarises the patterns of the past timesteps. Therefore, the output at time t depends not only on the current input, but also captures relevant information from past timesteps.

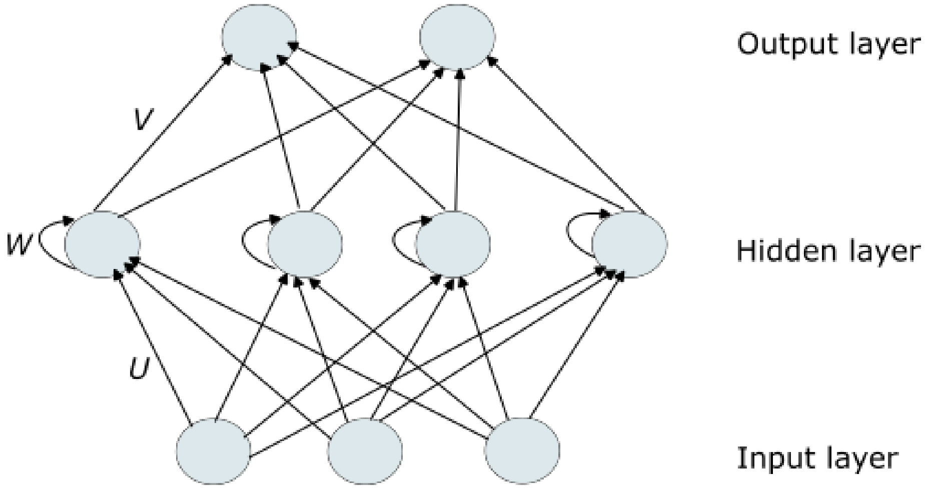


Figure 2.2: **Conventional RNN Architecture.** This figure shows the architecture of the conventional Recurrent Neural Network with one hidden layer. The input layer feeds the input vector of the current timestep through a set of connections which are parametrised by a set of weight U. Additionally, each unit in the hidden layer is fed with the activations from the corresponding unit of the previous timestep (self-connections of the hidden layer units) which is parametrised by a different set of weight W. The hidden units apply linear combinations of these distinct two incoming connections followed by an activation function. The activations from the hidden layer are then fed to the output layer which is parametrised by the weights V.

The RNNs are trained in a supervised learning manner in which the network is presented with a set of (input-output) pairs and the weights vectors (U, W, and V) are optimised by minimising a relevant objective function. The optimisation process of RNNs is similar to the feed-forward networks which involves applying a series of forward and backward passes until the convergence is reached. The key difference in the RNN is that these passes involve computations from multiple timesteps in order to learn the past dependencies. Consider an input sequence of length T, the forward pass involves calculating the activations of the hidden units at layer h at timestep t according to the following equation:

$$h_t = \theta(Ux_t + Wh_{t-1} + b) \tag{2.1}$$

Where θ is a nonlinear activation function such as the hyperbolic tangent activation function (tanh), x_t is the input vector at timestep t, h_{t-1} is the activations from the hidden layer of the previous timestep. U, W, and b are the input weight, recurrence weight, and bias vectors of the hidden layer respectively. For a single input sequence, the equation 2.1 is repeated over the full sequence by incrementing t starting from t=1 until t=T by feeding the recently evaluated h_{t-1} at each timestep.

Depending on the task, the output can be either computed simultaneously at each timestep or, as in most sequence classification problems, the output is computed after evaluating the full input sequence in which the activation of the last timestep h_T is passed to the output layer which computes the output o_T from:

$$a_T = Vh_T + c \tag{2.2}$$

$$o_T = \theta(a_T) \tag{2.3}$$

Where V and c are weight and bias vector parameters. For classification problems, θ is typically a softmax function that converts the linear combinations obtained by the output units to values in the range [0,1] which can be interpreted as probabilities of a certain input belonging to a certain class $P(C_k|x)$. In multiple classification problems, the target is coded using 1 of K vector representation where all elements of the vector are zero except the element that corresponds to the true class (set to one). Thus, the softmax activation function assigns an output probability value to a class k according to the following equation:

$$P(C_k|x) = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}} \tag{2.4}$$

where K is the number of the output units.

The objective function can be estimated either from one input sample, multiple samples, or the full training set depending on the mode of learning. Many optimisation algorithms apply a Stochastic Gradient Descent (SGD) method which estimates the error over a random subset of the training set. This method is proven efficient as computing the error function over the entire training set can be computationally expensive. One of the most applied optimisation algorithms is *Adam* optimiser [83] which is an adaptive learning rate optimisation algorithm that adapts the learning rate of the parameters to be proportional to their derivatives.

The backward pass involves calculating the derivatives of the error function with respect to the weights parameters in the network in order to update these weights to minimise the overall error in the next training step. One of the main algorithms that were derived to calculate the derivatives efficiently is the Back-propagation Through Time (BPTT) algorithm [84] which extends over the standard back-propagation algorithm by taking into account the influence of the previous timesteps to the current output error function. This means that for each unit at the hidden layer, we need to compute the gradients by recursively accumulating the gradients of the units that follow it (either output units or next-timestep's hidden units). In this way, all the weights are adjusted through their influence on both output units at time t and hidden units at time $t+1$. This is implemented efficiently by applying the chain rule for partial derivatives starting from the output layer and proceeding to the input layer. The derivative of the error function $L(x,y)$ with respect to w_{ij} parameter is given by:

$$\frac{\partial L(x,y)}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial L(x,y)}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} \tag{2.5}$$

which is evaluated over all the timesteps starting from $t = T$. The first term then can be evaluated recursively for a single timestep as follows:

$$\frac{\partial L(x,y)}{\partial a_j^t} = \theta'(a_j^t)(\sum_{k=1}^{K} \delta_k^t w_{jk} + \sum_{h=1}^{H} \delta_h^{t+1} w_{jh}) \tag{2.6}$$

where

$$\delta_j^t = \frac{\partial L(x,y)}{\partial a_j^t}, \tag{2.7}$$

and δ_k^t is the delta terms computed over output units a_k, and δ_h^t is the delta terms computed over the hidden units a_h. H and K are the number of hidden units and output units, respectively.

The standard RNN has been proven effective in a wide range of applications that require sequential processing such as speech recognition and text generation. However, its effectiveness is largely determined by the length of the temporal dependencies present in the input sequences. When these dependencies span over a short timestep sequence, it can be captured effectively through the standard RNN. However, as these dependencies get longer (more than 5-10 timesteps [85]), the learning capabilities deteriorate considerably.

This problem essentially arises from having to propagate the derivatives of the error function over a large number of timesteps and it has been known in the literature as vanishing and exploding gradients [86]. When the gradients are too small the weight parameters are adjusted by small quantities which slow the learning process of the network. Conversely, when the gradients are too large the update causes the weight parameters to oscillate rapidly and make it harder to converge to the optimal parameters. A number of solutions have been proposed to tackle this problem and the most notable solution is LSTM [87] which is considered to be the most effective solution to the problem (under appropriate parameterisation) [88] and it will be described more in-depth in the next section.

Various other variants of the conventional RNN architecture have been proposed to tackle various types of problems e.g Bidirectional RNN [89] which allows to learn dependencies from future input timesteps as well the past timesteps which are proven effective in contexts such as machine translations. Another variant is Echo State Network (ESN) [90] which is proposed to tackle the vanishing and exploding gradients by reducing the number of trainable connections.

2.2.2 Long Short Term Memory (LSTM)

As mentioned above, the LSTM architecture was proposed to tackle the main limitation of the standard RNN which is known as the vanishing and exploding gradients problem, and therefore to enhance the network's ability to capture dependencies that span over longer sequences. The proposed architecture of LSTM is based on constraining the gradient flow to a constant value so it neither vanishes nor explodes which is achieved through the unit's self-connections and by further introducing the input and output multiplicative gate units to control the content of the network. Figure 2.3 shows the architecture of one unit (memory cell) of an LSTM network.

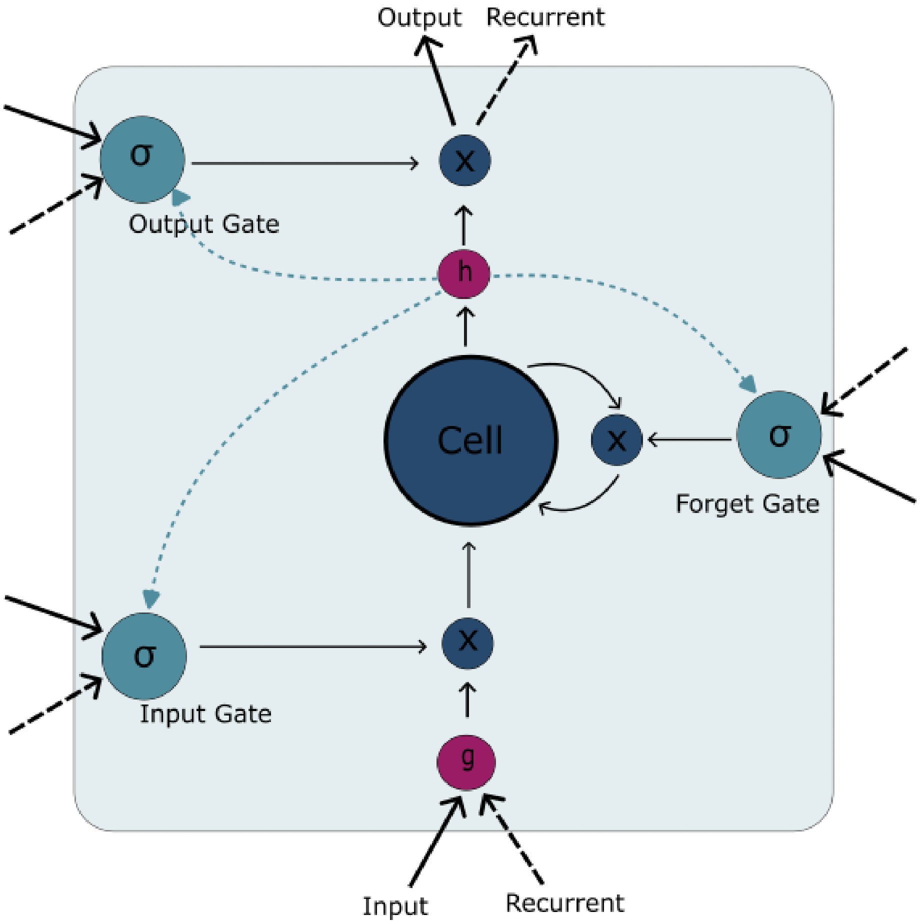


Figure 2.3: **The Architecture of an LSTM Memory Cell.** This figure shows the architecture of one memory cell of an LSTM network. Each unit receives input and recurrent connections shown in solid and dashed black arrows. The dark blue circle represents the memory cell where the summation of the incoming inputs from other cells is calculated. The activations in g and h can be either *sigmoid* or *tanh* functions. The green circles indicate the multiplicative gates where each gate takes the input and the recurrent connections. The activations in these gates are usually the *sigmoid* function. The green dashed connections show the peephole through which each gate can inspect the content of the cell at the current timestep. The incoming input to the cell is controlled by the input gate and similarly, the output is controlled by the output gate. The forget gate controls the content of the cell.

In this architecture, the hidden layer consists of a set of memory cells and the main difference between an ordinary summation unit and a memory cell is the multiplicative gates (input, output, and forget gates). The role of the input gate unit is to protect the content of a memory cell from irrelevant input rather than updating the hidden representation at each timestep as in the conventional RNN, thus allowing it to store content for a longer period of time. Similarly, the output gate prevents perturbing the network output with the irrelevant contents of the memory cell. An additional forget gate was proposed by [85] in which the network is augmented with the ability to learn to reset (forget) the content of the memory cells once it becomes irrelevant. The significance of the forget gate was shown in [91] where the authors found that the forget gate along with the output activation function are the most essential parts of the LSTM architecture and removing them from a memory cell affects the performance significantly. In [92], ”peephole” connections were introduced into memory cells which are weighted connections between the gates and the cells that allow these various multiplicative gates to inspect the content of the cell (even when the output gate is closed) instead of basing their decisions only on the cell output of the previous timestep.

According to [85, 87], LSTM architecture has the ability to capture patterns with dependencies that span over 1000 timesteps which is a tremendous improvement over the standard RNN architecture. However, this architecture can only mitigate the vanishing gradients problem [93] and therefore LSTM network may still suffer from the exploding gradients [86] although the exploding problem rarely occurs in practice [88].

2.2.3 Deep Recurrent Neural Networks

Deep learning models are developed to capture more complex functions compared to traditional shallow machine learning algorithms by having multiple layers of linear and nonlinear units that can learn higher-level representations. Deep models have successfully achieved state-of-the-art performances in many classification and clustering problems.

The same motivation of having the ability to learn more complex functions to learn higher-level representations is extended to the RNNs. However, the concept of depth is ambiguous in the case of RNNs because these models are already deep when unfolded in time i.e., hidden layer from the previous timestep feed to the hidden layer of the current timestep through the recurrent connections. Few attempts have been found in the literature [94, 95] that proposed a deep RNN architecture. In [95], the authors explored several architectures of constructing deeper RNNs by adding extra computational units in various transition connections and they found that these deeper architectures achieve better performances in major sequence classification problems.

One of their proposed architectures is based on stacking multiple levels of recurrent layers where each level operates on (takes as input) the hidden representation at the lower level of layers which encourages the top levels to operate at different timescales. Figure 2.4 depicts the architecture of the stacked RNN with two LSTM hidden layers unfolded in time. The core concept is that the top level of layers z takes the hidden representation of lower-level layers h as input to learn higher-level representations.

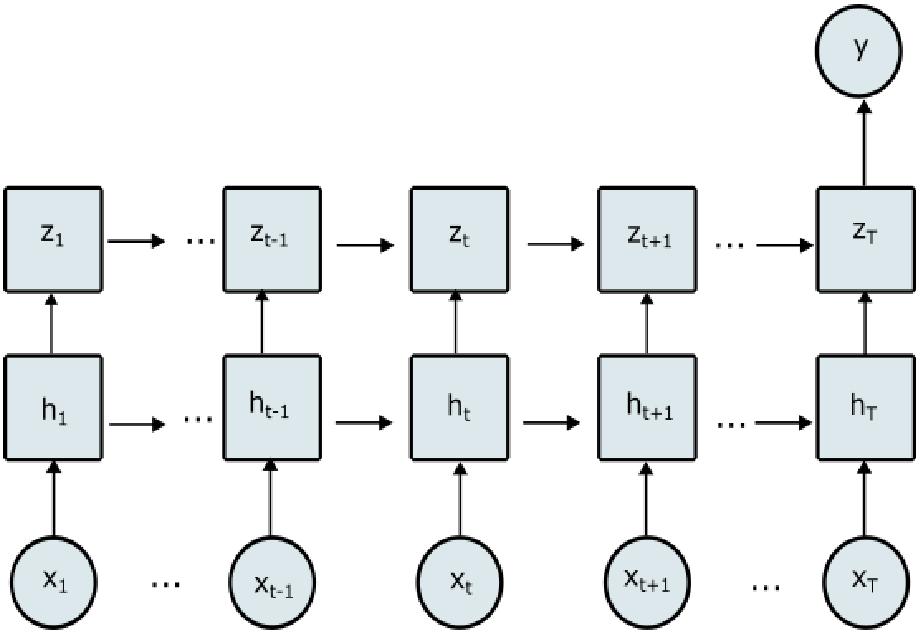


Figure 2.4: **Stacked LSTM.** This figure shows the structure of the 2-stacked LSTM for classification problems. Boxes correspond to hidden layers and circles denote input and output layers. The sequence of T input vectors (X) is fed to the first row of hidden layers h where each layer learns a representation h_t that depends on the current input x_t and the previous activation h_{t-1}. Each layer in the second row z learns a higher level representation z_t that depends on the lower representation h_t and the previous activation z_{t-1}. The output layer y takes as input the final hidden representation hidden layer z_T which summarises the full sequences and applies the softmax activation function for prediction.

2.2.4 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) have yielded remarkable breakthroughs over the past years in the fields of computer vision and natural language processing [96–98]. Inspired by the biological vision systems, the core process of the CNN is the *convolution* operation which is a mathematical operation that refers to combining two functions to produce another output function. More specifically, CNN convolves a set of learnable weights (called filters or kernels) with local patches of the input data in order to extract meaningful patterns for the task at hand. Depending on the shape of the data and the underlying task, filters in CNN can take various dimensional shapes for example, in image processing tasks, filters often take 2-Dimensional shapes whereas in time series signals filters can have 1-Dimensional shapes.

The appealing property of using the convolution operation is that these learnable weights are shared across the full input representation meaning that instead of assigning a separate weight for each input element as in conventional feedforward networks, CNN used the same set of weights across the input data. This reduces the computational complexity considerably depending on the size of filters relative to the input data size. The shared weights also allow to detect the similar patterns over the full input for example, in image processing one filter can detect edges over the full input image.

By training an adequate number of filters over the same input, a wide range of patterns can be detected which enhances the learning capabilities considerably compared to traditional feature extraction methods. These capabilities can be further enhanced by adding several layers of convolving filters that detect higher levels of representations.

CNN in NLP

Recently, several studies proposed applying the CNN architecture for text modelling in various NLP problems such as text classification [99, 100], sentence modelling [101], and other automatic text-based tasks [102]. In [99], the authors showed that one layer of convolution on top of pre-trained word embeddings *word2vec* has achieved a remarkable performance on various sentence classification tasks compared to conventional methods. Additionally, a very deep CNN was proposed in [100] in which the best performance was achieved with a deep model of 29 convolutional layers and a max pooling layer, improving over the state-of-the-art on several text classification tasks.

One of the key elements in text modelling is the choice of word representation methods which can have a significant effect on the overall performance and the complexity of the approach. Several word representation methods have been proposed in the literature such as Bag-Of-Words, One-Hot encoding, and word embeddings. Word embeddings have the property of projecting words' high-dimensional sparse representation into a dense fixed-length representation with lower dimensions. Additionally, embeddings can capture the semantics of words by projecting words with similar meanings closer to each other in the new space which makes them feature extractors in themselves [99].

There are two main ways to learn these embedding representations: they can be learned jointly from scratch over the dataset of the specific task where in this case, an embedding layer is cascaded on top of the classification model and trained in an end-to-end manner. Instead, pre-trained embeddings can be used which are trained over a rich text dataset and it can be also fine-tuned with the task dataset. The most commonly used pre-trained word embeddings are word2vec [103], and glove [104]. Learning the embeddings from scratch for a specific task can have the advantage of better capturing the semantics that are specific to the dataset.

Another significant advantage of using word embeddings is that it produces an image-like representation which makes it possible to be processed by a CNN model. The standard architecture of such a model is depicted in Figure 2.5. The input embeddings of size $n*m$ are convolved with k filters. The size of the filter matrices in NLP is usually set to have the same width as the embeddings m whereas the height r is set as a hyper-parameter. This makes filters move in one direction hence, it captures word relationships, as each word is represented in one row in the input embeddings. Moreover, filters can also have varying sizes which enables capturing multiple types of features. Each convolution operation results in a real-valued scalar and the full pass of one filter over the input representation results in a feature map vector of size $(n-r+1)$[10].

By convolving an adequate number of filters, rich information can be extracted which often can be high dimensional. Therefore, pooling methods can downsample the feature maps while preserving the most salient features such as max pooling and average pooling. Higher-level representation can be further learned using an additional dense layer and a softmax layer is used in classification problems to assign a probability value to each class.

[10] For valid padding and a Stride=1.

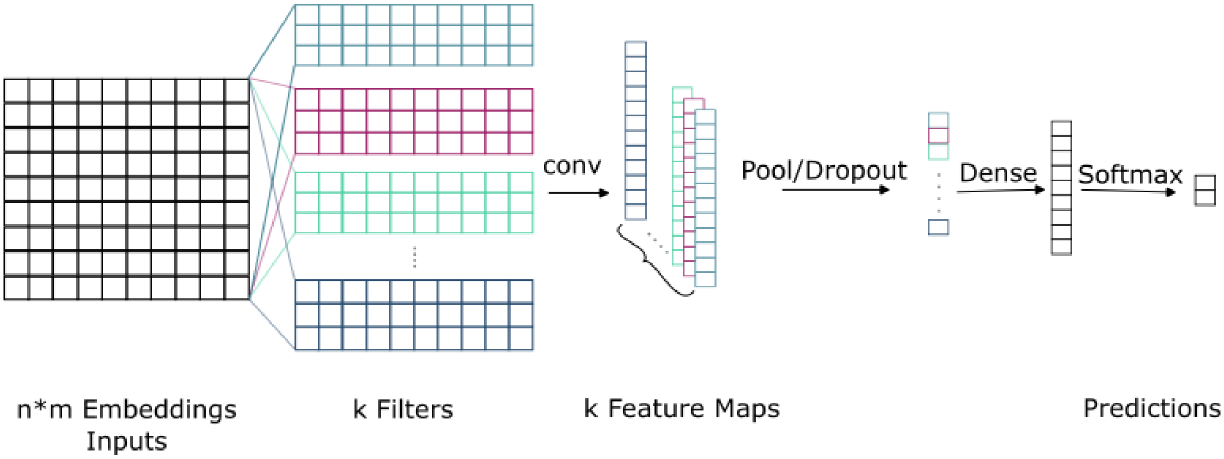


Figure 2.5: **CNN Standard Architecture for NLP Problems.** This figure shows the architecture of the standard one-layer CNN for text classification problems. The input text is represented by $n*m$ embeddings where n represents the number of tokens in the input text and m represents the dimensions of the embedding representation. Then, the embedding is convolved with k filters in which the size of the filters is set to $r*m$ where r is the filter size and it is set as a hyperparameter and m is the size of the embeddings. Thus, the filters are moving in one direction hence it is called 1-Dimensional CNN. Convolving all k filters results in a set of k vectors called feature maps. Feature maps can often be high dimensional which can be reduced using a pooling strategy followed by a dropout to enhance the generalisation. A dense layer is applied to learn a higher-level representation which then can be classified with a softmax layer.

2.2.5 Logistic Regression (LR)

Logistic regression (LR) is a probabilistic machine learning model widely used for classification problems. Its output can be interpreted as estimating the probability of a given input x belonging to one of the classes, for example, a binary logistic regression model estimates the quantity $p(y=1|x)$ which is given by:

$$p(y=1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^{\mathbf{T}}\mathbf{x}+b)} \tag{2.8}$$

where $\mathbf{x}$ is the input vector, $\mathbf{w}$ and b are weight and bias vectors, respectively. The quantity $\sigma(a) = \frac{1}{1+\exp(a)}$ is a sigmoid function that serves as a squashing function that converts a real value input a to a value bounded between 0 and 1. The parameters $\mathbf{w}$ and b are estimated by optimising an objective function corresponding to maximising the negative log-likelihood (over the N data points) which gives rise to the cross entropy loss function given by:

$$L(y,\hat{y}) = -\log p(y|x) = -\sum_{n=1}^{N}\{y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\} \tag{2.9}$$

where $\hat{y}_n = \sigma(-w^T x_n + b)$ and y_n is the true labels. This function can be optimised with an iterative method based on Newton-Raphson optimisation techniques such as the L-BFGS[11] method. A regularisation term such as L1 and L2 norms can be added to the above loss in order to improve the generalisation of the LR model and a hyper-parameter λ has to be predefined to specify the strength of the regularisation effect. Thus, the L2-norm regularised loss function is given by:

$$L(y, \hat{y}) = -\log p(y|x) = -\sum_{n=1}^{N} \{y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)\} + \frac{\lambda}{2} ||\mathbf{w}||^2 \tag{2.10}$$

2.2.6 Ensemble Learning

Ensemble learning in classification problems aims at combining decisions made by multiple classifiers trained individually over a pattern X to enhance the performance of the individual classifiers. These multiple classifiers can be obtained by training on different representations of the same input pattern X, using different classification methods or weights of the same X representations, or trained over different training sets [105]. Ensemble methods are only expected to lead to better performance than its individual members if those members perform well individually and if they are diverse or statistically independent meaning that they make different mistakes over different input samples [106, 107]. This is possible because in this case different classifiers may offer complementary information about the same pattern which in turn lead to better identification rate. If the members are not diverse it means that it is very likely that they will agree on the same mistakes which in turn results in an ensemble with similar performance as the members. Several methods have been suggested in the literature to ensure that the ensemble of classifiers is as diverse as possible such as bagging and boosting [107].

Given multiple classifiers, various methods have been proposed to create classifier ensemble such as Product Rule, Sum Rule, Mean Rule, and Majority Voting which are examples of non-trainable combiners. According to [107], the simplest method which is based on aggregating the votes made by the ensemble members (i.e., Majority Voting) can lead to performance higher than the recognition rate of an individual classifier. Moreover, they proved that, under the statistical independence assumption, increasing the members of the ensemble can decrease the empirical

[11] Limited-memory (Broyden–Fletcher–Goldfarb–Shanno algorithm).

mean squared error MSE.

Kittler et al. [108] showed that using the above combining methods can be viewed as building a single compound classifier $F(X)$ in which its class posterior probability can be jointly estimated from applying the above methods to the estimations made by individual members $f(x_1)....f(x_R)$. Under various assumptions and approximations, they were able to derive these combining rules and further, they constructed a framework to derive other possible combining rules. Moreover, they showed experimentally that Sum Rule outperforms other rules, although it was developed under the most restrictive assumption (see below), due to its resilience to the estimation error of the class posterior probability.

In the following, a formal brief description of the three most commonly used combining rules is presented: given an input pattern X, the goal is to assign X to one of m possible classes $(y_1...y_m)$. Number of classifiers R aim to assign X to target class by representing X using a different representation x_i i.e., each individual classifier $f(x_i)$ estimates the j^{th} class posterior probability over its own input representation $p(y_j|x_i)$. Thus, the ensemble function $F(X)$ is modelling posterior probability given by $p(y_k|x_1...,x_R)$.

Product Rule

By assuming that different representations of X are conditionally statistically independent, the product rule estimates the posterior probability by multiplying the posterior probability ”support” made by individual classifiers $f(x_i)$ and assigns the input pattern to the class with the highest support. More formally, the product rule assigns a pattern X to a class y_k if

$$\prod_{i=1}^{R} p(y_k|x_i) = max_{j=1}^{m} \prod_{i=1}^{R} p(y_j|x_i) \tag{2.11}$$

It can be observed that this rule is sensitive to near zero values [109] as it is sufficient for one classifier to assign zero to the true class for the ensemble to make the wrong decision therefore, this can lead to poor performance in practice compared to other rules.

Sum Rule (SR)

Sum Rule is one of the most widely used classifiers combiners along with majority voting rule (discussed below). It estimates the joint posterior of the ensemble by summing the support to each class made at each individual classifier i.e., it assigns a pattern X to class y_k if

$$\sum_{i=1}^{R} p(y_k|x_i) = max_{j=1}^{m} \sum_{i=1}^{R} p(y_j|x_i) \tag{2.12}$$

According to [108], SR is derived under the assumption that the posterior probability deviates from the prior only by a small fraction meaning that $P(y_k|x_i) = P(y_k)(1+\delta_{ki})$. While this is a very strong assumption, it can be readily satisfied especially in domains where the datasets are highly noisy and ambiguous.

Majority Voting (MV)

This method aims at combining the decisions made at each classifier rather than the posterior probabilities where each component of the ensemble votes for a particular class and the ensemble decision is made by choosing the class with the highest number of votes. More formally, MV estimates the joint posterior probability of the ensemble by approximating the component posterior probability to produce a binary values according to the following formula:

$$\Delta_{ij} = \begin{cases} 1 & \text{if } p(y_j|x_i) = \max_{k=1}^{m} p(y_k|x_i), \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

Subsequently, MV assigns X to class y_k if:

$$\sum_{i=1}^{R} \Delta_{ik} = max_{j=1}^{m} \sum_{i=1}^{R} \Delta_{ij} \tag{2.14}$$

Chapter 3

Attachment Assessment and Data

3.1 Overview

This chapter describes the attachment assessment tool and the dataset that was adopted for the experiments. The assessment tool is based on the Manchester Child Attachment Story Task (MCAST) [11], the standard attachment assessment tool for school age children which is based on a vignette-completion and doll-playing method. This tool has been previously automatically administered through the School Attachment Monitor (SAM) [82] in which the MCAST is fully administered and recorded automatically prior to the assessment process. In 87% of the cases, the system was capable of administrating the interview sessions and capturing recordings that were informative enough for human assessors to be able to rate these recordings in terms of the attachment styles. This constitutes a major element of the automation process where the workload of the administration process is reduced considerably.

Several attachment assessment tools have been proposed previously in the literature such as Strange Situation procedure (SSP) for infants [4], Attachment Doll-Play Interview [10], separation-reunion story completion task [110] for preschoolers (3-6 years old), and Adult Attachment Interview (AAI) [77] for adults. The common element between such tools that were developed for pre-schoolers is that they are all based on introducing children to stressful doll-play scenarios to elicit various attachment-related behaviours. MCAST complements these previous tools with the aim to develop an attachment assessment tool in young school-age children (5-9 years old) by relying on similar vignette-completion and doll-playing methodologies which are believed to be the most appropriate eliciting tools for this age group.

The rest of this chapter is organised as follows: first, a description of the MCAST is presented in section 3.2, then, in section 3.3, SAM system will be described. Next, the description of the participants is presented in section 3.4. Lastly, section 3.5 describes the recordings and presents statistical information about the dataset.

3.2 Manchester Child Attachment Story Task (MCAST)

The MCAST [11] is designed to assess the attachment styles in school-age children (5-9 years old) and is based on a vignette-completion and doll-playing method. Children are introduced to a playing doll house and provided with two dolls and asked to choose one doll to represent the child and the other to represent the primary caregiver (typically the mother). Then, they are asked to listen to five different scenarios in which the child is faced with varying degrees of stressful situations while the caregiver is close but not proximate. These scenarios are chosen to be age appropriate and are expected to elicit attachment-related behaviours. This allows the children to exhibit proximity-seeking behaviours through the way they complete the story, play with dolls, and describe their feelings.

The five scenarios are as follows:

1. Breakfast (BF): the mother wakes the child up in the morning while she is preparing breakfast.

2. Nightmare (NM): the child awakes at night alone after having a nightmare and she calls her mother for comfort.

3. Tummy-ache (TA): the child was sitting at home watching TV and felt abdominal pain and asked the mother for assistance.

4. Hopscotch (HS): the child was playing outside when she fell and hurt her knee.

5. Shopping Mall (SM): the child finds herself lost at a shopping mall and tries to establish contact with her mother again.

The first story (BF) is intended to act as an introduction to the task where no distress or conflict is being conveyed. However, it might still induce attachment related behaviours [11]. The remaining stories are chosen so they convey different distress situations which may invoke different attachment related behaviour to allow for a fine grain analysis of the IWMs.

Each scenario consisted of two consecutive phases, in the first phase (induction phase), the interviewer describes the scenario while amplifying the intensity of the distress situation to the point where the child is clearly engaged with the story. In the second phase (completion phase), the interviewer asks the child to complete the story using the dolls. After completing the story, the interviewer asks the child to describe how the child and the mother feels: 'Can you tell me how the child/parent doll is feeling now?'. This allows one to provide more clarification about intentions and feelings behind the story completions. The interview takes between 20 to 30 minutes to administer and between one to two hours to assess depending on the complexity of the materials.

This instrument was validated on a sample of 53 participants achieving an inter-rater agreement at distinguishing between secure vs insecure of 94%. Several behaviours were found more predictive than others such as proximity seeking, degree of assuagement, engagement in the interview, and narrative coherence [11]. Moreover, the MCAST shows concurrent validity against other well-validated attachment tools such as AAI [111]

3.3 School Attachment Monitor (SAM)

Attempts to automate the MCAST administration process have been made through Computerised MCAST (CMCAST) [112], and School Attachment Monitor (SAM) [12, 82]. This eliminates the need for trained professionals to conduct these interviews and therefore reduces the workload of health practitioners considerably as it typically takes between 20 to 30 minutes per child to conduct these assessment interviews. Figure 3.1 shows the setup of the SAM system. As shown in the figure, the system consists of five main elements: a computer screen, a button to signal the end of each stage of the interview, two dolls to represent the child and the caregiver, and a toy house with the main furniture in which children act out the story completions. The setup is also equipped with a camera to record the upper body of children while interacting with the system. The interview session is conducted in the following consecutive steps:

1. The participant watches a video on the screen where an actor shows the vignette and asks the participant to complete the story using dolls;

2. The participant completes the story using the dolls and presses the button to signal the end of the story;

3. Pressing the button in the previous step activates another video where the actor asks the participant to describe how the child doll feels;

4. The participant starts explaining how the child of the story feels and presses the button to signal the end of the task;

5. A video is activated where the actor asks the participant to describe how the caregiver doll of the story feels;

6. The participant explains how the caregiver doll feels and presses the button to signal the end of the task;

7. The system starts again at step 1 for the next vignette until all five vignettes have been presented to the participant.

This system was validated on a dataset of 120 participants to administer the MCAST [80], where each child was rated as secure, insecure, or non-assessable[1] by a pool of 4 trained assessors and in the case of disagreement between assessors, consensus has been reached through discussions. The results show that the attachment style is identified in 104 out of 120 meaning that SAM was able to administer MCAST with a success rate of 86.7%. In other words, the system succeeded at administering the MCAST by collecting enough materials for the assessors to be able to rate the recordings in 104 out of 120 cases.

3.4 Participants

The total number of participants in the experiments is 104 whose assessment interviews were successfully administered through SAM. The age range of the participants is 5-9 years and they were recruited from primary schools in Glasgow [80]. Table 3.1 shows the distribution of

[1] Corresponding to the cases in which the recorded materials did not provide enough information for the assessment.

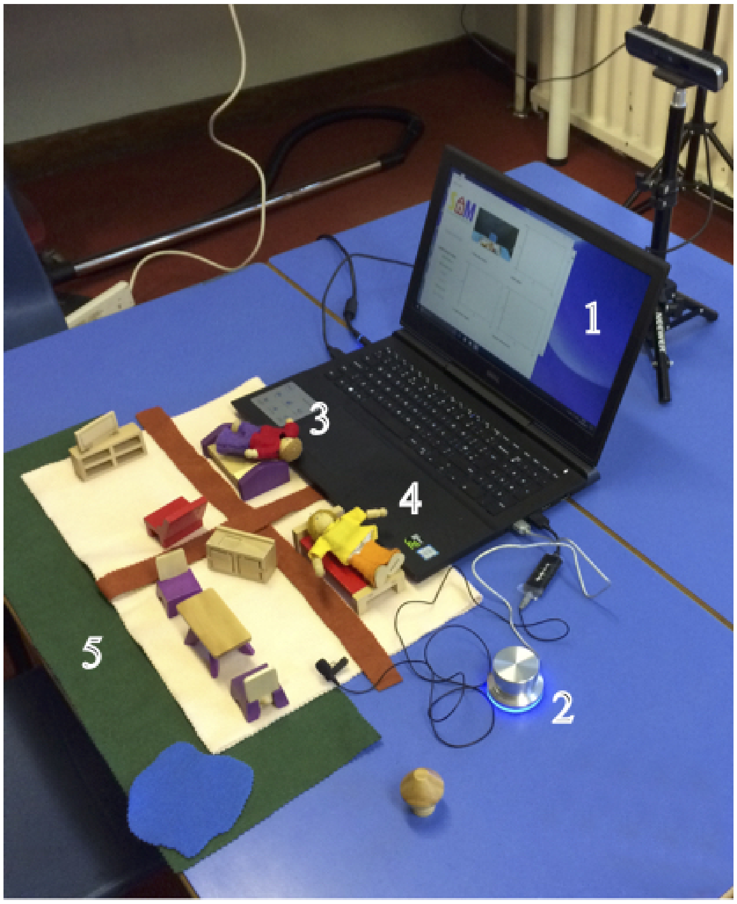


Figure 3.1: **Setup of SAM System.** The figure shows the setup of the SAM system. Element 1: a computer screen where the actor presents the story stems and guides the children through the system's various stages. Element 2: a button that participants are instructed to press to signal the end of each stage. Element 3, 4: two dolls to represent the child and the caregiver. Element 5: a toy house with main furniture. In the back, a camera is set to capture the upper body of the participants while interacting with the system.

participants across school levels, gender groups, and attachment styles. According to χ^2 test, the attachment distribution in the participants is within a statistical fluctuation with respect to the attachment distribution in the general population [13, 14]. The attachment labels were made by a pool of four trained assessors in which each participant was assessed independently by two random members from the pool. In the case where there is an agreement between the assessors, the decision is accepted, otherwise, all the members in the pool discuss the case to reach a consensual decision.

As shown in the table, there are more children from levels P2 and P3 which account for two-thirds of total number of participants. In addition, the fraction of insecure children in levels P1 and P2 is higher than other levels. Regarding gender distribution, the percentage of female and

Level	P1 (5-6)	P2 (6-7)	P3 (7-8)	P4 (8-9)	Total
Female	9	22	15	11	57
Male	10	18	14	5	47
Secure	9	22	18	10	59
Insecure	10	18	11	6	45
Total	19	40	29	16	104

Table 3.1: **Participants Distribution.** Participants distribution across school levels. The numbers between parentheses indicate the age range of the corresponding school level. For each level, gender and attachment styles distributions are also provided.

male participants is 55% and 45%, respectively, with the biggest variations appears in level P4. According to χ^2 test, such a distribution is the same as in the general population.

3.5 Data

The data is comprised of video recordings that were collected using SAM system. The total length of the recordings is 18 hours, 30 minutes, and 34 seconds with the average length being 640.7 ± 273.5 seconds. The video length varies across participants as shown in Figure 3.2. The average video length for the secure group is 650.2 ± 232.7 whereas it is 628.3 ± 316.2 in the insecure group and according to two-tailed t-test, there is no statistically significant difference in video length between the attachment groups. The video resolution is 1920 x 1080 pixels, the frame rate is 30 frames per second, and the sampling rate of the audio stream is 44100 Hz.

The average length of the recordings that corresponds to each story is shown in Table 3.2. As shown in the table, The induction phase varies between stories with SM having the longest induction phase. Moreover, in the completion phase, the BF is longer, on average, than other stories to a statistically significant extent ($p < 0.05$) whereas there are no significant differences among other stories. A possible reason for this difference is that BF is the first story in the interview and therefore it might take longer for children to familiarise themselves with how to interact with the system.

For the language-based classification approach, the interviews were automatically transcribed using the online transcription tool *sonix*[2] resulting in a total number of words in the dataset

[2] https://sonix.ai

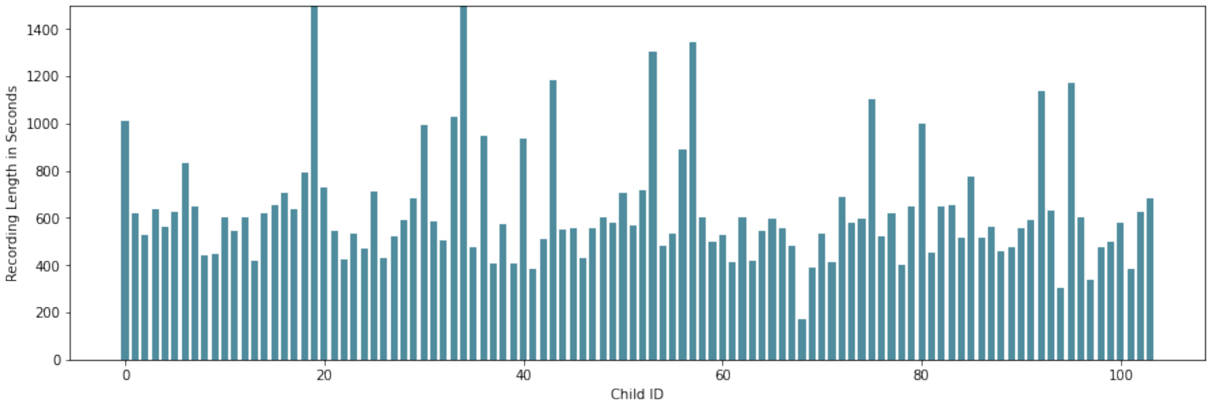


Figure 3.2: **Recording Length Distribution.** The figure shows the length of the full recordings for all participants in the dataset in seconds. For clarity, two samples whose lengths are above 1500 are truncated and the corresponding lengths are 1942 *secs* and 1822 *secs* .

Story	BF	NM	TA	HS	SM
Induction Phase	24	33	36	36	78
Completion Phase	113.0 ± 78.3	84.6 ± 73.6	75.3 ± 58.2	80.4 ± 66.4	80.5 ± 78.5
Total Length	137.0 ± 78.3	117.6 ± 73.6	111.3 ± 58.2	116.3 ± 66.4	158.5 ± 78.5
Amount of Words	115.6 ± 124	79.3 ± 111.1	69.3 ± 78.8	79.1 ± 94.2	84.5 ± 143.6

Table 3.2: **Average Recording Length and Amount of Words per Story.** This table shows the average recording length measured in seconds and the amount of words per story. The length of each one of the two phases is also provided.

of $43,725$, where the average number of words per participant is 420.4 ± 443.2. The average number of words per secure and insecure group is 474.8 ± 392.3 and 349.1 ± 488.7, respectively. According to t-test, there is no statistically significant difference between attachment groups concerning the amount of words. Additionally, the number of words per each story is shown in Table 3.2, according to t-test, The amount of words is higher in BF compared to the three lowest stories (NM, TA, and HS) to a statistically significant difference ($p < 0.05$).

3.6 Chapter Summary

This chapter described the MCAST assessment tool along with the automatic administration system SAM that was adopted to collect the dataset of this research. A description of the participants and distribution information is also provided. The dataset is also described along with various statistical information in terms of the recording lengths and the amount of words.

Chapter 4

Unimodal Approaches

4.1 Overview

This chapter presents the three unimodal approaches used in the experiments. For each approach, the methodology is described and the results are presented. These approaches are based on three different behavioural channels, namely: facial expressions, paralanguage, and language. These behavioural channels are considered to be the main channels that people use to express their emotions and inner states and therefore it is worthwhile exploring what these channels can convey about the attachment in children. Most of the results in this chapter (and the next chapter) were previously presented and published in major venues relevant to this work [113–115].

The following three sections will be dedicated to describe each approach individually. Each approach addresses a classification problem where the video recordings are mapped to one of the two attachment classes (negative class: secure attachment, and positive class: insecure attachment). The key steps in each approach are extracting a relevant set of features and building a classification model that can learn patterns from these extracted features. Moreover, the effect of gender and age variations in the performance are also presented which can provide insights on how children with different attachment styles manifest their condition.

The section that follows provides comparisons between these three approaches concerning overall performance, performance obtained at each MCAST story, and performance for specific age and gender groups. Various conclusions have been drawn from these comparisons which can provide insights and suggestions on how to improve the attachment recognition. For instance, one approach was found to work better for children from lower age groups while a different one works better for children from greater age groups indicating the possibility of adopting different approaches for different age groups.

To assess the effectiveness of the approaches i.e., whether the approaches are capturing attachment-relevant cues, the results were compared to the random guess classifier which assigns samples to class c according to a priori probability $p(c)$, such a classifier would have an accuracy that can be computed from $\sum_{c \in c_0, c_1} p(c)^2$ (for binary classifications). Due to the imbalance between attachment groups in our dataset (and in the general population), additional performance metrics are also reported in the results to account for this imbalance i.e., to provide a better idea of how the approaches perform at individual classes. These metrics include precision, recall, and F1-score with their random guess performance can also be derived from classes prior probabilities $p(c)$. More specifically, the number of true positive TP and true negative TN predictions of the random guess classifier can be derived from:

$$TP = n \times p(c_1)^2$$
$$TN = n \times p(c_0)^2$$

where n is the total number of samples in the dataset, $p(c_1)$ and $p(c_0)$ are the prior probabilities of the positive class and negative class, respectively. Accordingly, the number of false negatives FN and false positive FP predictions can be calculated from:

$$FN = P - TP$$
$$FP = N - TN$$

where P and N is the number of positive and negative samples in the dataset. This allows to calculate the three above metrics of the random guess classifier from:

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.2 Facial Expressions Based Approach

The first unimodal approach is based on analysing facial expressions and how are they related to differences in attachment styles. Facial expressions have been long explored due to their relevancy to affective and emotional states and how they can reveal various emotions. One of the earliest and most prominent works that investigated facial expressions and their association with emotions was written by Charles Darwin and was first published in 1872 (*'The Expression of the Emotions in Man and Animals'*). In this book, Darwin explored facial expressions in humans and how some emotions have universal facial expressions and also suggested that facial expressions of emotions are evident in some animals [116]. This is one of the main works that influences the modern-day research in facial recognition that was pioneered by Paul Ekman and his colleagues since the 1970s. One of the most important Ekman's works was the development of a coding system to represent facial expressions, the Facial Action Coding System (FACS) [117]. This system provides a universal comprehensive coding for facial expressions and it is applied in most facial expressions research. Another coding system for facial expressions is the Facial Animation Parameters (FAPs) [118, 119].

Since the coding systems are made available, the literature shows that emotions, psychological, and affective disorders can be detected from facial expressions [120]. For this reason, in this study, we attempt to implement an approach that uses FACS to detect attachment styles in children.

4.2.1 Facial Action Coding System (FACS)

A brief description of FACS will be provided below as it is the core element of our facial expression approach. Ekman and Friesen proposed describing facial expressions by the corresponding facial muscle movements that cause these expressions. They studied every facial muscle movement and noted the visible associated effects, which they called Action Units (AU). Accordingly, they identified all possible AUs and assigned them a unique code such as AU3 and AU12. Table 4.1 shows examples of AUs along with their corresponding movement muscles [121]. Moreover, given this coding, the basic emotions can be also described with one or more AUs, for example, happiness expression can be mainly characterised by two action units AU12 (Lip Corner Puller), and AU25 (Lips part) whereas anger expression can be characterised by three muscle movements, AU4 (Brow Lowerer), AU7 (Lid Tightener), and AU24 (Lip Pressor) [122].

Action Unit	Description	Muscles	Example Image
AU1	Inner brow raiser	Frontalis, pars medialis	
AU4	Brow Lowerer	Depressor Glabellae, Depressor Supercilli, Currugator	
AU7	Lid Tightener	Orbicularis oculi, pars palpebralis	
AU23	Lip Tightener	Orbicularis oris	

Table 4.1: **Examples of Facial Action Units.** The table shows examples of facial action units along with their facial muscles. Example images were taken from [123].

In about two decades after the development of the FACS, the efforts have been directed towards automating the facial Action Units Detection (for a survey, see [124]). In this respect, the *Openface* toolkit is one of the most popular publicly available toolkits for facial behaviour analysis which can detect the intensity and the presence of various action units [125, 126]. For intensity detection, a real value in the range [0, 5] is assigned to the AU to reflect the intensity of the muscle movement, whereas in presence detection, a binary value is used to indicate the presence or absence of a certain AU. Its implementation relies on appearance features based on HoG descriptors and geometry features based on the landmark locations [127].

4.2.2 Approach

Figure 4.1 depicts the overall architecture of the facial expressions-based approach. The approach consists of three main steps: 1) features extraction, 2) attachment recognition, and 3) aggregation. The first two steps are applied to the segment of the video recordings that corresponds to a single story, this yields five predictions per participant which are then aggregated to make the final prediction (participant-level prediction). This means that a separate classification model is trained over the material that corresponds to one story. This decision is made based on the rationale of having multiple different stressful scenarios in the MCAST assessment tool which suggests that having different attachment-related scenarios can invoke attachment-related behaviours to different extents. Thus, training a separate classification model on the material that corresponds to one story can capture different behavioural cues that are related to the specific scenario rather than learning general behaviours which might be harder to detect. This in turn can create an ensemble of classifiers that is expected to lead to a better performance.

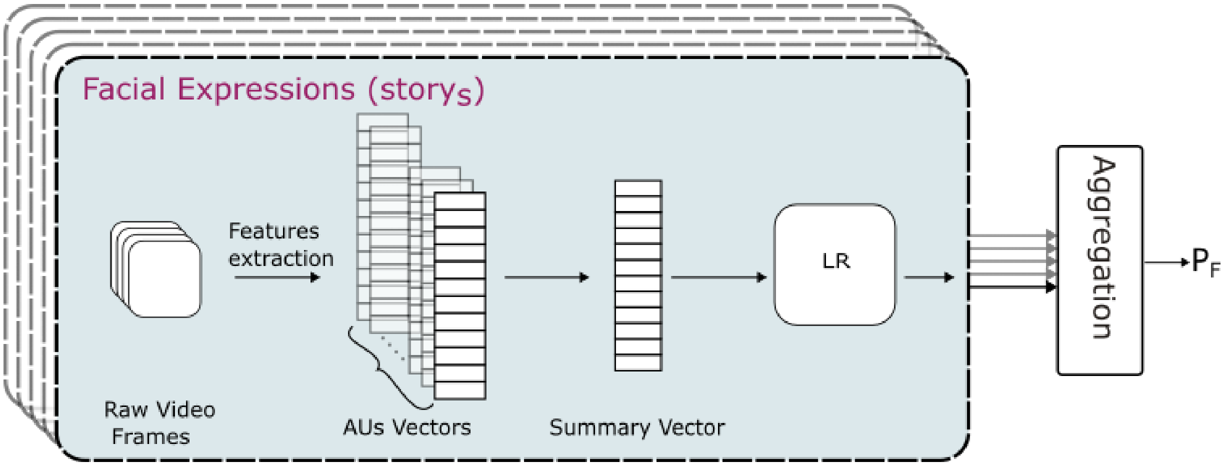


Figure 4.1: **Architecture of the Facial Expressions-based Approach.** This figure shows the overall architecture of the facial expressions approach. The features are extracted from video segments corresponding to a single story s. Each video frame results in one AUs vector. A summary vector is derived where the i^{th} element corresponds to the i^{th} AU by calculating the fraction of frames with AU values that are greater than θ_i. LR refers to the Logistic Regression classification model which results in story-level predictions. The above steps are repeated for all stories hence the five boxes in the figure. This results in five predictions per participant which then are aggregated using aggregation methods to give the final participant-level prediction P_F.

Features Extraction

In the first step, facial action units (AUs) were extracted from each video frame using the *Openface* toolkit [126, 127]. In this study, we extracted the intensity values of 17 AUs comprising the set of all AUs that can be recognised with *Openface* which are listed in Table 4.2. In the extraction step, this set of AUs is assigned real values in the range [0, 5] to reflect the intensity of the muscle movements where 0 indicates no movement, 1 indicates a movement with minimum intensity, and 5 indicates a movement with the highest intensity. This results in a sequence of feature vectors where each vector encodes one video frame.

After extracting these AUs, a threshold value (θ) per AU can be estimated from the training set. This θ value corresponds to the intensity value that discriminates between the highest 5%

AU	Description	Face Area
AU1	Inner Brow Raiser	Eyes Area
AU2	Outer Brow Raiser	
AU4	Brow Lowerer	
AU5	Upper Lid Raiser	
AU6	Cheek Raiser	
AU7	Lid Tightener	
AU45	Blink	
AU9	Nose Wrinkler	Nose Area
AU10	Upper Lip Raiser	Mouth Area
AU12	Lip Corner Puller	
AU14	Dimpler	
AU15	Lip Corner Depressor	
AU17	Chin Raiser	
AU20	Lip stretcher	
AU23	Lip Tightener	
AU25	Lips part	
AU26	Jaw Drop	

Table 4.2: **List of Action Units.** The table lists the description of the 17 facial Action Units (AUs) that were extracted by *OpenFace* with their corresponding face area. Each AU is presented with the corresponding muscle movement description as defined in [121].

intensity values and the lower values. Therefore, having this θ value, it is possible to obtain the fraction of frames f in a given video recording in which a certain AU is above the corresponding θ. In other words, this process obtains the fraction of frames where a corresponding AU is being the furthest from the neutral expression meaning that each recording is being represented by how often the AUs are being displayed.

This decision is motivated by the fact that the extracted feature vectors mostly correspond to neutral expressions. This means that the sequence-based approaches might not be well suited to capture any related information. Therefore, representing the input videos by taking into account only the video frames in which an expression is being displayed can be more informative (separable) for our classification task. Thus, at the end of the extraction step, each video recording that corresponds to one story is converted to a feature vector of dimension D=17 where the i^{th} element corresponds to the fraction of frames f_i in which the intensity of the i^{th} AU is above its corresponding threshold θ_i.

Attachment Recognition

The next step is attachment recognition, where the the above resultant summary features vectors are fed to a Logistic Regression classifier which maps these vectors to the attachment types according to equation 2.8:

$$p(y=1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^{\mathbf{T}}\mathbf{x}+b)}$$

where $\mathbf{x}$ is the feature vector corresponding to one input sample, $\mathbf{w}$ is the weight vector, and b is the bias parameter. The above formula produces a value between 0 and 1 that can be interpreted as a probability value. The training involves minimising the loss function to find the optimal parameters $\mathbf{w}$ and b which corresponds to maximising the cross entropy function which is given by equation 2.9. This loss is also penalised using an L2-norm regularisation $\frac{\lambda}{2}||\mathbf{w}||^2$ in order to enhance the generalisation where λ specifies the regularisation strength (equation 2.10). Optimising this error function is implemented using the L-BFGS which is a second-order optimisation algorithm belonging to the class of Quasi-Newton iterative methods.

The above two steps result into five predictions per participant that have to be aggregated. This brings us to the final step of our approach.

Aggregation

In this step, the predictions that were made in the previous stages are aggregated to make a decision at the participant level, meaning that participants are assigned to one of the two attachment classes (Secure and Insecure). In this study, two aggregation methods were applied: Majority Voting (MV), and Sum Rule (SR) (see section 2.2.6). MV aggregates the decision labels that were made at the story level in which the participants are assigned to the class where the majority of their stories are assigned to.

In the second aggregation method, the posterior probabilities that were estimated by the LR model at each story are summed and the participants are assigned to the class with the highest value. One advantage of SR over MV is that predictions with higher probability can have more influence on the aggregation results. The choice of this late fusion scheme instead of an early fusion for stories aggregation lends itself naturally to the way the assessment process is conducted in the manual mode. Specifically, according to [11], each story is assessed and rated individually and the full interview is rated based on the predominant classification across all stories.

The above classification model has led to a significant improvement in the performance compared to other various models including: a Deep RNN model with frame-based AUs features, an SVM model to classify either the vector of AUs or each AU separately, and an SVM model with a summary vector as the one illustrated above.

4.2.3 Experiments and Results

Experimental Design

The experiments were implemented with the k-fold protocol where k=10, in which participants were randomly assigned to one of the k disjoint groups, with k-1 folds were used as a training set and the remaining fold was used as a testing set. The results from each fold-based iteration were combined to compute the total classification results. This is to overcome the limited number of samples in the dataset. Moreover, the samples in each fold were assigned using a stratified method to maintain the class distribution across different folds. This is to tackle the class imbalance in the dataset to ensure that no class is being underrepresented in any fold. Additionally, each experiment was repeated $R = 10$ times and the final results are presented in terms of averages and standard deviations over these R repetitions. This is to account for the randomisation involved in the initialisation of the LR parameters and the generation of the k folds.

An L2 regularisation was implemented in the LR model to reduce the generalisation error for the unseen data. In this respect, a hyper-parameter C (inverse of the regularisation strength) has to be set to determine the strength of the regularisation effect. In this study, C was optimised by a nested cross-validation scheme and grid search method, meaning that the model with the best C value on a validation set is selected for the final predictions. The number of folds of the nested cross-validation were set to k=10, where k-1 folds were used as training set and the remaining fold as a validation set. The grid search was set over the following discrete set $C \in \{0.05, 0.1, 0.5, 0.75, 1, 5, 10\}$.

Evaluation Metrics

The results in all experiments are presented in terms of four performance metrics common in classification problems: Accuracy, Precision, Recall, and F1-score, where each metric provides a different perspective on how well an approach performs. In Accuracy, the performance is assessed as the fraction of correctly predicted cases among all cases. In clinical applications, the rate of identifying positive cases (in this study, children who have insecure attachment style) is of significant importance because this is when the recognition approaches provide information about who needs treatment, this is where Precision, Recall, and F1-score can give a better estimate on how a system performs. In particular, precision provides a measure of how well the approach correctly identified the positive cases among all positively identified cases, whereas recall measures how well the approach identifies positive cases among all positive cases in the test set. F1-score combines these two measures in one metric using the harmonic mean.

The following equations define these metrics in terms of the predicted cases: TP, TN, FP, and FN denoting true positives, true negatives, false positives, and false negatives respectively.

$$\begin{aligned}
&Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\\
&Precision = \frac{TP}{TP+FP}\\
&Recall = \frac{TP}{TP+FN}\\
&F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}
\end{aligned} \tag{4.1}$$

Statistical Significance Tests

The statistical significance of the results were examined by utilising the t-test and the significance level is reported in three different p-values ($p < 0.05$, $p < 0.01$, and $p < 0.001$). T-test assumes that the samples follow a normal distribution and the variances of these samples are homogeneous. In our experiments, the comparisons were made between the averages of 10 repetitions of each experiment and according to *Central Limit Theorem (CLT)*, the averages of large enough samples [1] tends to approximately follow a normal distribution. Moreover, a normality test was conducted for the performance results obtained from these repetitions, and the

[1] The sample size should be at least 30, the larger the sample size the closer the sample averages to a normal distribution.

corresponding p-values are reported in Table A.1 (see Appendix A for more details). In addition, no major differences in the variances are noted between different samples and thus these assumptions are not violated.

Results

Table 4.3 presents the results of the facial expressions-based approach in terms of the above four performance metrics along with the random guess baseline. As the table shows, all experiments perform better than the baseline to a statistically significant extent in all metrics ($p < 0.001$, single-tailed t-test), except for the recall in the HS story. The best performance is achieved by majority vote (MV) with accuracy of 64.3% and F1-score of 56.1% and this improvement is statistically significant with respect to any individual story for accuracy, precision ($p < 0.01$), and F1-score ($p < 0.05$). This suggests that models that were trained over different individual stories are diverse, i.e., they tend to make different mistakes over different samples, an essential property for multimodal combination to improve over their unimodal components [106] (see section 2.2.6).

Story	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Breakfast (BF)	59.4 ± 1.2	52.9 ± 1.4	51.9 ± 2.5	52.3 ± 1.6
Nightmare (NM)	61.9 ± 2.3	56.5 ± 2.9	52.2 ± 3.2	54.3 ± 3.0
Tummyache (TA)	61.4 ± 2.2	55.9 ± 2.9	49.8 ± 2.5	52.6 ± 2.4
Hopscotch (HS)	57.4 ± 2.0	50.2 ± 2.7	44.3 ± 3.1	47.0 ± 2.3
Shop. Mall (SM)	58.9 ± 2.1	51.9 ± 2.5	50.1 ± 4.4	51.1 ± 3.4
All (MV)	64.3 ± 1.4	60.1 ± 2.3	52.7 ± 2.1	56.1 ± 1.3
All (SR)	64.0 ± 1.9	60.2 ± 2.9	50.4 ± 2.4	54.8 ± 2.1
Random	51.0	43.0	43.0	43.0

Table 4.3: **Facial Expressions Approach Performance.** This table presents the performance results of the facial expressions-based approach. The results are presented in terms of averages and standard deviations over R repetitions. (All) stands for the participant-level predictions.

Both aggregation methods (MV and SR) improve over the best individual stories in accuracy and precision (NM and TA) ($p < 0.01$). However, in recall, both methods fail to outperform the best stories (NM and BF) meaning that the different classifiers are not making different mistakes in identifying the positive cases. Additionally, in F1-score, only MV method performs better than the best story (NM) ($p < 0.05$). This finding suggests that insecure children are less likely to change their behaviours in response to different attachment scenarios, at least when it comes to using facial expressions.

Another important finding is that the approach achieves its worst performance in recall, meaning that it is generally harder to identify insecure children. One possible reason for the low recall performance is that the dataset has fewer insecure children (45 insecure compared to 59 secure) and therefore fewer materials for the training process to capture behaviour cues relevant to insecure attachment.

It is also shown in the table that the performance varies between different stories, where the best-performing stories are NM and TA which both achieve accuracies higher to a statistically significant extent than other stories. In F1-score, all stories perform better than HS ($p < 0.01$). This suggests that different stories are more likely to induce attachment-related behaviours to different extents which is the motivation behind having multiple stories in MCAST.

Effects of Age and Gender Variations

In this section, the effect of age and gender variations on the performance of the face-based approach is discussed. Table 4.4 presents the performance with respect to each school level as each level is associated with a specific age range as shown in Table 3.1. These results are obtained per each level after aggregating the predictions from all stories using the SR aggregation method. The SR method is used instead of MV (although MV performs better in this approach) to ensure that the results are compatible across all unimodal approaches for later comparisons.

Level	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
P1 (5-6)	53.7±4.8	60.1±8.7	37.0±6.7	45.5±6.5
P2 (6-7)	56.5±3.8	51.9±4.4	45.0±6.7	48.1±5.2
P3 (7-8)	76.6±1.5	71.6±3.3	63.6±0.0	67.3±1.4
P4 (8-9)	72.5±5.3	63.8±8.1	65.0±5.3	64.1±5.4

Table 4.4: **Age-Based Performance of the Facial Expressions Approach.** This table presents the performance of the facial expression-based approach with respect to each school level. Each level is shown with the associated age range between parentheses. The results are shown in terms of averages and standard deviations over the R repetitions.

The most important finding in this table is that the performance in accuracy, recall, and F1-score is significantly higher for older children (7-9 years) compared to their younger counterparts (5-

7 years) ($p < 0.001$, t-test). This suggests that older children are more likely to express their attachment condition through facial expressions compared to younger children. One plausible explanation is that according to [128], facial expressions for this age group are still developing in which the ability to recognise and produce facial expressions improves between the age of 3 to 10 years. Therefore, older children may be more capable of producing facial expressions that reflect their inner emotional state. However, the distribution of attachment styles in each age group is not balanced; there are fewer children in levels P1 and P4 with even fewer insecure children in level P4, so this finding could likely result from having more negative cases[2] in older age groups, therefore, this finding is worth further investigation with a more age-balanced sample.

In precision, the pattern is slightly different than what is observed in the other metrics, where level P2 is significantly lower than all groups ($p < 0.01$) whereas level P3 is higher than other groups ($p < 0.01$). High precision indicates less false positives meaning that the positive predictions can be trusted to better extent for P3 compared to other groups.

The other factor that may interplay with performance is the gender variations, for this reason, Table 4.5 shows the performance with respect to each gender group. As shown in the table, in all metrics, the performance is higher for female participants compared to their male counterparts to a statistically significant extent ($p < 0.001$). In fact, the performance with regard to the male participants hardly improves over the random guessing for recall metric, meaning that it is unlikely to identify insecure male children in this approach. This suggests that female children are more likely to express their condition through facial expressions. This finding is consistent with previous studies which found that females are more facially expressive than males [129, 130] and as such our approach is more likely to capture relevant cues to attachment styles for female participants.

Group	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Female	70.0±1.3	73.3±3.3	54.2±2.8	62.2±1.6
Male	56.8±4.1	46.8±5.2	45.3±4.4	45.9±4.1

Table 4.5: **Gender-Based Performance of the Facial Expressions Approach.**

[2] For which the approach has better identification rate.

4.3 Paralanguage Based Approach

The second unimodal approach is based on analysing the nonverbal aspects of speech such as pitch, intonations, and loudness which are also referred to as paralanguage. In other words, paralanguage refers to how something is said instead of what is being said. These aspects can convey information about the internal state of individuals such as their current emotions, for example, it can be easily distinguishable to human ears between happy or angry individuals only by hearing their voices. Moreover, past studies have shown that these aspects of speech carry important cues about a wide range of emotional and psychological traits [120, 131]. For these reasons, this study attempts to detect the differences in attachment styles in children by analysing the paralanguage aspects of speech signals.

The proposed recognition approach is based on extracting various paralanguage features from the recordings of the MCAST interviews. The set of features used in the experiments is chosen due to its relevance to capturing emotional and psychological traits. These features are typically extracted from a short analysis window that lasts for 20-40 ms which results into sequences of vectors. This requires applying a classification model suitable for the sequential nature of the data to capture temporal information. For this reason, we proposed using a stacked RNN model to learn relevant representations that can be mapped to attachment styles (see section 2.2.3). As in the case of the previous approach, the classification model is trained over the segments of recordings that correspond to each one of the five stories and therefore aggregation methods are applied to make the final predictions. In the following, a brief description of the speech features is provided.

4.3.1 Speech Based Features

Features extraction consists of converting digital speech signals into a form that is suitable for processing. The core concept of signal processing is the Fourier Transform (FT) which is based on the fact that every signal can be decomposed into a set of sinusoids scaled and shifted appropriately. This transforms the signal from time domain[3] representation to the frequency domain which reveals aspects of the signal that were not visible in the time domain representation. Moreover, the FT decomposition allows for visualising the speech signals both in time and fre-

[3] Where signals are shown as a function of time.

quency through a *spectrogram*. Figure 4.2 shows an example of an audio raw signal and its corresponding spectrogram representation.

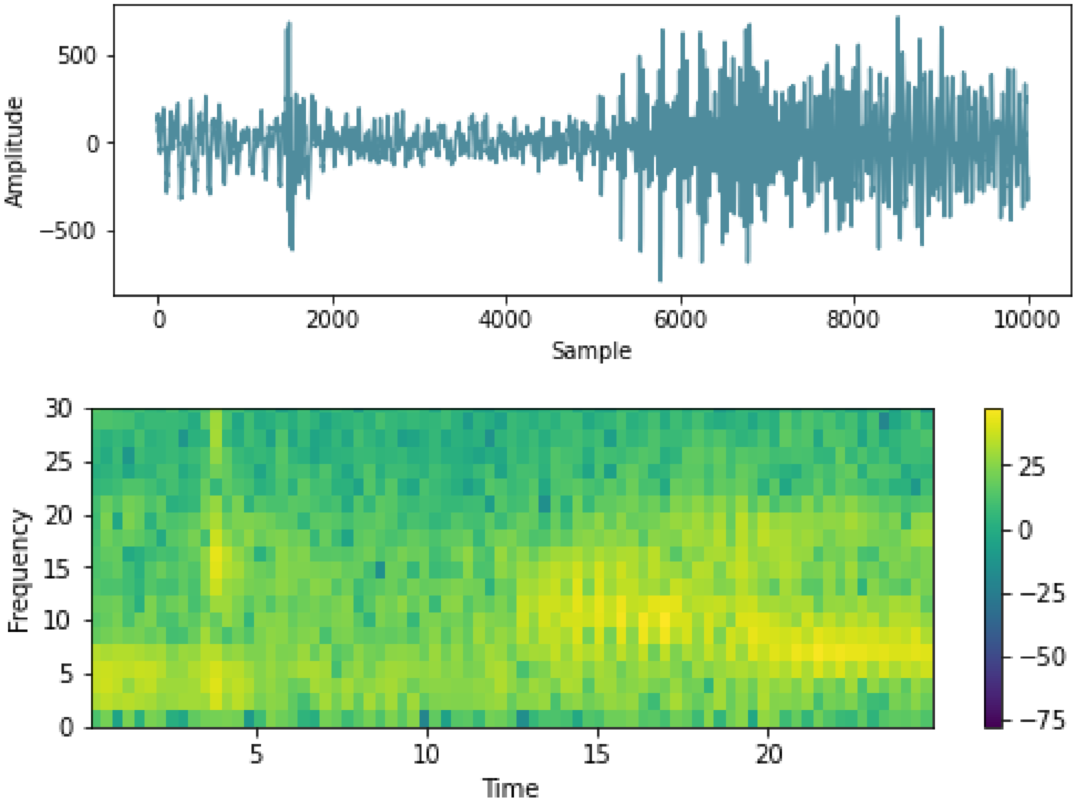


Figure 4.2: **Spectrograms.** The top pane shows the amplitude of an audio signal excerpt in the time domain while the lower pane shows the same excerpt in both time and frequency domains. Brighter regions indicate higher amplitude.

Various speech features, which are also known as low-level descriptors (LLDs) can be extracted from the time domain (temporal features) such as Root Mean Square Energy (RMSE) and Zero Crossing Rate (ZCR), whereas others can be extracted from the frequency domain (spectral features) such as fundamental frequency F_0 and Mel-Frequency Cepstral Coefficients (MFCC). These features are often extracted over short-time analysis windows using a windowing function such as rectangular, Hamming, or Hann window where the Hamming window is often used for frequency domain features and the rectangular window is used for time domain features.

The LLDs used in the experiments are:

- Root Mean Square Energy (RMSE): it accounts for loudness and magnitude of the signal;
- Zero Crossing Rate (ZCR): it accounts for the smoothness of the signal, it provides information about the frequency distribution and it is computed as the number of zero crossings

per time unit;

- Mel Frequency Cepstral Coefficients (MFCC) (1-12): A set of coefficients extracted from Mel-Scale Frequency Spectrogram[4]. In speech recognition, the coefficients 0-12 are often used [132], in which coefficient 0 accounts for the signal energy while coefficients 1-12 for the phonetic content of speech;
- Fundamental Frequency (F_0): it accounts for the frequency of the highest energy and it contributes to defining how the voice sounds;
- Voicing Probability: it accounts for the presence of silences.

Other features can be derived from the above LLDs such as delta regression coefficients which are calculated by finding the local slope over 50-100 ms of the signal, and it measures how the features change over time which, in some cases, makes it easier for a classifier to detect patterns compared to the direct LLD features [133].

OpenSMILE

OpenSMILE [134, 135] is an open-source software for extracting speech features. It was the official toolkit for feature extractions for the INTERSPEECH 2009 Emotion Challenge [136]. Various types of LLDs are supported along with several statistical functions. Its incremental processing allows for efficient implementation as each feature extractor can be used as an input to other feature extractors thus ensuring that no feature has to be computed twice.

4.3.2 Approach

Figure 4.3 depicts the overall architecture of the paralanguage-based approach. As shown in the figure, the approach consists of three main steps: 1) features extraction, in which the raw audio signals are converted into sequences of feature vectors, 2) attachment recognition, in which a classification model maps these sequences into one of the attachment classes, and finally 3) the aggregation, in which the predictions that were made at story level are aggregated to make the final prediction (participant-level predictions).

[4] A spectrogram representation with frequency scales corresponds to the human auditory system.

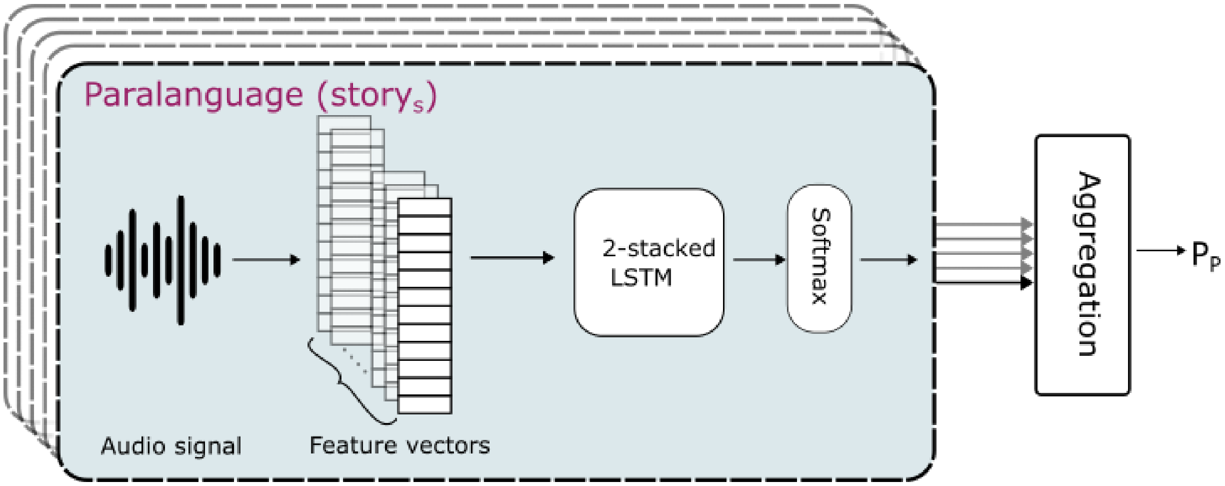


Figure 4.3: **Architecture of the Paralanguage-Based Approach.** This figure shows the overall architecture of the paralanguage approach. A set of speech features are extracted from the audio signals corresponding to a single story s. This results in a sequence of feature vectors. These vectors are fed a stacked LSTM classification model followed by a softmax layer to predict the attachment labels. This process is repeated for all five stories resulting in five predictions which then are aggregated to make the final participant-level prediction P_P.

Features Extraction

In the first step, various speech-based features were extracted from the raw audio signals using the OpenSMILE toolkit. The set of features adopted in this work covers the main aspects of non-verbal behaviours in speech and it was the baseline for the emotion recognition challenge [136] and since has been shown to be effective in speech-based recognition domains. Specifically, 16 basic features along with their delta regression coefficients were extracted. The basic features are Root mean square of the energy (1 feature), Mel Frequency Cepstral Coefficients (12 features), Zero Crossing Rate (1 feature), Voicing probability (1 feature), and Fundamental frequency (1 feature).

This results in a set of 32 features which are extracted over 33 ms non-overlapping analysis windows using the hamming windowing function (an analysis window corresponds to one video frame). Moreover, the feature values were smoothed by averaging over every three consecutive windows. More formally, the extraction step converts the raw audio recordings into a sequence of T vectors of dimension $D = 32$, $X = (\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_T})$ where T is the number of analysis windows per recording.

Attachment Recognition

The next step is attachment recognition where a classification model is trained to map these extracted set of features to the attachment labels. In this case, as the features set denotes sequences, an approach based on a Long Short Term Memory (LSTM) network appears to be suitable for the task for its ability to capture dependencies across sequences while having an enhanced memory capability to allow for processing longer sequences compared to the conventional RNNs (see section 2.2.1). Moreover, a deeper architecture is employed where two LSTM layers are stacked as shown in Figure 2.4, and following the architecture proposed in [95]. This architecture proposed feeding the hidden states of the base layers to the top layers which allows for capturing higher level representations. The output of the last time step at the top layers is fed to the softmax layer which assigns a probability value corresponding to the attachment labels.

As shown in section 3.5, the average length of the recordings per each story in the data set is between 111 and 158 seconds which results in very long feature sequences[5] and therefore, (despite having an enhanced memorisation capability which allows for capturing temporal dependencies that span over 1000 time steps [85]), training an LSTM model on very long sequences can give rise to vanishing and exploding gradients problem [86]. To tackle this problem, the feature sequences are segmented into non-overlapping segments in which each segment contains L vectors and then each segment is treated as a separate training example labeled by the same label assigned to the full recording. To classify new input during the test phase, the full sequence is assigned to the class where the majority of its segments are assigned to i.e., the full sequence is assigned to the class $\hat{c}$ that satisfies the following: $\hat{c} = argmax_{c\in\{c1,c2\}} f(c)$, where $f(c)$ is the fraction of segments in a given sequence that are assigned to class c.

Aggregation

As each story is expected to elicit attachment-related behaviours to different extents, the above two steps (feature extraction and attachment recognition) are repeated for each one of the five stories stems as depicted in Figure 4.3, this results in five predictions per a participant which then are aggregated using two aggregation methods MV and SR (as in the facial expression approach). In the case of SR, the posterior probability has to be estimated at the full sequence which has multiple segments, each associated with a prediction (with a corresponding estimation

[5] Correspondingly, the average sequence length is between 3600 and 5200 vectors per recording.

of the posterior probability at segment level). One way to estimate the posterior probability for the full sequence is by calculating the fraction of the segments assigned to the predicted class.

4.3.3 Experiments and Results

Experimental Design

The experiments were performed according to a k-fold protocol ($k = 10$) in which the participants were assigned randomly to one of the k folds, meaning that all materials that belong to the same participant are assigned to the same fold, thus ensuring that the same participant never appears in both training and testing sets. This is to ensure that the approach recognises the attachment styles, not the participant's identity. Moreover, the folds are stratified so the percentage of classes was maintained across all folds as in the initial full dataset. As such, $k-1$ folds were used for training, while the remaining fold is used for testing.

To account for the randomisation involved in the experiments (initialisation of the RNNs and assigning participants to the folds), each experiment was repeated $R = 10$ times and the results are presented in terms of averages and standard deviations over these repetitions. Moreover, the performance is reported by four different evaluation metrics (see section 4.2.3).

Other hyper-parameters were set to values that were considered to be standard in the literature. In particular, the dimension of the hidden states was set to $D = 70$, the learning rate to α=1e^{-3}, and the number of training epochs to E=50. The training was performed with the mini-batch strategy with $B = 512$ and the risk of overfitting was reduced by applying the L2-norm regularisation (the λ parameter was set to 1e^{-2}). The length of the audio segment was set to L=128 which corresponds to 4.2 seconds of the recording which is considered to be short enough to alleviate the vanishing and exploding problem and long enough to allow to capture relevant cues.

Results

The results are shown in Table 4.6 in terms of four performance metrics. As shown in the table, in all experiments, the improvements over the random baseline are always statistically significant for accuracy, precision, and F1-score ($p < 0.01$, t-test) meaning that the approach is capturing paralanguage-related cues that correlate with attachment styles. The best performance is achieved by aggregating classifiers of all stories using the SR method with accuracy and F1-

score of 69.3% and 60.2%, respectively.

Story	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Breakfast (BF)	64.4 ± 3.0	60.9 ± 4.5	48.4 ± 4.1	53.9 ± 3.8
Nightmare (NM)	59.7 ± 2.7	54.4 ± 4.0	44.0 ± 6.9	48.4 ± 4.8
Tummyache (TA)	62.9 ± 2.6	59.1 ± 4.5	46.4 ± 4.9	51.8 ± 3.9
Hopscotch (HS)	64.8 ± 3.2	62.2 ± 5.1	46.1 ± 8.8	52.4 ± 6.5
Shop. Mall (SM)	64.8 ± 3.5	61.0 ± 5.2	48.6 ± 7.1	53.9 ± 5.8
All (MV)	66.5 ± 2.6	65.1 ± 3.6	48.9 ± 5.7	55.7 ± 4.4
All (SR)	69.3 ± 1.7	68.8 ± 3.4	53.8 ± 4.3	60.2 ± 2.7
Random	51.0	43.0	43.0	43.0

Table 4.6: **Paralanguage Approach Performance.** This table shows the performance of the paralanguage approach. The results are shown in terms of averages and standard deviations over 10 repetitions.

However, in Recall, the improvements over the baseline are statistically significant only in 3 out of 5 stories and the aggregation of all stories ($p < 0.05$). Whereas in two stories (NM and HS), the classifiers do not improve over the baseline, meaning that these two stories are less likely to elicit attachment behaviours compared to other stories. Interestingly, while NM achieves low performance across all metrics, HS has the best accuracy and precision score, this means that the HS story has the biggest variations in terms of recognising cases from different classes i.e., it seems that HS performs relatively well at recognising negative cases and performs poorly in recognising positive cases. In general, recall has the lowest performance compared to other metrics both at the story level and the aggregation methods. One possible reason is that there are less positive cases (insecure children) in the dataset and thus the approach is less likely to capture related cues.

Another important aspect of the results is that different stories achieve varying performances which also vary across different metrics. For instance, the difference between the highest performing story in accuracy (64.8% for HS and SM) and the lowest one (59.7% for NM) is statistically significant ($p < 0.001$) while in F1-score the highest performing story (53.9 % for BF and 53.9% for SM) is higher than (48.4% for NM). This confirms the fact that different stories are more likely to elicit detectable attachment-related behaviours than others.

Aggregating the predictions from all stories improves the performance over the best individual one to different extents. In the case of the MV method, the improvement is not significant in any metric suggesting that the classifiers that are trained over individual stories are not diverse, i.e., they do not tend to make different mistakes over different children [106]. On the other hand, the SR method has improved over the best-performing story and MV significantly in all metrics ($p < 0.05$). In the SR method, the posterior probabilities are estimated by the fraction of segments assigned to a certain class and therefore the improvements over the MV method suggest that when the classifier assigns segments to the right class, it tends to do it to a greater extent. In other words, the fraction $f(c)$ of segments assigned to class c tends to be higher when c is the right class.

Effects of Age and Gender Variations

In this section, the effect of age and gender on the paralanguage approach is discussed. Table 4.7 shows the performance in each age group. As the table shows, different patterns are noted for different metrics. In particular, the accuracy is higher in levels P2 and P3 compared to the remaining levels to a statistically significant extent ($p < 0.05$). In precision, P2 achieves higher performance than the other groups ($p < 0.01$) while P4 achieves the lowest performance ($p < 0.01$). In recall, P1 achieves the best performance compared to other groups ($p < 0.01$), while P3 has the lowest performance ($p < 0.01$). In F1-score, both P1 and P2 perform better than P3 and P4 to a statistically significant extent ($p < 0.001$).

Level	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
P1	66.8±4.3	70.8±3.9	63.0±9.5	66.3±6.6
P2	71.5±2.9	75.8±4.2	53.9±5.3	62.9±4.6
P3	70.7±2.4	67.1±6.4	45.5±6.1	53.9±4.7
P4	64.4±7.2	54.3±10.6	53.3±7.0	53.1±5.9

Table 4.7: **Age-Based Performance of the Paralanguage Approach.** This table presents the performance of the paralanguage-based approach with respect to each age group. The performance is presented in terms of averages and standard deviations over 10 repetitions.

In general, as shown in F1-score, the approach seems to perform better for younger groups (P1 and P2), however, the behaviour of the approach is different over these groups, as it seems that the approach is identifying positive cases in P1 to better extent while it identifies the negative cases better in P2. Additionally, the overall high performance in P2 and P3 likely results from the higher number of samples in both groups (40 and 29 respectively) compared to the remaining groups as shown in Table 3.1, while the low recall performance in P3 could be a result of the low positive cases in this group. This implies that the approach achieves higher performance with a higher amount of training materials.

The performance with respect to the gender groups is shown in Table 4.8. As shown in the table, the performance is higher for male participants compared to female counterparts in accuracy ($p < 0.01$), and precision ($p < 0.05$). Conversely, in recall and F1-score, the performance is higher for female participants compared to male counterparts ($p < 0.01$). This suggests that the paralanguage approach performs better at identifying secure male children whereas it performs better at identifying insecure female children.

Gender	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Female	67.5±2.1	67.3±4.2	56.9±6.0	61.4±3.3
Male	71.5±3.2	71.3±5.1	49.5±6.2	58.3±5.3

Table 4.8: **Gender-Based Performance of the Paralanguage Approach.** This table presents the performance of the paralanguage-based approach for each gender group.

4.4 Language Based Approach

This section presents an approach to detect the attachment styles based on the linguistic contents of the MCAST interviews by applying Natural Language Processing (NLP) methodologies. The linguistic contents can reveal rich information about an individual's inner feelings and mental states. In this respect, past studies show that emotions can be detected from text [137] and therefore this was extensively used in areas such as sentiment analysis, mental states detection, and emotion recognition. Moreover, a previous study found that some linguistic features are strong predictors of attachment styles in adults [81]. Another study concluded that the linguistic coherence is one of the main predictors for attachment styles in children as assessed by human raters [11]. These reasons motivated the decision to apply a language-based approach which exploited the advancements in NLP methodologies.

In this study, an approach based on word embeddings representation and 1D-CNN classification model (see section 2.2.4) is proposed and in the next sections the approach and the results will be presented.

4.4.1 Approach

The architecture of the language-based approach is depicted in Figure 4.4. The audio recordings are first transcribed automatically using the online transcription tool *Sonix*[6]. Then, the approach is implemented in three consecutive steps: preprocessing, attachment recognition, and aggregation. In the first step, the transcripts are preprocessed to eliminate any non-word symbols such as punctuations, numbers, and other symbols. In addition, all words are converted to lowercase to ensure that the same words are mapped to the same representation. Moreover, the English stopwords are removed using the Natural Language Toolkit (NLTK)[7], a publicly available package for text processing.

Then, the resultant list of words is lemmatised meaning that all word's morphological variants are converted to their root word. This is to ensure that these word variants are all mapped to the same word representation. The segment of the text that corresponds to the story induction phase is removed because the focus is on the text that the participants uttered. Moreover, this phase is identical in all samples so it does not carry any relevant cues for our classification problem. Next, the resulting terms are tokenised and the word vocabulary is constructed, in which every word is assigned a unique integer index. Finally, all token vectors are zero-padded to a common length L (sequences that are longer than L are truncated). In summary, the preprocessing step converts the raw text transcripts to token vectors of length L in which each element corresponds to a numerical word index from the vocabulary, a suitable form for further processing.

The resulting token vectors are fed to a deep learning classification model which consists of three main parts (Embedding layer, 1D-CNN/Max-pool/Dropout/Dense, and a Softmax layer). In the embedding layer, each token is mapped to an embedding vector of a fixed length dimension M, thus the input vectors are represented with a matrix of dimensions $L \times M$. These matrices are then passed to a 1-D convolutional neural network model (1D-CNN) in which K filters are convolved with the embeddings to extract relevant features from words' n-grams (see section 2.2.4). This is followed by a max-pooling layer that aims at preserving the most salient features resulting from the convolution step. Next, to enhance the generalisation of the approach a dropout layer is applied followed by a dense layer to learn higher-level representation, and finally, a softmax layer is used to map these representations to one of the attachment labels.

[6] https://sonix.ai
[7] https://www.nltk.org

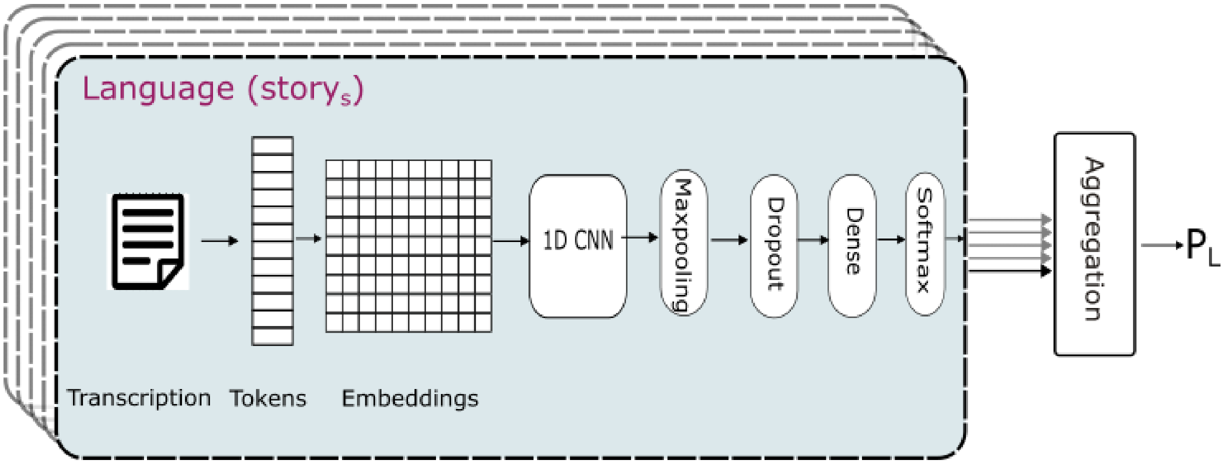


Figure 4.4: **Architecture of the Language-Based Approach.** This figure shows the overall architecture of the language-based approach. The transcript that corresponds to story$_s$ is first preprocessed to obtain a list of tokens (word indices) which is then padded to a length L. Then, this tokens vector is fed to the classification model in which the tokens are first mapped to word embeddings and then fed to a CNN model which is composed of a 1D-CNN layer followed by maxpooling, dropout, and dense layer. Finally, a softmax layer is applied to classify the learned representations to the attachment labels. The above process is repeated over the five stories resulting in five predictions per participant which then are aggregated to make the final prediction P_L.

As depicted in the figure, the above steps are repeated over the transcripts that correspond to each story individually. As a result, there are five predictions for each participant which are aggregated using both MV and SR aggregation methods to make the final prediction (participant-level prediction) P_L.

4.4.2 Experiments and Results

Experimental Design

The experiments were implemented in a k-fold protocol with $k = 10$ by assigning children randomly to the folds where $k-1$ folds were used for training and the remaining one is used for testing. All text sequences were padded to the same length $L = 100$ and sequences that are longer than L were truncated, where L is chosen empirically in which, approximately, 95% of the samples' tokens are shorter than L. The word embeddings dimension was set to $D = 300$ and the number of filters in the convolution layer was set to $F = 64$. Three different filter sizes were explored (3, 5, and 7) with 5 giving the best performance, and the dropout rate was set to $p = 0.2$. As in the previous approach, the training was performed with a mini-batch strategy to limit the computational issues with each mini-batch including ten input sequences, and the number of the training epochs was set to $T = 20$. All experiments were performed $R = 10$ times and, at every repetition, the layers' weights were initialised randomly. All performance metrics are presented in terms of averages and standard deviations over the R repetitions.

Results

Table 4.9 presents the performance of the language-based approach obtained over every story and over their aggregations through MV and SR. According to the t-test, all experiments perform, to a statistically significant extent, better than the random classifier ($p < 0.01$) in all metrics except for recall in TA story. The best result is achieved by the SR aggregation method in accuracy and precision achieving 72.1% and 72.6% respectively, however, the difference with respect to the best-performing stories is not significant (71.2% for accuracy, and 70.6% for precision). Whereas in recall, the BF story outperforms both individual stories and the aggregation methods to a significant extent ($p < 0.05$). For F1-score, BF and SR achieve comparable performances and both outperform other experiments significantly ($p < 0.05$).

Story	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Breakfast (BF)	71.2 ± 2.0	69.3 ± 3.1	59.5 ± 2.2	64.0 ± 2.2
Nightmare (NM)	65.1 ± 2.9	62.3 ± 4.6	49.3 ± 4.2	55.0 ± 4.0
Tummyache (TA)	64.4 ± 2.3	62.7 ± 4.1	43.4 ± 2.5	51.3 ± 2.9
Hopscotch (HS)	70.8 ± 2.0	70.6 ± 2.8	54.1 ± 3.7	61.2 ± 3.0
Shop. Mall (SM)	69.5 ± 1.9	66.7 ± 2.8	57.3 ± 3.0	61.6 ± 2.5
All (MV)	69.3 ± 1.9	69.7 ± 3.1	51.6 ± 2.7	59.2 ± 2.7
All (SR)	72.1 ± 1.7	72.6 ± 2.6	57.1 ± 2.3	63.9 ± 2.3
Random	51.0	43.0	43.0	43.0

Table 4.9: **Language Approach Performance.** This table shows the performance of the language-based approach. As in the other approaches each experiment was repeated R times and the results are presented in terms of averages and standard deviations.

Both aggregation methods did not improve over the best performing story in any metric. In fact, except for precision, the difference between the best performing story (BF) and the MV aggregation is significant ($p < 0.05$) suggesting that in this approach, the classifiers that are trained over different stories are not diverse and hence does not lead to better performance. Moreover, SR performs significantly higher than MV in all metrics ($p < 0.05$), however, it still fails to outperform the best performing story in any metric. The improvement of SR over MV suggests that the approach tends to be more confident (by assigning higher posterior probability) when making the right decision.

As in the previous approaches, recall achieves the lowest performance among all metrics meaning that it is harder to identify insecure children in this approach too. Moreover, the performance varies across individual stories in this approach too, where in accuracy, the difference between the highest performing story (71.2% for BF) compared to the three least performing stories (69.5% for SM, 65.1% for NM, and 64.4% for TA) is statistically significant ($p < 0.05$). This result confirms that some stories are more likely to elicit detectable attachment-related behaviours compared to other stories.

Effects of Age and Gender Variations

In this section, the effect of age and gender variations on the language approach is discussed. Table 4.10 shows the performance of the language approach with respect to age variations. As the table shows, the performance is significantly higher for P1 and P2 compared to P3 and P4 ($p < 0.001$, for recall and F1-score) and ($p < 0.05$, for precision). In accuracy, a slightly different pattern is noted, where the best performance was achieved in P2 which is significantly higher than P1 and P3 ($p < 0.01$). These findings seem to suggest that younger children are more likely to express their attachment condition through language compared to their older counterparts. Moreover, the improved accuracies for older children (P3 and P4) compared to other metrics suggests that the approach is identifying negative cases (secure children) to better extents. In other words, it seems that the approach is more likely to misclassify older insecure children. One possible reason is the limited number of insecure children in older groups (only 38% of all insecure participants belong to these groups) which may provide limited training materials to capture any relevant cues.

Level	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
P1	70.5 ± 3.7	80.4 ± 6.9	59.0 ± 3.2	67.9 ± 3.2
P2	74.3 ± 2.4	74.8 ± 2.3	64.4 ± 4.7	69.2 ± 3.5
P3	69.7 ± 2.2	63.2 ± 2.9	47.3 ± 5.7	54.0 ± 4.8
P4	73.1 ± 3.0	70.5 ± 7.2	50.0 ± 0	58.4 ± 2.6

Table 4.10: **Age-Based Performance of the Language Approach.** This table presents the performance of the language approach in each age group in terms of averages and standard deviations over R repetitions.

The performance with respect to gender variations is presented in Table 4.11. As the table shows, the performance for male participants is significantly higher than for females in all metrics ($p < 0.05$, for precision) and ($p < 0.001$, for other metrics). The biggest difference between groups appears in recall suggesting that insecure male children are more likely to express their condition through language compared to females, however, this may be an effect of the age variations observed above as there are more male participants in younger groups and thus further investigation is needed to confirm this effect. The relatively high precision for female participants (71.5%) compared to recall indicates that the approach is making fewer false positive errors i.e., for female participants, the positive predictions can be trusted to a greater extent compared to negative predictions.

Gender	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Female	68.4 ± 1.4	71.5 ± 2.1	51.2 ± 2.6	59.6 ± 2.2
Male	76.6 ± 2.5	73.9 ± 3.5	65.3 ± 3.7	69.3 ± 3.3

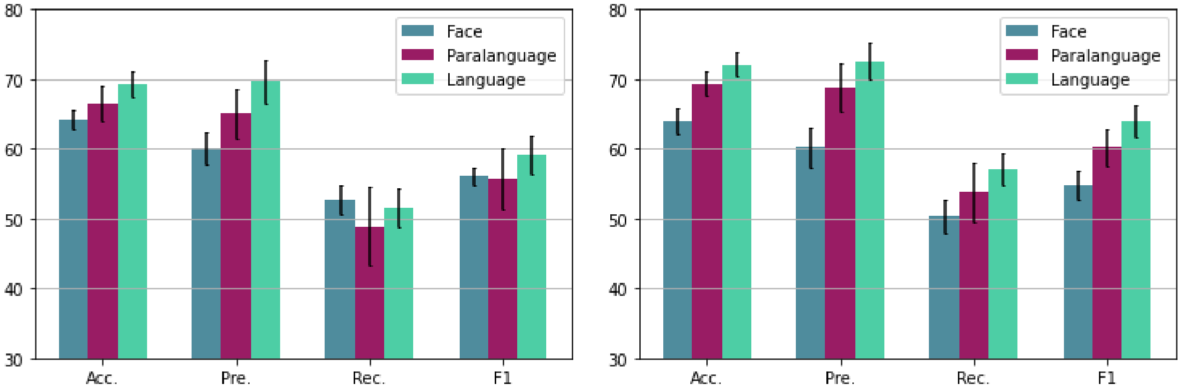
Table 4.11: **Gender-Based Performance of the Language Approach.** This table presents the performance of the language approach for each gender group in terms of averages and standard deviations over R repetitions.

4.5 Performance Comparisons

In this section, a comparative discussion of the performance of the above three unimodal approaches is provided. This includes comparisons of these approaches in terms of overall performance, performance at the story level, and performance with respect to age and gender variations. The primary goal of this section is to highlight the similarities and differences between approaches which may provide useful insights on how to better exploit these relationships across different approaches.

4.5.1 Overall Performance

Figure 4.5 presents the overall performance of the unimodal approaches after aggregating all stories using MV and SR aggregation methods with respect to all four evaluation metrics. The figures show that the language approach outperforms other approaches in almost every case to a statistically significant extent ($p < 0.05$) (the only exception is in MV for recall). In addition, the paralanguage approach outperforms the facial expressions approach in the SR method in all metrics ($p < 0.05$), whereas in MV, it only improves over facial expressions in accuracy and precision ($p < 0.05$). Facial expressions, in contrast, only improves over paralanguage in the MV method for recall ($p < 0.05$).



(a) MV

(b) SR

Figure 4.5: **Overall Performance of Unimodal Approaches.** The figures show the performance of the three approaches at the participant level obtained by the aggregation methods, where (a) presents the results obtained with MV, and (b) presents the results obtained with SR.

The figure also shows that in both paralanguage and language-based approaches, the SR method improves over the MV method to significant extents. However, the SR never improved over MV in the facial expressions approach, in fact, the performance in the recall is higher in MV than SR and this improvement is statistically significant ($p < 0.05$). The improvement of SR implies that the approach tends to assign a higher posterior probability to a certain class c when c is the true class i.e., the approach tends to be more confident when making the right decisions. Hence, it seems that both paralanguage and language approaches are more confident while this is not the case for the facial expression approach (the lower performance in SR-recall suggests that the approach tends to assign higher posterior probability to the wrong class).

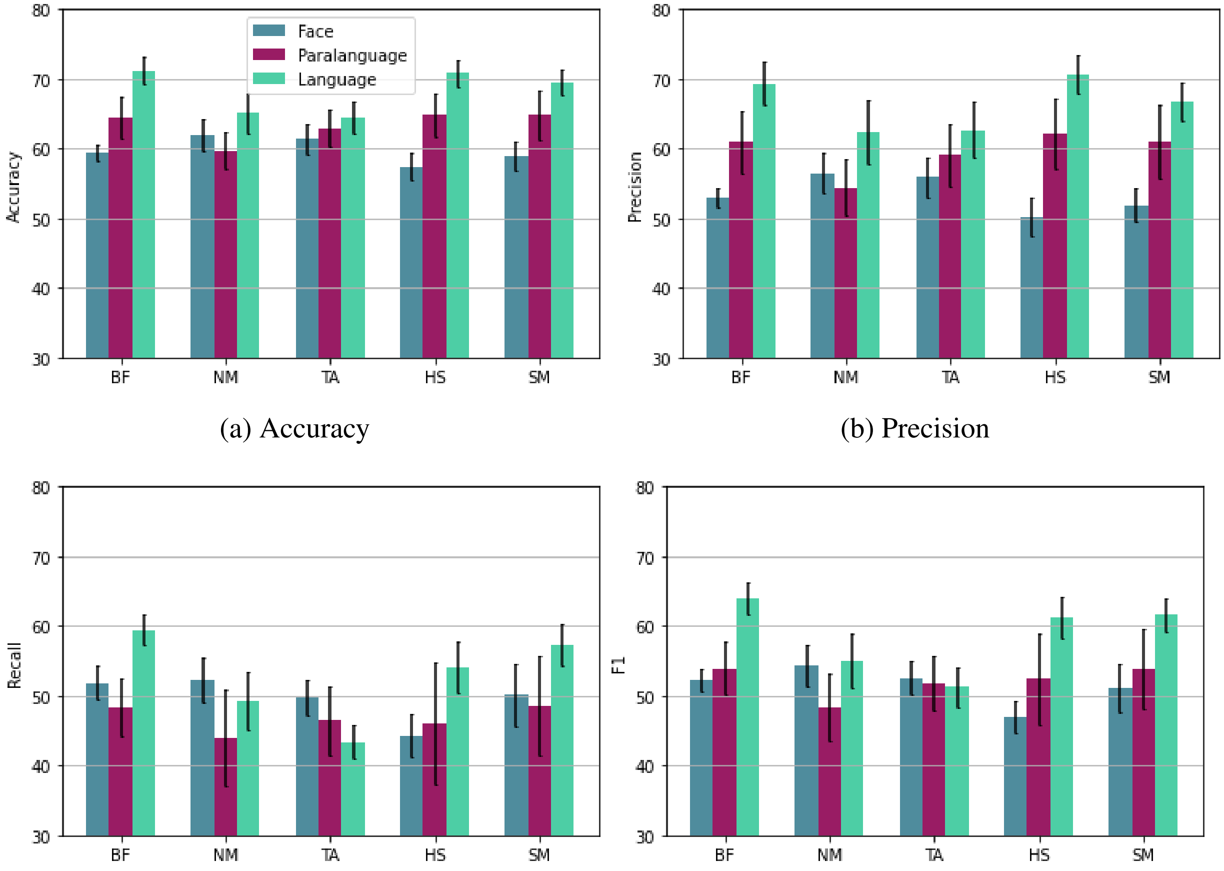
Another aspect of this figure is that the paralanguage approach has higher standard deviations (error bars) compared to other approaches, which is especially apparent in the recall, suggesting that the approach is more sensitive to weight initialisations.

4.5.2 Story-Level Performance

This section provides a discussion on how individual stories perform across all approaches and with respect to the various metrics. This may help provide a better understanding of what types of scenarios are more likely to elicit attachment cues for each behavioural channel and whether different stories are more likely to elicit cues from different attachment classes. Figure 4.6 shows the performance obtained over individual stories in all approaches. The most apparent finding in these charts is that the language approach achieves statistically significant improvements compared to other approaches ($p < 0.01$) in at least 3 out of 5 stories (BF, HS, and SM) in all metrics, in fact in precision, the language approach improves over other approaches in all stories ($p < 0.05$). The only case where other approaches outperform the language approach significantly is in the recall metric in which facial expressions outperforms the language in two stories (NM, and TA) ($p < 0.05$, $p < 0.001$) respectively. This suggests that the language approach is capturing more relevant attachment cues, especially in three stories (BF, HS, and SM).

When taking into account the two remaining approaches, paralanguage and facial expressions only, different patterns are noted across different metrics. For instance, in accuracy and precision, paralanguage outperforms facial expression in 3 and 4 stories respectively (BF, HS, and SM) plus TA in precision ($p < 0.01$). However, in recall, facial expressions outperforms paralanguage in three stories (BF, NM, and TA) ($p < 0.05$) while in the remaining two stories both approaches achieve similar performances. In F1-score, facial expressions achieves higher performance in NM ($p < 0.01$) while paralanguage achieves better performance in HS ($p < 0.05$), other stories achieve similar performance in both approaches. This contrasting pattern of performance implies that these two approaches tend to make different types of mistakes over different stories, for instance, paralanguage seems to make more false negative predictions over NM (lower recall) compared to facial expressions, while it makes lower false positive errors over HS (higher precision). This may constitute a source of diversity on the classifiers from these two approaches.

A major finding that emerges from the figures is that the individual stories that perform better than others are similar in language and paralanguage (BF, HS, and SM) this especially appears in accuracy and precision, however, by looking into the face approach, a different finding is noted where the best-performing stories are (NM and TA) i.e., the worst performing stories in



(a) Accuracy

(b) Precision

(c) Recall

(d) F1-score

Figure 4.6: **Story-Level Performance in All Approaches.** The figures show the performance of individual stories in all approaches with respect to each one of the four evaluation metrics where each subfigure presents one metric. The legend of subfigure (a) is shared across all subfigures.

paralanguage and language approach is the best performing in facial expressions approach. As both former approaches rely on speech data for recognition (what they say and how they say it), possible reason for low performance is that children are less interested in interacting with NM and TA stories verbally[8], while they may still show significant facial expression behaviours during both induction and completion phases.

Another interesting observation is that in all stories except TA, there are statistically significant differences between the performance of certain approaches in all metrics meaning that the approaches are capturing different attachment cues over different stories. However, this is not the case in the TA story as it seems that all approaches achieve comparable performance in accuracy in F1-score.

[8] This is where both paralanguage and language achieve their lowest performances.

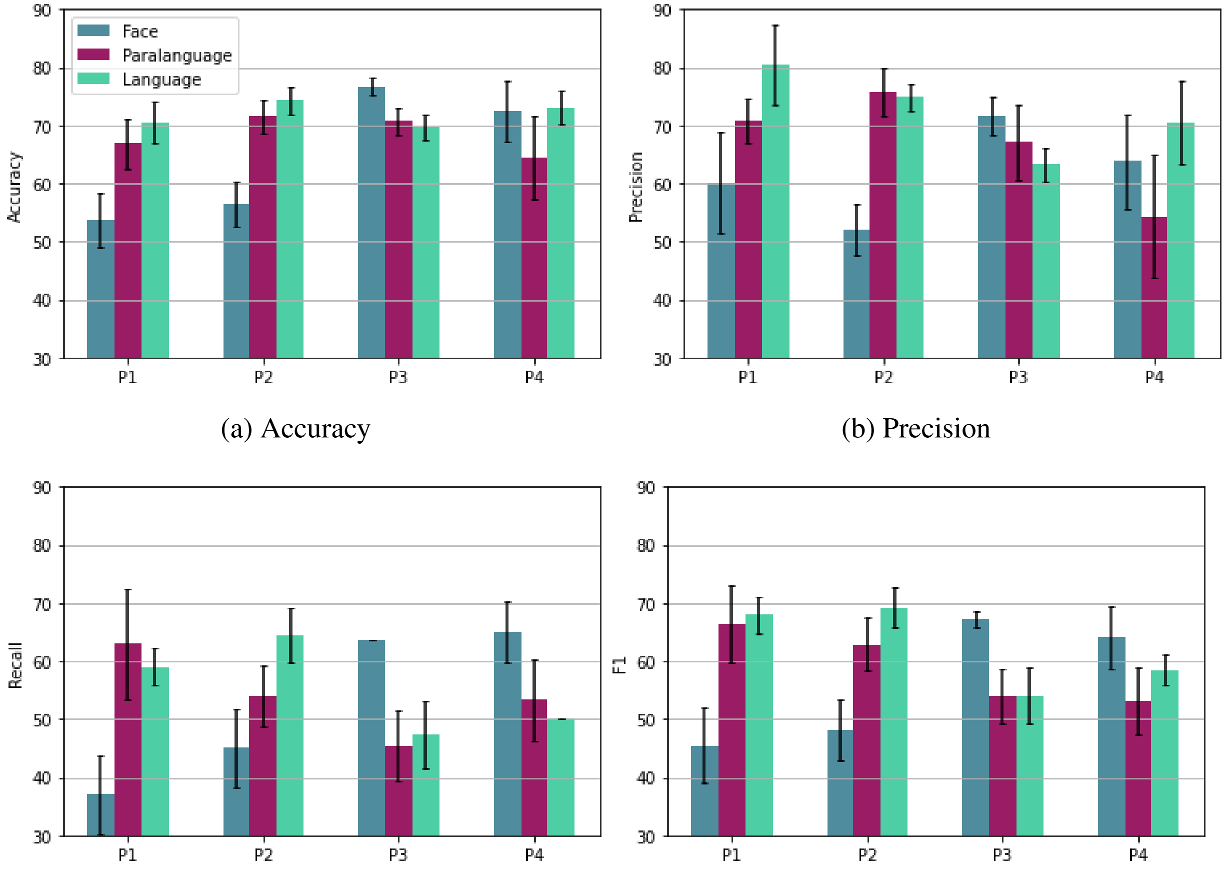
4.5.3 Age-Based Performance

This section compares the performance in each age group in all approaches with respect to different metrics. As discussed above, The age of the children that participated in this study ranges from 5 to 9 years old, and hence it covers a wide range of cognitive developmental stages. This difference may have an influence on the performance of our approaches as children may show different capabilities and understandings in the way they respond and interact with the system.

Figure 4.7 shows the performance of all approaches with respect to each age group. As the figure shows, there are, almost in every case, statistically significant differences between different approaches for certain groups and metrics. The most prominent result of this figure is that in levels P1 and P2 (younger children, 5-7 years old), both paralanguage and language perform better than facial expressions in all metrics ($p < 0.01$). Additionally, facial expressions exhibit a sharp decrease in recall for those groups, suggesting that younger children are more likely to express their condition through language and paralanguage. In contrast, in levels P3 and P4 (older children, 7-9 years old), the facial expressions approach performs significantly better than other approaches for recall and F1-score ($p < 0.01$).

The results above suggest that younger children are more likely to express their condition through speech (what they say and how they say it), while older children are more likely to express it through facial expressions. One possible explanation is that older children may feel too old for the doll-playing task and therefore may be less likely to interact verbally during the sessions, however, they may still show significant facial behaviours. On the other hand, as discussed earlier, younger children may not be developed enough to show meaningful facial expressions (related to emotions) compared to their older counterparts [128].

For older groups (P3 and P4), the improvements of the facial expressions approach over other approaches are more pronounced in the case of recall (compared to other metrics), suggesting that facial expressions approach is more likely to identify the positive cases of these groups. On the other hand, paralanguage and language are more likely to identify negative cases of these groups as indicated by the higher accuracy and lower recall.



(a) Accuracy

(b) Precision

(c) Recall

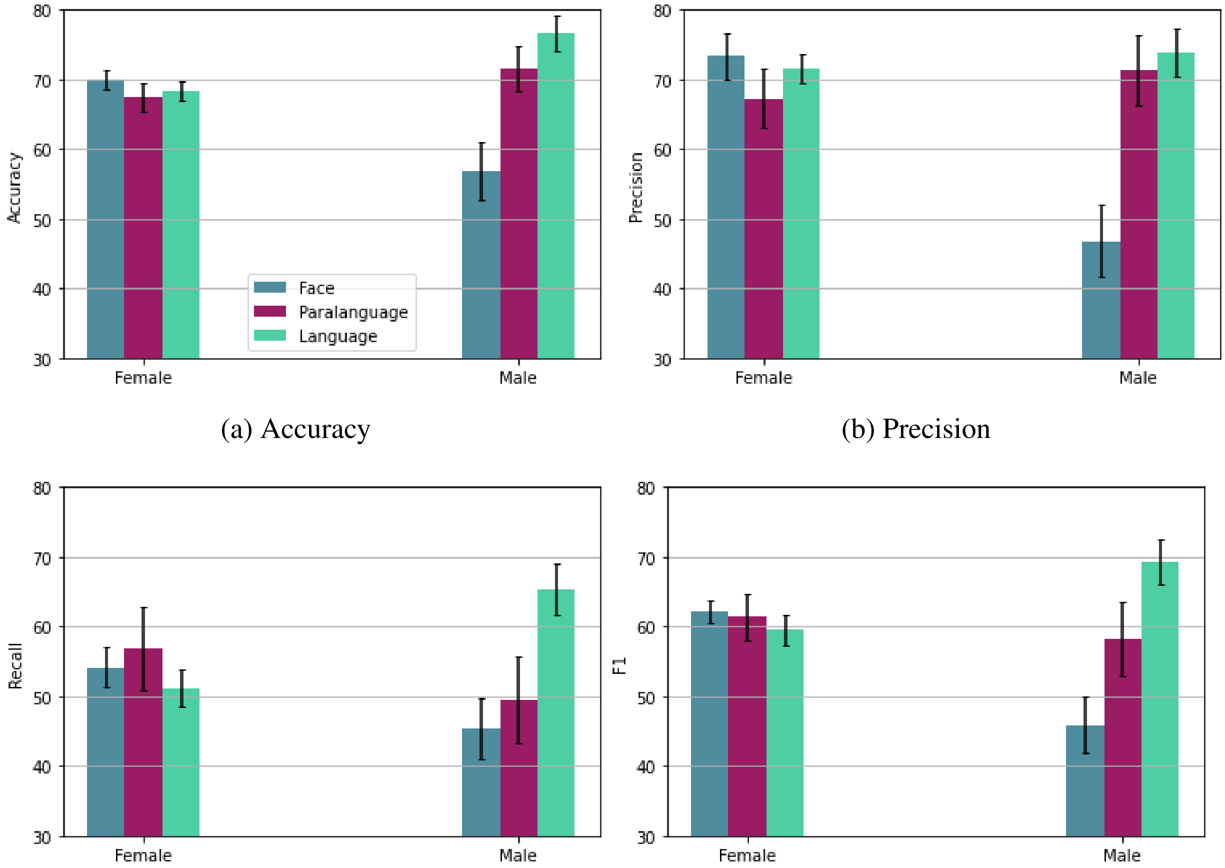
(d) F1-score

Figure 4.7: **Age-Based Performance in All Approaches.** This figure shows the performance of different age groups in all 3 approaches with respect to the 4 evaluation metrics. The legend of the first subfigure is shared across all subfigures.

4.5.4 Gender-Based Performance

This section provides a comparative discussion about the effects of gender variations on the performance of the approaches. Figure 4.8 shows the performance of all approaches with respect to each gender group. As the figure shows, there are more variations in the performance among different approaches for male children compared to female counterparts. For instance, for male children, both paralanguage and language outperform facial expressions in accuracy, precision, and F1-score ($p < 0.001$), and language outperforms other approaches in recall ($p < 0.001$). On the other hand, for female children, there are less variations among different approaches where the biggest variation is in precision between facial expressions and paralanguage ($p < 0.01$). Interestingly, for female children, the facial expressions approach is the best or comparable to the best-performing approach in all metrics whereas it is the worst-performing approach for male children in all metrics. This result suggests that female children are more likely to express their attachment condition through facial expressions while male tends to express it more through language.

By taking into account only the recall metric, the best approach for male participants (language) is significantly better than the best approach for females (paralanguage) ($p < 0.001$) meaning that insecure male children can be identified to a better extent compared to their female peers.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1-score

Figure 4.8: **Gender-Based Performance in All Approaches.** The figure shows the performance of different gender groups in all 3 approaches with respect to the 4 evaluation metrics. The legend of the first subfigure is shared across all subfigures.

4.6 Chapter Summary

This chapter presented several automatic approaches to detect attachment styles in school-age children while undergoing the MCAST assessment test. These approaches are based on analysing the main behavioural channels that convey information about the inner feelings and mental states. These channels are facial expressions, paralanguage, and language. The approaches are validated on a dataset of 104 children and the results show that it is possible to detect attachment in children with an accuracy and F1-score of 72.1%, and 63.9% respectively, which confirms that attachment styles leave machine-detectable behavioural cues.

The results also show that some stories are more likely to elicit attachment cues depending on the different approaches. This confirms the motivation behind having multiple story stems in the MCAST. Interestingly, while the breakfast story (BF) initially aimed to serve as an introduction to the task in which children familiarise themselves with the interviews and was not expected to elicit any attachment behaviour, our results show that, in many cases, BF outperforms other stories in different metrics which mostly appears in the language approach. As no distress scenario is being conveyed in this story, this result suggests the validity of designing non-distress scenarios for attachment detection [11].

Various conclusions were drawn regarding the effect of age and gender variations. For instance, the results suggest that younger children are more likely to show their attachment through paralanguage and language while older children are more likely to express it through facial expressions. Moreover, insecure male children can be identified to a better extent compared to insecure female children. These findings imply that it might be worthwhile assigning different approaches for different age, or gender groups. Finally, the differences in the performance over different stories and with different metrics imply that the approaches are capturing different patterns. In other words, it is likely that different modalities carry complementary information which motivates conducting multimodal experiments. This will be the focus of the next chapter.

Chapter 5

Multimodal Approaches

5.1 Overview

This chapter focuses on multimodal approaches that are based on combining the three unimodal approaches described in the previous chapter. Four different approaches were attempted comprising all possible combinations of the three unimodal approaches namely, facial expressions, paralanguage, and language. The chapter starts by presenting the overall multimodal approach and the results. Then, it describes each approach individually by looking into how individual stories perform and how this performance changes with respect to the unimodal components. The last few sections deal with the effect of age and gender variations on the performance.

The main motivation behind employing a multimodal approach is that humans often express their internal states through a combination of behavioural channels, for example, uttering the same words with different intonations or facial expressions can convey a different message to the listeners. Therefore, in the areas that deal with emotional and mental state recognition, better recognition rate is often achieved by building an approach that takes into account the behavioural cues emitted from multiple behavioural sources. In addition, previous studies showed that multimodal combinations have led to better performances compared to their unimodal components [138, 139]. Moreover, it is shown mathematically in [107] that a simple majority voting scheme is guaranteed to improve over the individual components given that these components are statistically independent. Furthermore, the results in chapter 4 indicate that different unimodal approaches are capturing different patterns over different stories (see section 4.5.2), implying that classifiers from each approach are diverse and the combination may lead to better performance.

5.2 Approach

The proposed multimodal approach combines the unimodal predictions using the Sum Rule (SR) method (see section 2.2.6) i.e., in the same way multiple predictions made at the story level were combined within the unimodal approaches to make the final participant-level prediction. Specifically, each unimodal approach resulted in five predictions per participant, and a late fusion scheme was implemented by combining the predictions from multiple modalities and assigning the participant to the class according to the SR method.

The reason for applying the SR method instead of a simple majority voting MV is due to the improved results of SR over MV that were observed in the previous chapter especially in paralanguage and language approaches so it is reasonable to base the combinations on the best results achieved at the unimodal levels. Moreover, the MV method can harm the performance when there is an even number of predictions (which is the case in most of our multimodal combinations which have 10 predictions per participant) because it can not break the ties in the case of an equal number of votes. In addition, using a method that takes into account the estimate of posterior probabilities produced by the recognition models may lead to better performance in some cases if the corresponding models are likely to assign a higher posterior probability to the true class (which seems to be the case in paralanguage and language, as the results of chapter 4 seem to indicate).

All possible multimodal combinations are attempted in this study. This creates four different multimodal approaches, namely: Face-Paralanguage (FP), Paralanguage-Language (PL), Face-Language (FL), and All modalities (FPL). Therefore, 10 to 15 predictions per participant were combined (depending on the number of modalities involved in the combinations). Figure 5.1 depicts the overall architecture of the multimodal approaches.

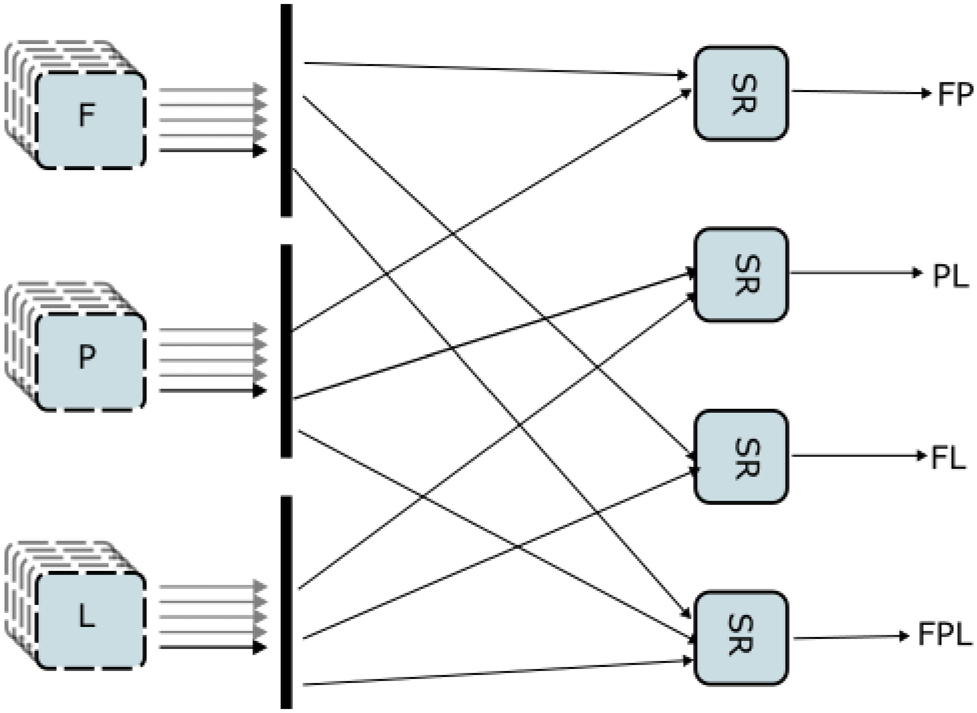


Figure 5.1: **Multimodal approaches Architecture.** This figure shows the overall architecture of the multimodal approaches. F, P, and L boxes stand for face, paralanguage, and language approaches, respectively. The vertical bold lines merge the multiple incoming paths into one path i.e., the first SR box has 10 incoming paths while the last one has 15 incoming paths. The SR takes the posterior probabilities made by each unimodal component and assigns the input sample to the class with the highest result according to equation 2.12.

5.3 Results

This section presents the results for each multimodal approach starting by presenting the overall results and subsequently presenting a detailed description of the performance of each multimodal approach where a story-level performance is discussed along with its unimodal components to inspect the effect of multimodal combinations at the story-level.

5.3.1 Overall Performance

Table 5.1 presents the overall results of each multimodal approach as well as the overall results of the unimodal approaches for the sake of clarity of discussions. Each experiment is presented in terms of four evaluation metrics to provide more accurate measures at class level given the imbalance of classes in the dataset. As shown in the table, the best overall performance was achieved through the combination of paralanguage and language modalities (PL), and the face and language modalities (FL), where the former yields 75.6% and 79.7% in accuracy and precision, and the latter yields 75.4%, 62.7% and 68.8% in accuracy, recall, and F1-score, respectively. Interestingly, both approaches achieve comparable performance in terms of accuracy but differ significantly in other metrics, in particular, PL performs better than FL in precision ($p < 0.01$), whereas FL performs better than PL in recall ($p < 0.001$), and F1-score ($p < 0.05$). Both approaches use the language component indicating the importance of including this modality.

Approach	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Face (F)	64.0 ± 1.9	60.2±2.9	50.4±2.4	54.8±2.1
Paralanguage (P)	69.3 ± 1.7	68.8 ± 3.4	53.8 ± 4.3	60.2 ± 2.7
Language (L)	72.1 ± 1.7	72.6 ± 2.6	57.1 ± 2.3	63.9 ± 2.3
Multimodal1 (FP)	72.2 ± 0.8	71.8 ± 1.1	58.9 ± 2.4	64.7 ± 1.5
Multimodal2 (PL)	**75.6 ± 1.0**	**79.7 ± 3.5**	58.7 ± 2.1	67.5 ± 1.0
Multimodal3 (FL)	**75.4 ± 0.9**	76.3 ± 2.1	**62.7 ± 2.3**	**68.8 ± 1.3**
Multimodal4 (FPL)	74.7 ± 1.4	77.8 ± 2.5	58.2 ± 1.8	66.6 ± 1.9

Table 5.1: **Overall Performance of the Multimodal Approaches.** Entries in bold indicate the best approaches in each one of the four metrics.

The combination of all three unimodal approaches (FPL) does not improve over the best multimodal approach in any metric, in fact, the difference with respect to the FL approach is statistically significant ($p < 0.01$) in recall and F1-score. However, in accuracy and precision, FPL achieves comparable performance to both PL and FL. Additionally, FPL achieves better performance compared to FP in three metrics ($p < 0.05$) and similar performance in recall suggesting that combining the language modality contributed to the performance of (FPL) by better identifying the negative cases (secure children). On the other hand, combining the third modality has not improved over PL and FL, in fact, adding paralanguage to FL harms the performance in recall and F1-score meaning that it leads to more false negative errors. A possible explanation of this effect on FL is that both language and paralanguage are detecting similar patterns over positive cases, so the contribution of the face is outweighed by the paralanguage. Moreover, FPL improves over the best unimodal component (the language approach) to a statistically significant extent ($p < 0.01$) in three metrics (all except recall).

Another interesting finding that emerges from the table above is that all of the four multimodal approaches have improved over their unimodal components to a statistically significant extent in accuracy, precision ($p < 0.01$), and F1-score ($p < 0.05$) suggesting that all unimodal components tend to convey complementary information, a condition necessary for modalities to mutually improve over each other. However, in recall, only FP and FL have improved over their best unimodal component ($p < 0.01$), meaning that in these two approaches there are complementary patterns that were captured over positive cases as opposed to the other approaches.

A possible explanation of why FL performs best in recall and F1-score is as shown in chapter 4, that face approach seems to capture different patterns compared to other modalities which can be seen from the performance of individual stories in recall (see figure 4.6) in which (in most cases) the highest performing stories in the face approach are the lowest performing in both paralanguage and language. Moreover, the language approach is generally the best-performing approach in most stories therefore, combining the best-performing modality with the one that captures the significant amount of complementary patterns can be the main reason for yielding the best performance.

5.3.2 Performance of FP Approach

The results of the multimodal combination of face and paralanguage approaches (FP) are presented in Table 5.2. As in the case of the unimodal approaches, the results for each story and their combination are shown in four performance metrics. The performance of the individual stories is made by aggregating the predictions made at the respective stories from each modality (using SR). For example, the performance of Breakfast is made by combining the predictions made at Breakfast in Face and Paralanguage unimodal approaches. Therefore, the story-level results are based on combining two classifiers whereas the overall results are based on combining 10 classifiers.

Story	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Breakfast(BF)	66.7 ± 2.5	62.8 ± 3.8	56.0 ± 3.4	59.1 ± 2.6
Nightmare(NM)	65.9 ± 3.4	62.3 ± 4.0	52.7 ± 7.3	57.0 ± 5.9
Tummyache(TA)	64.8 ± 2.3	61.1 ± 3.0	50.9 ± 4.7	55.3 ± 3.7
Hopscotch(HS)	63.4 ± 2.0	59.9 ± 3.8	44.1 ± 3.4	50.7 ± 2.7
Shop. Mall(SM)	62.0 ± 2.5	57.5 ± 3.6	42.0 ± 5.1	48.5 ± 4.5
All (SR)	72.2 ± 0.8	71.8 ± 1.1	58.9 ± 2.4	64.7 ± 1.5
Random	51.0	43.0	43.0	43.0

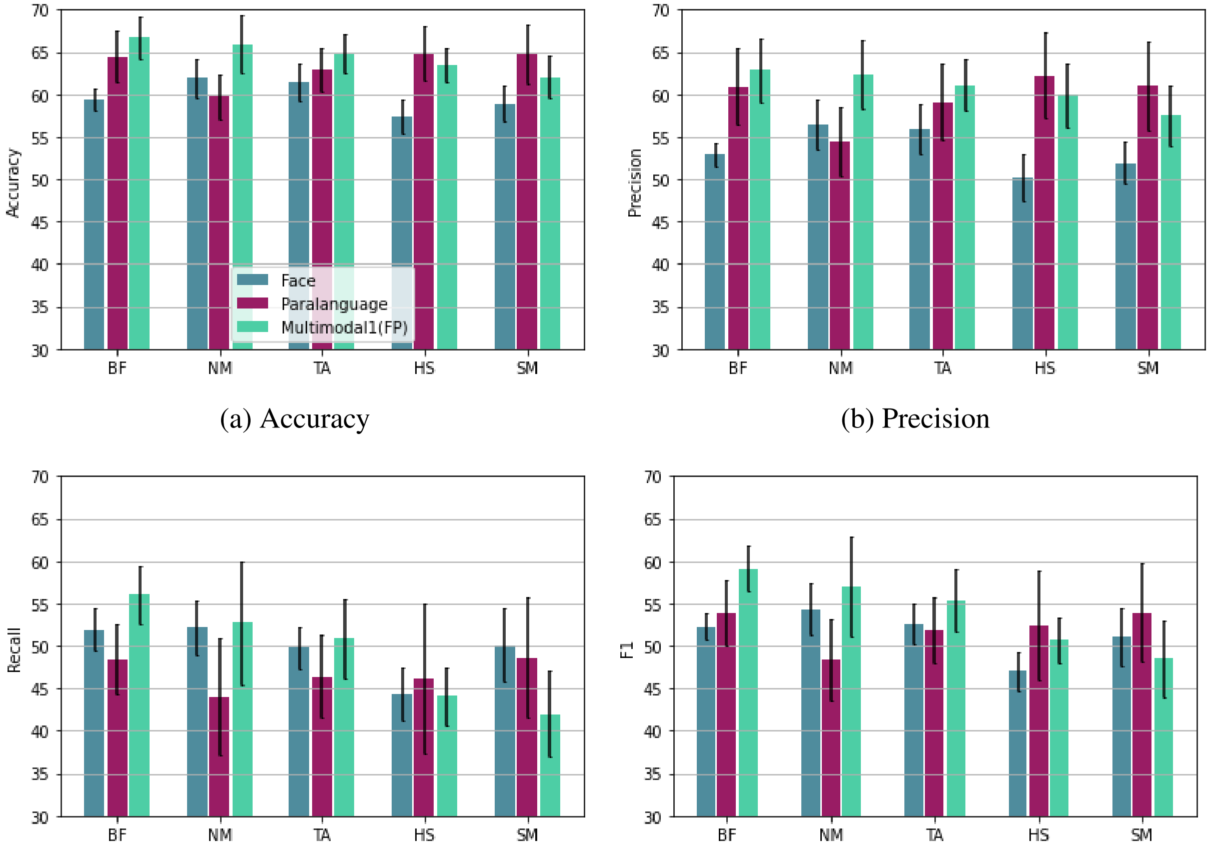
Table 5.2: **FP Overall Performance.** This table presents the performance of the multimodal approach (FP), both at story-level and at participant-level.

As shown in the table, all experiments perform better than the random baseline to a statistically significant extent in all metrics ($p < 0.001$), except in recall for two stories (HS and SM). Moreover, different stories vary in their performance, in particular, BF performs better than the three lowest-performing stories (TA, HS, and SM) in three metrics. On the other hand, in precision, all stories achieve comparable performance except for SM. The combination of all stories from both modalities has improved the performance over the best-performing story in all metrics ($p < 0.001$ for accuracy, precision, and F1-score, $p < 0.05$ for recall) suggesting that different stories from both channels are conveying complementary information. It is worth noting that the improvement in recall is more pronounced in Table 5.1 compared to the present table, suggesting that the improvement in recall is mostly obtained by combining different modalities, rather than by having different stories.

Another way to look at how the multimodal combination behaves is to inspect the story-level performance along with its respective unimodal components. Figure 5.2 presents the performance of each story from both the unimodal components and its multimodal combination. It is important to note that the combination at story-level is based on two classifiers only (one from each modality) and therefore their contribution to enhance the performance could be less noticeable than in the case of participant-level performance. On the other hand, the participant-level combination is based on 10 classifiers, and therefore the improvement can be more observed. However, the results still give useful indications of which stories benefit more from the multimodal combination.

The most apparent aspect that emerges from the figure is that in three stories (BF, NM, and TA), the multimodal combination has improved over the unimodal components at least in two metrics ($p < 0.05$). This is important because it confirms that combining these behavioural channels has contributed to the overall performance i.e., the improvement is not only a result of having multiple attachment scenarios[1]. In recall, the only story that improves over its respective unimodal component is BF. Moreover, NM improved significantly in precision compared to other stories and in any other metrics. Both HS and SM never improved over their individual unimodal-based classifier, in fact, the performance of SM dropped significantly in recall and F1-score (that drop might result from the wrong classifier being more confident in its prediction). These results suggest that this multimodal combination behaves differently at different stories

[1] Which is already confirmed in the previous chapter.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1

Figure 5.2: **FP story-level performance.** Story level performance of the FP multimodal approach along with its unimodal components (Face and Paralanguage). Each subgraph presents the performance with respect to one metric. The legend is shared across all subgraphs.

which is likely to be the result of the fact that different stories elicit different attachment-related patterns.

5.3.3 Performance of PL Approach

Table 5.3 presents the results of the combination of paralanguage and language-based approaches (PL) at story-level and their combination (participant-level). As shown in the table, all experiments perform better than the random guessing classifier in all metrics. The best-performing stories are HS in accuracy and precision, and BF in recall and F1-score, both performing significantly higher than the lowest three stories ($p < 0.05$). The combination of all classifiers has led to a significant improvement over stories in all metrics ($p < 0.01$) except recall which performs as well as the BF story. Additionally, TA and NM are the worst-performing stories in recall ($p < 0.05$) compared to the remaining stories.

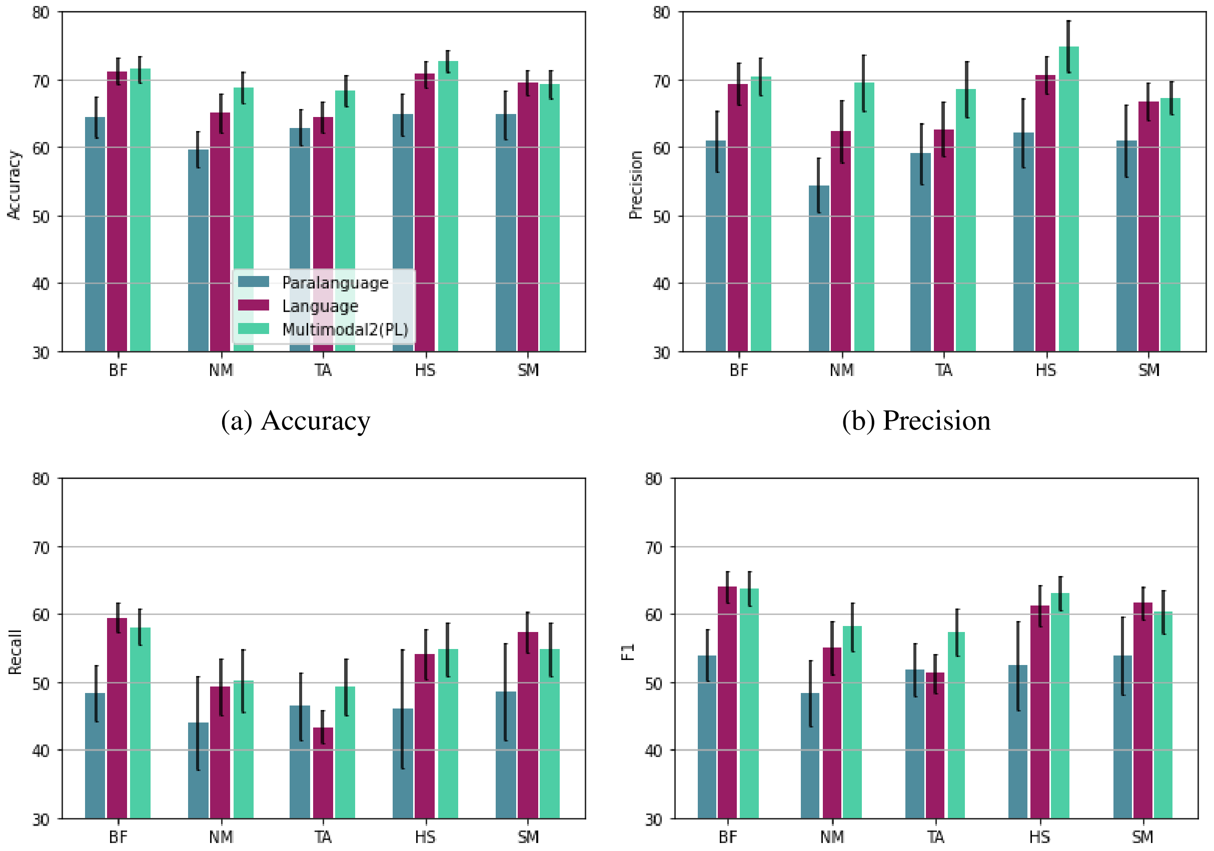
Story	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Breakfast	71.5 ± 2.0	70.4 ± 2.8	58.1 ± 2.7	63.7 ± 2.5
Nightmare	68.8 ± 2.2	69.5 ± 4.2	50.2 ± 4.7	58.2 ± 3.6
Tummyache	68.3 ± 2.3	68.6 ± 4.2	49.3 ± 4.2	57.3 ± 3.4
Hopscotch	72.7 ± 1.7	74.8 ± 3.8	54.8 ± 3.9	63.1 ± 2.5
Shop. Mall	69.3 ± 2.0	67.3 ± 2.5	54.8 ± 3.9	60.4 ± 3.2
All (SR)	75.6 ± 1.0	79.7 ± 3.5	58.7 ± 2.1	67.5 ± 1.0
Random	51.0	43.0	43.0	43.0

Table 5.3: **PL Overall Performance.** This table presents the performance of the multimodal approach (PL), both at story-level and at participant-level.

Figure 5.3 shows how the individual stories perform compared to their respective stories in unimodal components i.e., how the multimodal combination behaves at story-level. Each subfigure presents the performance with respect to one of the performance metrics. As the figure shows, two stories (NM and TA) have improved over their unimodal components in three metrics (all except recall) ($p < 0.05$), while HS has improved in two metrics ($p < 0.05$). Moreover, in recall, none of the stories have improved over the components suggesting that the approaches are capturing the same patterns over positive cases in all stories. The remaining stories (BF and SM) have never improved in any metrics.

5.3.4 Performance of FL Approach

Table 5.4 presents the results of the FL multimodal approach which is based on combining face and language unimodal components. As shown in the table, all experiments perform better than the random guess classifier, and the combination of all stories (participant-level) has led to better performance in three metrics ($p < 0.05$), where in recall the aggregation does not outperform the best individual story (BF). Moreover, there are variations in the performance of individual stories where BF is the best-performing story in all metrics compared to other stories ($p < 0.01$) except in precision. For precision, both BF and HS outperform other stories to a significant extent ($p < 0.05$). Furthermore, NM story is among the worst-performing stories in all metrics. TA and HS show opposite performances, in which TA achieves better recall whereas HS achieves better precision and both have comparable performances in accuracy and F1-score, suggesting that these two stories tend to make different types of mistakes, in other words, TA identifies positive cases to a better extent whereas HS identifies negative cases to a better extent. As in



(a) Accuracy

(b) Precision

(c) Recall

(d) F1

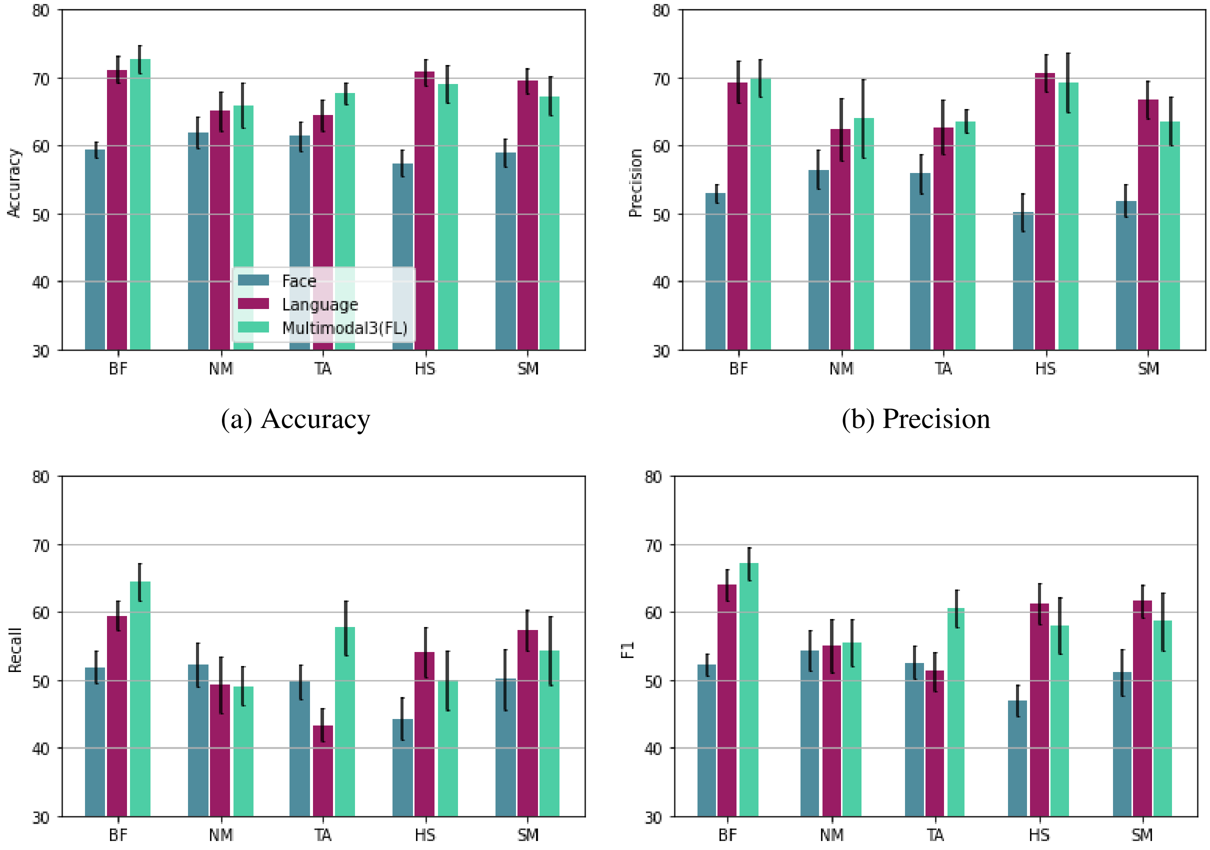
Figure 5.3: **PL story-level performance.** Story level performance of the PL multimodal approach along with its unimodal components (Paralanguage and Language). Each subgraph presents the performance with respect to one metric. The legend is shared across all subgraphs.

the FP approach, Table 5.1 suggests that the improvement in recall of the FL approach seems to result from having different modalities rather than having different stories.

Story	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Breakfast	72.8 ± 2.1	70.0 ± 2.8	64.4 ± 2.7	67.1 ± 2.5
Nightmare	65.9 ± 3.3	64.0 ± 5.7	49.1 ± 2.9	55.5 ± 3.4
Tummyache	67.6 ± 1.6	63.6 ± 1.7	57.7 ± 4.0	60.5 ± 2.8
Hopscotch	69.1 ± 2.8	69.2 ± 4.4	50.0 ± 4.4	58.0 ± 4.2
Shop. Mall	67.3 ± 2.9	63.6 ± 3.5	54.3 ± 5.1	58.6 ± 4.3
All (SR)	75.4 ± 0.9	76.3 ± 2.1	62.7 ± 2.3	68.8 ± 1.3
Random	51.0	43.0	43.0	43.0

Table 5.4: **FL Overall Performance.** This table presents the performance of the multimodal approach (FL), both at story-level and at participant-level.

Figure 5.4 shows how individual stories perform compared to their respective stories in the unimodal components across all metrics. As the figure shows, the patterns at the story level are slightly different from what was observed in the above two multimodal approaches. In particular, two stories (BF and TA) have improved in all metrics except precision ($p < 0.05$), whereas the remaining three stories have not improved in any metric, in fact, the difference between the language component and FL is statistically significant for (HS and SM) in at least two metrics ($p < 0.05$). In precision, none of the stories outperforms the best unimodal (language) suggesting that the FL combination performs better at identifying positive cases (insecure children). Interestingly, the biggest improvement was achieved in TA, especially in recall, although the language component performs only at random level. In recall, BF and TA perform better than the unimodal and these two stories could be the main reason for the overall performance of the FL approach (best multimodal approach in recall and F1-score).



(a) Accuracy

(b) Precision

(c) Recall

(d) F1

Figure 5.4: **FL story-level performance.** Story level performance of the FL multimodal approach along with its unimodal components (Face and Language). Each subgraph presents the performance with respect to one metric. The legend is shared across all subgraphs.

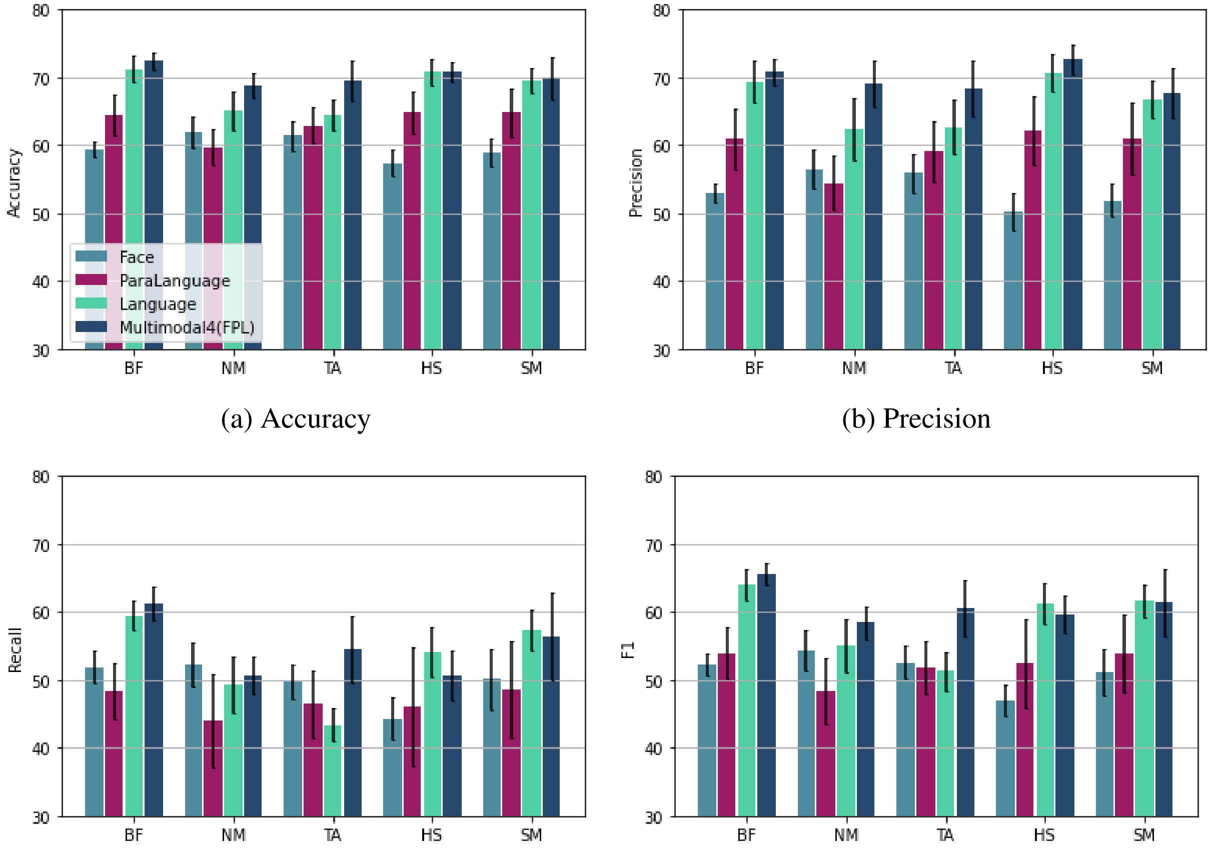
5.3.5 Performance of FPL Approach

Lastly, Table 5.5 presents the results of the FPL approach which is based on combining all of the three modalities. As the table shows, there are variations in the performance across different stories where BF is the best-performing story in three metrics (all except precision) ($p < 0.05$). Combining all the stories improved the performance in accuracy and precision ($p < 0.001$), however, in recall and F1-score, the combination has not improved over the BF story. In fact, the performance in BF is significantly higher than the combination in recall ($p < 0.01$). These findings suggest that combining the three modalities improves the precision to a greater extent i.e., it makes fewer false positive errors.

Story	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Breakfast	72.4 ± 1.2	70.8 ± 2.0	61.2 ± 2.5	65.6 ± 1.6
Nightmare	68.8 ± 1.8	69.0 ± 3.4	50.7 ± 2.7	58.4 ± 2.4
Tummyache	69.5 ± 2.9	68.4 ± 4.1	54.5 ± 5.0	60.6 ± 4.1
Hopscotch	70.8 ± 1.5	72.7 ± 2.2	50.7 ± 3.7	59.6 ± 2.8
Shop. Mall	69.9 ± 3.1	67.7 ± 3.7	56.4 ± 6.4	61.4 ± 5.0
All (SR)	74.7 ± 1.4	77.8 ± 2.5	58.2 ± 1.8	66.6 ± 1.9
Random	51.0	43.0	43.0	43.0

Table 5.5: **FPL Overall Performance.** This table presents the performance of the multimodal approach (FPL), both at story-level and at participant-level.

Figure 5.5 shows how individual stories perform compared to the respective unimodal components. As the figure shows, TA story has improved in all metrics ($p < 0.01$), whereas (BF and NM) have improved in at least two metrics ($p < 0.05$). Moreover, HS improves in precision ($p < 0.05$) however in recall it drops significantly compared to the language component. Again, SM has never improved in any metric over the best unimodal component (language). The most significant improvement is yielded by TA although the performance has not exceeded other stories such as (BF and HS). These results seem to suggest that the approaches are capturing the most complementary information over TA story. However, the relatively poor performance at unimodal components over this story has led to overall less or similar performance compared to other stories.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1

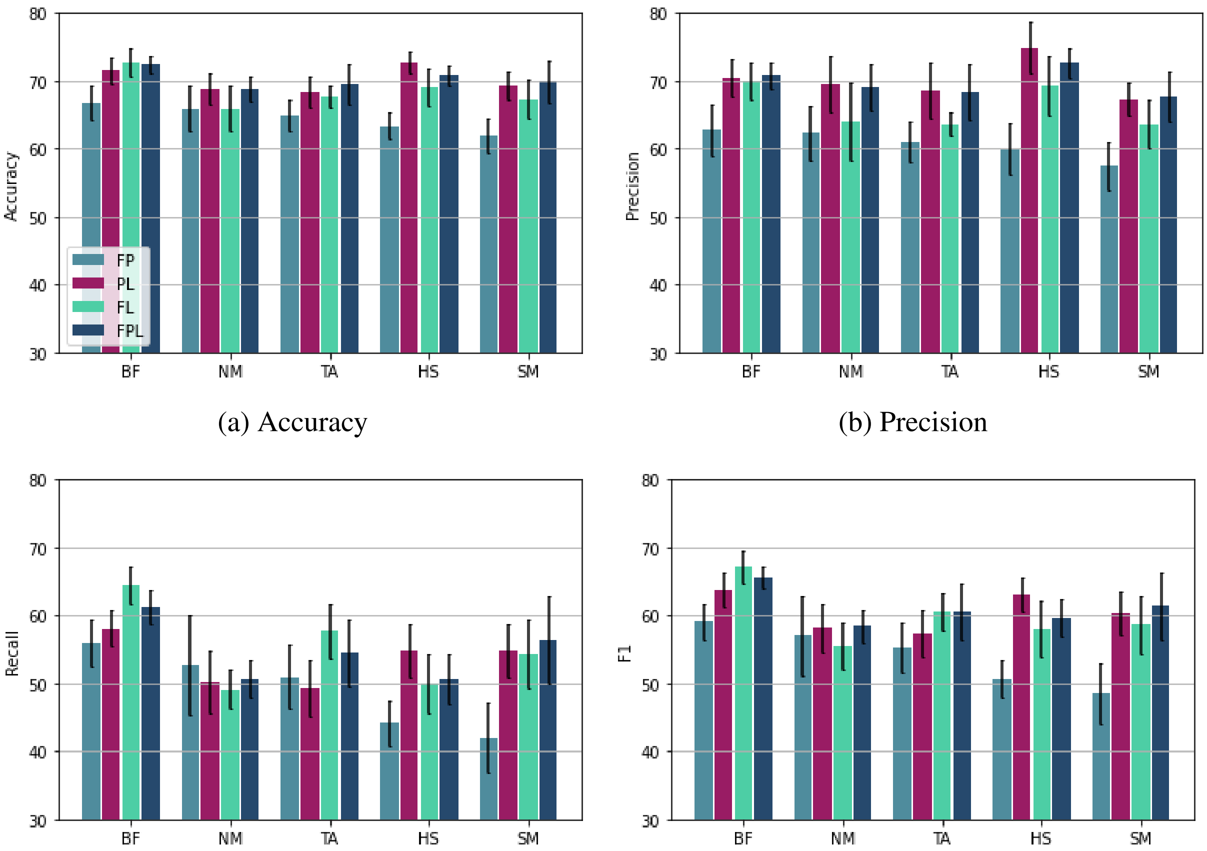
Figure 5.5: **FPL story-level performance.** Story level performance of the FPL multimodal approach along with its unimodal components (Face, Paralanguage, and Language). Each subgraph presents the performance with respect to one metric. The legend is shared across all subgraphs.

5.4 Multimodal Approaches Comparative Analysis

Following the same way as in the previous chapter, this section is dedicated to comparing the performance of the different multimodal approaches with respect to three different aspects: story-level, age-based, and gender-based performance. Certain variations in the behaviours of the approaches are expected because of how these aspects vary in the unimodal approaches. Moreover, another source of variation is expected to result from the multimodal combinations.

5.4.1 Story-Level Performance

The first set of comparisons is performed between the performance of each story in various multimodal approaches. These comparisons are carried out with respect to each evaluation metric. Figure 5.6 shows the performance of each story and each subfigure shows the performance with respect to one metric. As shown in the figure, in all cases there are significant variations among certain approaches in different stories.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1

Figure 5.6: **Story-level Performance in Multimodal Approaches.** This figure shows the performance of individual stories for all multimodal approaches. Each subfigure corresponds to one metric. The legend is shared across all subfigures.

Firstly, the FP approach is generally the worst performing approach especially regarding three stories (BF, HS, and SM) highlighting the importance of including the language approach in these stories. For the remaining stories, FP remains one of the worst performing approaches in all metrics. The only case where this approach is among the best performing ones is in NM in both recall and F1-score. This is not surprising considering the overall low performance achieved in this approach as shown in section 5.3.2.

PL approach (best overall performance in precision) is among the best performing approaches in all stories for accuracy and precision. For other metrics, PL is among the best performing approaches in three stories (NM, HS, and SM). Interestingly, PL is always the best performing approach in (HS) which significantly outperforms other approaches in three metrics ($p < 0.05$)

(all except F1-score). One advantage of this approach is its relatively higher precision indicating that one can trust its positive predictions to a better extent.

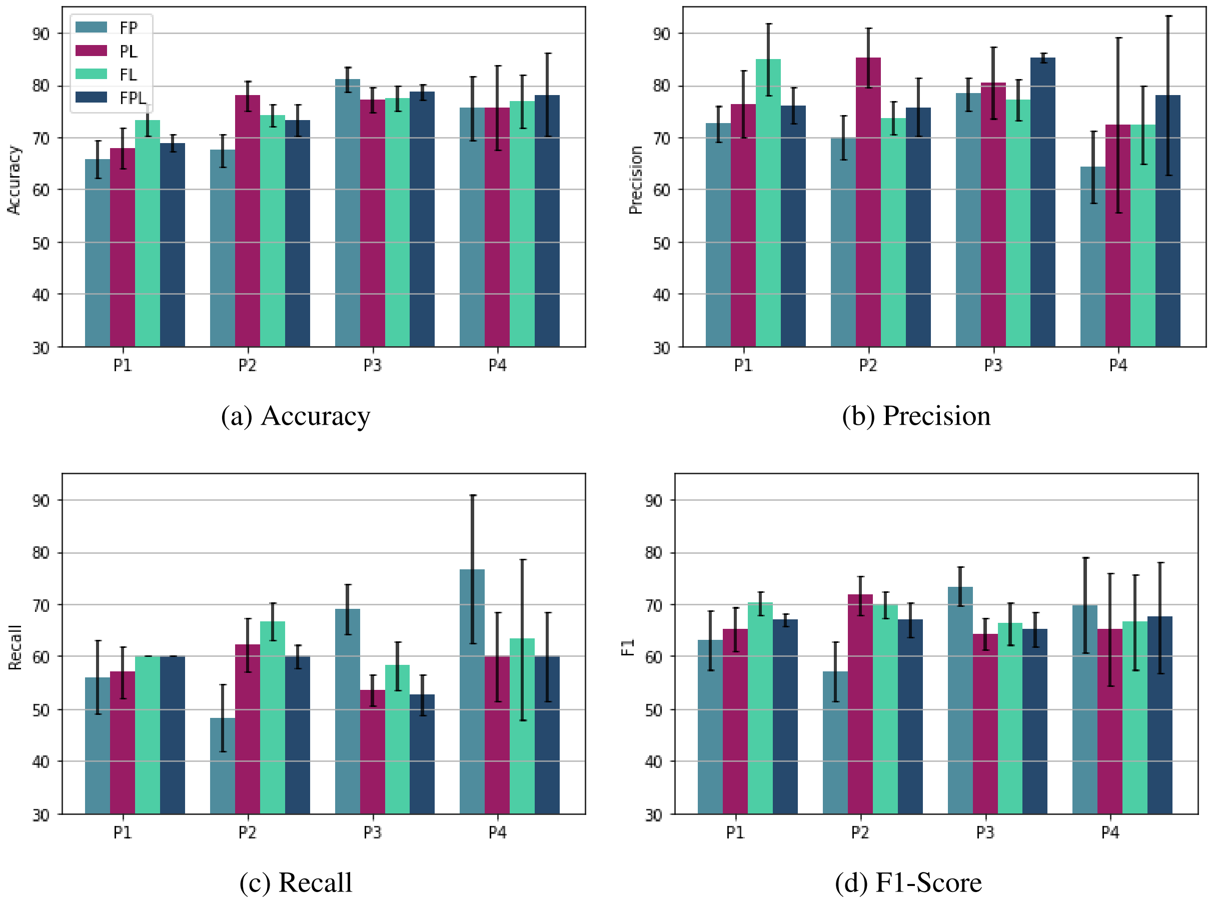
FL (best overall performance in recall and F1-score) is among the best performing approaches in fewer cases compared to PL. Where in accuracy and precision, FL is among the best performing only in (BF). In recall, FL outperforms other approaches in one story (BF) ($p < 0.01$), whereas in (TA) it achieves slightly better performance compared to other approaches. As shown in the previous section, this result also suggests that these two stories are the biggest contributors to the overall higher recall of the FL approach. The relatively lower precision, suggests that this approach tends to make more false positive errors, however, in medical diagnosis systems these errors are less risky than the opposite type of errors.

The last approach FPL is among the best approaches in all but a few cases, in particular, in (HS) this approach performs less than PL in all metrics suggesting that combining face component harms the performance of this story. Additionally, the approach performs worse than FL in (BF) for recall and F1-score. Interestingly, the approach always outperforms FL and FP in accuracy and precision, indicating that combining paralanguage and language improves the ability to identify the negative cases (secure children), on the other hand, (in some cases) the approach outperforms PL in recall and F1-score, suggesting that combining face improves the ability to identify the positives cases (insecure children).

5.4.2 Performance with Age Variations

The age variations among participants are also expected to affect the performance of the multimodal approaches due to the effects that were observed in the unimodal approaches and another source of variation may emerge from the combination. To verify this effect, the performance in all metrics of each approach is presented in Figure 5.7 where each subfigure shows the performance with respect to one metric.

The figure shows that in most cases, there are statistically significant differences between certain approaches in certain age groups meaning that the approaches do not perform equally well in all groups. For instance, in accuracy, FL performs best for P1, PL performs best for P2, and FP performs best for P3 ($p < 0.01$), and all approaches achieve similar performances for P4. Combining all modalities in FPL has never outperform any approach at any age group. Precision



(a) Accuracy

(b) Precision

(c) Recall

(d) F1-Score

Figure 5.7: **Age-Based Performance in Multimodal Approaches.** Performance across different age groups in multimodal approaches. Each subfigure presents one metric. The legend is shared across all subfigures.

shows the same pattern as in accuracy for levels P1 and P2. For P3, FPL works better than other approaches ($p < 0.05$). For the first three levels, the best precision is above 85% meaning that, by using different combinations, the positive predictions can be trusted to a better extent compared to the negative predictions.

In recall, P1 shows less variation between approaches compared to other levels where FL and FPL outperform other approaches ($p < 0.05$). This is not the case for other levels where a higher degree of variation is noted. In P2, the FL approach performs better than other approaches ($p < 0.05$). The most interesting finding is that in P3 and P4, FP performs significantly higher than other approaches despite the overall low performance. This is interesting, as for those levels the face approach achieved higher recall compared to other groups (see figure 4.7) and by combining paralanguage, the identification rate of positive cases has increased significantly indicating the substantial amount of complementary patterns detected by these unimodal components over these groups. Finally, the performance in F1-score shows approximately a similar pattern as in the accuracy in the sense that for each level, the best performing approach is the same in both accuracy and F1-score.

In general, it seems that for younger children it is best to utilise the language approach whereas for older children it is best to utilise facial expressions. Moreover, the best approach for younger groups P1 and P2 corresponds to the one that achieves the best precision while for P3 the best approach corresponds to the approach with the best recall. The figures also show higher standard deviations in P4 in all metrics which is attributed to the lower number of participants belonging to this group (P4 accounts for 15% of the participants) with even lower positive cases.

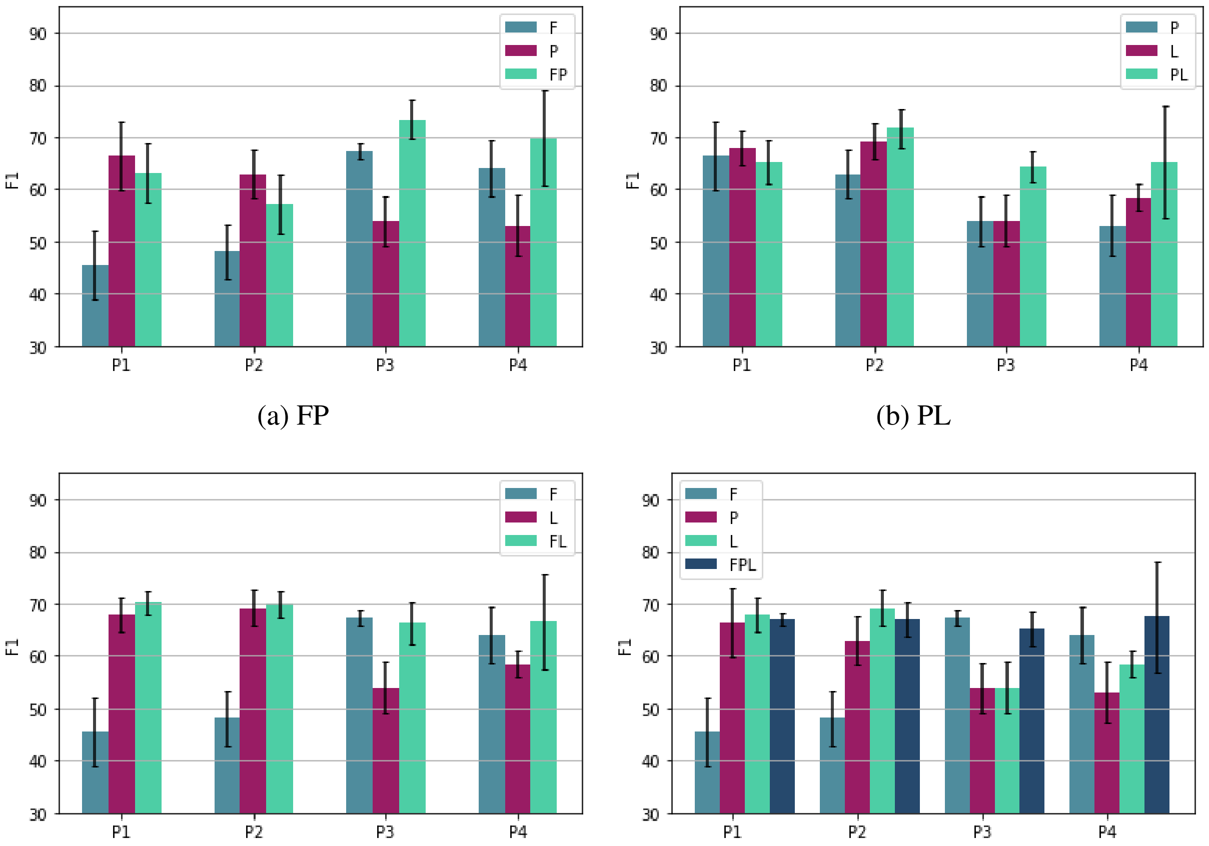
Another way to examine the effect of age variations is to inspect how each multimodal approach performs compared to its unimodal components (for brevity, this is only shown for one metric, F1-score) which is presented in Figure 5.8. As shown in the figure, in FP, only level P3 improves over the components ($p < 0.001$) while in PL, both P3 and P4 improve over the components ($p < 0.001$, $p < 0.05$). FL only improves for level P1 ($p < 0.05$) while FPL does not outperform the best component in any levels.

In addition, both PL and FL did not lead the performance to degrade at any level. This is in contrast to FP, where it performs worse than paralanguage for level P2, and FPL which performs worse than face in level P3 ($p < 0.05$). Thus, given the observation that was discussed in section 4.5.3 which suggests that older children tend to express their attachment condition more through facial expressions while younger one tends to do it through speech[2], it seems that the FL is the ideal multimodal approach to account for age variations because it performs as well as the best unimodal approach in each age level (language for younger children and face for older ones). This can be seen by the less spread of the green bars in subfigure (c) compared to the other subfigures.

5.4.3 Performance with Gender Variations

This section presents the effect of gender-based variations in the multimodal approaches. Figure 5.9 presents the results for all approaches and with respect to each metric. As the figure shows, there are significant variations in how the approaches perform in both gender groups. For in-

[2] Corresponding to both paralanguage and language.



(a) FP

(b) PL

(c) FL

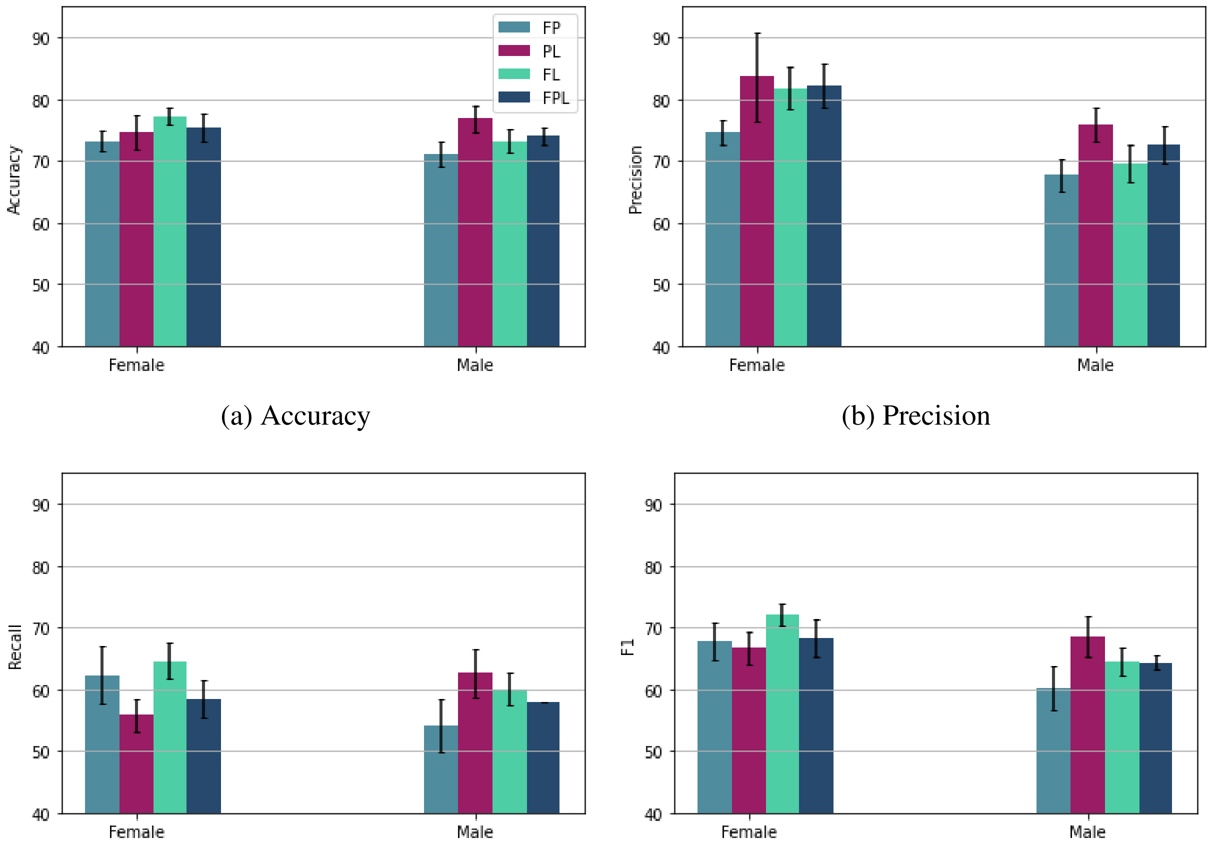
(d) FPL

Figure 5.8: **Age-Based Performance of Multimodal Approach and Unimodal Components.** Performance across different age groups measured in F1-score. Each subfigure presents one multimodal approach along with its unimodal components.

stance, in accuracy, FL performs better than any other approach for female children ($p < 0.05$). Whereas for male children, PL approach performs significantly better than any other approach ($p < 0.01$).

An interesting finding appears in precision where it seems that all approaches perform better for females than for males (the difference is more apparent in FL) meaning that all approaches tend to make fewer false positive errors for female children and thus positive predictions can be trusted for female children compared to their male counterparts.

Recall shows lower performance in all approaches for both gender groups. Both FP and FL perform better for females compared to males whereas it is the opposite for PL. FPL achieves similar performances but significantly lower than the best approach. Subsequently, F1-score shows approximately the same pattern of variations as what was observed in accuracy in which



(a) Accuracy

(b) Precision

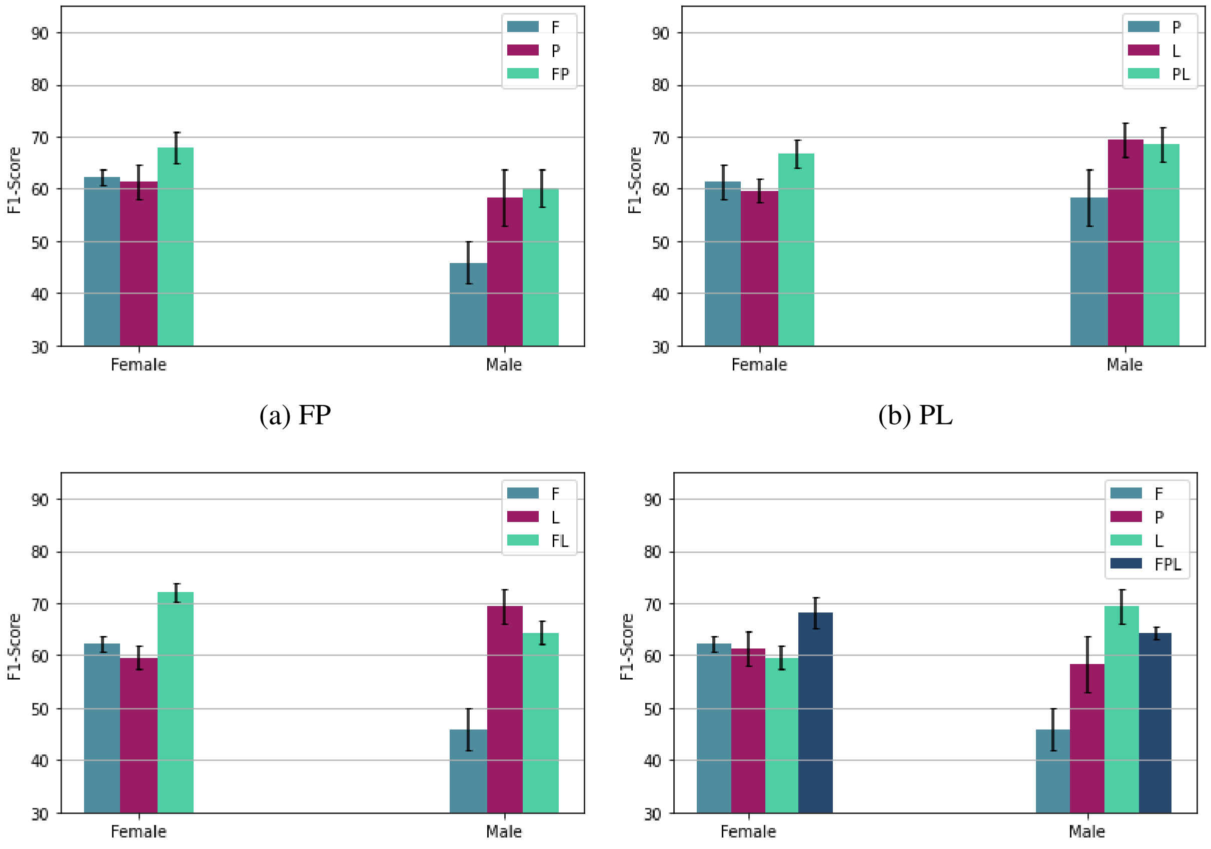
(c) Recall

(d) F1

Figure 5.9: **Gender-Based Performance in Multimodal Approaches.** The legend is shared across all subfigures.

FL works best for females whereas PL works best for male children. In general, it seems that (in all metrics) the FL approach works best for female children (or comparable to the best) which leads to an F1-score of 72.1%, while PL works best for males which leads to an F1-score of 68.5%.

Figure 5.10 presents the gender based performance of each multimodal combination along with their unimodal components (for brevity, this is only shown for F1-score). This is to get a clearer idea of whether gender variations are associated with the amount of complementary patterns detected by different unimodal components. The most interesting finding of the figure is that the combinations for female children always improve over their unimodal components ($p < 0.001$). Whereas for male children, the combination performs as well as the best unimodal component in FP and PL while it performs worse than the best component (language) in FL and FPL ($p < 0.001$). The results suggest that the approaches are capturing enough complementary patterns for female children while this is not the case for males. Moreover, in three approaches (all except PL), the overall performance is higher for female children than males ($p < 0.001$) which might resulted from the relatively poor performance of the face approach for male children.



(a) FP

(b) PL

(c) FL

(d) FPL

Figure 5.10: **Gender-Based Performance of Multimodal Approaches and Unimodal Components.** Performance across different gender groups measured in F1-score. Each subfigure presents one multimodal approach along with its unimodal components.

5.5 Discussion

The primary motivation for conducting the multimodal experiments is to exploit the complementary information that might have been captured with the unimodal approaches as these were found to perform differently over different stories and with different metrics. The results in Table 5.1 show that the multimodal combinations have led to significant improvements in most cases which confirms that different modalities are capturing different patterns regarding the attachment styles. Various conclusions were drawn from the results regarding what is the most beneficial modality and which attachment group is more identifiable by our approaches. These findings can help inform a decision on what type of approaches are more appropriate depending on the circumstances.

It is also shown that the combinations have led to improvements in some stories but not in others. In particular, one important finding is that the approaches achieve lower performance in recall, however, it is shown that BF and TA have led to better recall after the multimodal combinations. As such, it might be worthwhile investigating (from a psychological point of view) why these stories work better for positive cases which may lead to suggesting other scenarios that might induce similar types of distress to be able to identify the positive cases to a better extent (by aggregating multiple classifiers that achieve good recall). The results also show that the SM story has never improved over the components and in some cases it performed significantly worse than the best component. This can result when one of the modalities tends to assign higher posterior probabilities to wrong predictions (more confident at wrong predictions). Thus, it might be useful to eliminate such a story as its inclusion probably harms the overall performance.

The results also show considerable variations among age and gender groups. It might be useful to adopt different approaches for different groups, for instance, it might be reasonable to adopt the FP approach for older children or for female groups. However, these conclusions need further confirmations (due to possible biasing from having lower positive cases in older groups).

5.6 Chapter Summary

This chapter presented multimodal approaches for attachment recognition based on a late fusion scheme. The results show that combining the three approaches have always led to better performance meaning that each unimodal approach is extracting complementary cues that enhance the multimodal predictions. The best performance was achieved by combining paralanguage and language (PL) yielding an accuracy of 75.6% whereas the best F1-score was achieved by combining face and language (FL) yielding 68.8%. Also, this chapter analyses how stories perform in different approaches and several conclusions have been drawn from this analysis. The effect of age and gender is also explored where it has been shown that the multimodal combinations behave differently with respect to these two factors.

Chapter 6

Performance Analysis

6.1 Overview

This chapter presents an in-depth analysis to gain better insights into the performance of our recognition approaches. The analysis was conducted from two different points of view. First, an analysis study has been conducted to gain insights on how different aspects of dataset variations affect the performance which is described in section 6.2, and second, a study has been made to investigate the effect of incorporating a confidence measure to increase the applicability of our approaches and it is described in section 6.3. Both analysis studies were conducted for all unimodal and multimodal approaches that were experimented in this thesis.

6.2 Effects of Dataset Variations

A closer inspection of the dataset recordings revealed that there are significant variations among participants with respect to two major aspects: length of recording, and amount of speech. The expectation was that these aspects may have an effect on the performance and, for this reason, several analyses were conducted to gain better insights on how and to what degree this occurs. The analysis was conducted by examining the change in the performance after varying only the aspect under examination. The following sections present the results with respect to each aspect and draw some conclusions.

6.2.1 Length of Recordings

As shown in Figure 3.2, the length of the recordings in the dataset varies considerably among participants which intuitively leads to expect that this may have an influence on the performance. The expectation was that the longer the recorded materials are, the more likely for classification approaches to detect cues that are related to the attachment types, meaning that longer recordings are expected to be classified correctly. To verify this relationship, the participants were sorted in a descending order depending on the full recording length (where all five stories are combined), then they were segmented into four disjoint groups in which the first group (Group 1) contains participants with the longest recordings and the last group (Group 4) contains participants with the shortest ones. This results in a number of 26 participants in each group and the distribution of each attachment class per group is shown in Table 6.1. As the table shows, there are more secure participants in Group 1 than insecure ones, while in the remaining groups, the number from each attachment class is approximately equivalent.

	Group 1	Group 2	Group 3	Group 4
Secure	18	14	13	14
Insecure	8	12	13	12
Random Accuracy	52.6	50.5	50.0	50.5
Average Length	993.7 ± 329.7	616.1 ± 23.6	538.3 ± 22.8	414.8 ± 66.1

Table 6.1: **Participants Distribution with respect to the Variations of Recordings Lengths.** The table also provides the random guess accuracy and the average recording length measured in seconds for each group.

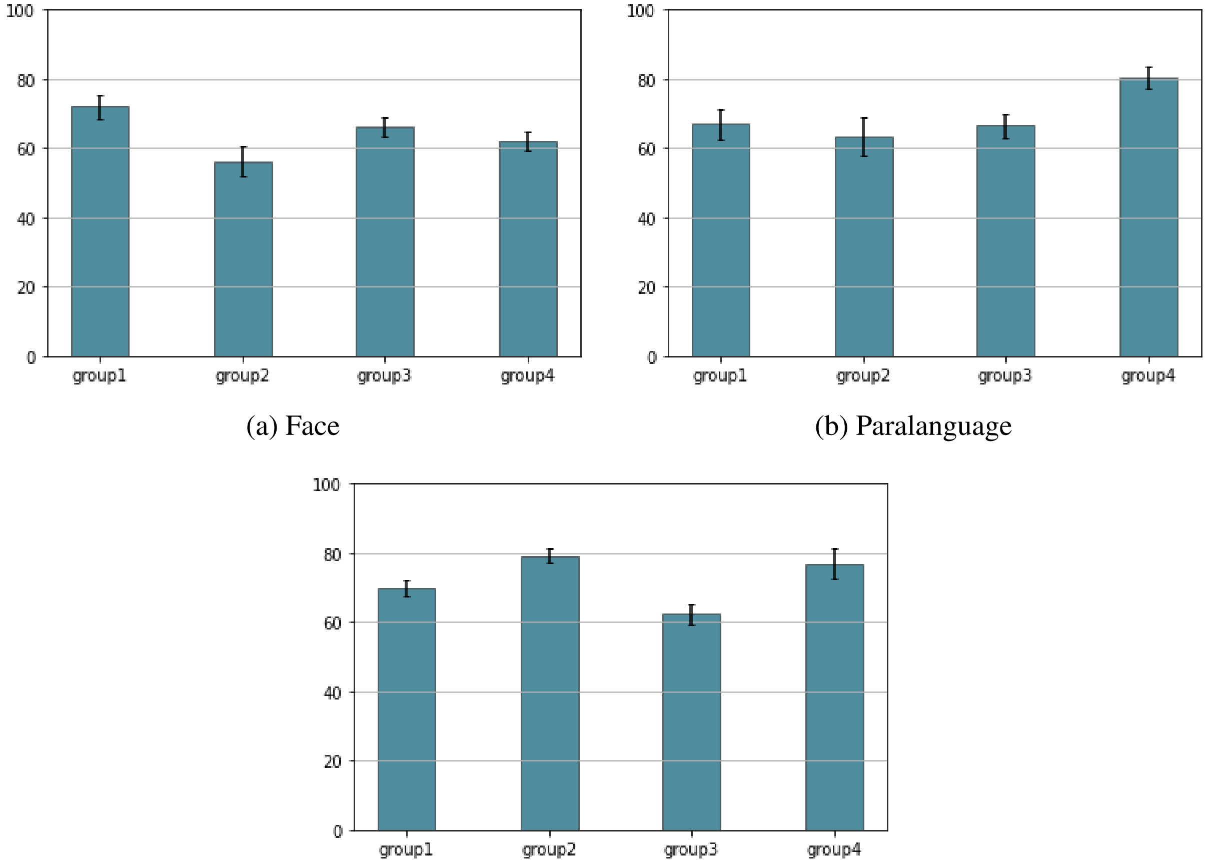
The performance with respect to each approach was obtained over each group separately in terms of accuracy and confusion matrices. Then, the performance was compared between groups to examine whether there are consistent patterns with respect to the length. The accuracy provides a general overview of how the performance changes while confusion matrices give a deeper look into the level of classes given that the dataset is imbalanced and that, as shown in previous chapters, the approaches perform differently for different classes. The following section provides the results for the unimodal approaches and the section after deals with the multimodal approaches.

Unimodal Approaches

Figure 6.1 and Table 6.2 present the performance in accuracy and the corresponding confusion matrices for each group in each unimodal approach. As shown in the figure, for the face approach, the performance is better for Group 1 (longest recordings) compared to the remaining groups whereas the worst performing group is Group 2 which hardly recognises more than half of the cases. The remaining two groups (Group 3 and Group 4) exhibit gradually decreased performance compared to the first group meaning that apart from Group 2, the performance gradually decreased from Group 1 (longest recordings) to Group 4 (shortest recordings) which suggests that face approach seems to achieve better performance over the longer recordings. The corresponding confusion matrices in Table 6.2 show that in all groups, the face approach approximately achieves similar performance in identifying the positive cases (I, insecure) as it approximately recognises half of the cases in each group. For the negative class (S, secure), the approach identifies most cases in all groups except in Group 2 where more negative cases are confused. This suggests that in face approach, the variations in recording length do not have an effect on the ability to identify the positive cases whereas there is a slight effect on identifying the negative cases as shown in Group 2.

An interesting pattern appears in the paralanguage approach as shown in Figure 6.1 as the overall accuracy is approximately similar for the first three groups but it gets significantly better for Group 4 (shortest recordings). By inspecting the corresponding confusion matrices in Table 6.2, in Group 1, the approach assigns most cases to the negative class (S, secure) confusing most of the positive cases, whereas for the lower groups, the approach gets better at identifying the positive cases while maintaining good performance at identifying the negative cases (except for Group 2). In Group 4, the approach identifies both classes to better extents compared to any other group achieving the least amount of confusions, and therefore the approach achieves its best overall performance over the shortest recordings. This suggests that the approach is associating long recordings to the secure class as can be seen from Group 1. Thus, the length has an effect on the performance in a way that is against the initial expectation in which the approach is achieving the best performance on the shortest recording besides associating long recordings with the secure class.

A different pattern appears in the language approach in Figure 6.1 where the overall accuracy



(a) Face

(b) Paralanguage

(c) Language

Figure 6.1: The figures show the performance in accuracy depending on the **length of recordings in unimodal approaches**.

varies considerably between groups as the highest performances are achieved in Group 2 and Group 4 while the worst performance is in Group 3. The corresponding confusion matrices in Table 6.2 show that in Group 1, the approach assigns most cases to the negative class whereas for the shorter groups, the approach identifies more positive cases, however unlike the paralanguage, this approach confuses more negative cases in the lower groups. In Group 4, the approach identifies almost all positive cases while making the biggest confusion for the negative cases. This suggests that the approach tends to assign long recordings to the secure class and short recordings to the insecure class. Moreover, in the middle groups (Group 2 and Group 3), the approach achieves better performance in the longer group for both classes meaning that unless the recording is excessively long or short, the language approach achieves better performance over the longer recordings. The source of this effect may arise from the variations in the amount of words i.e., longer recordings are more likely to contain a higher amount of words.

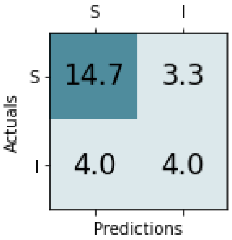
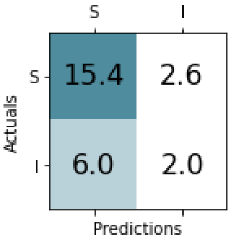
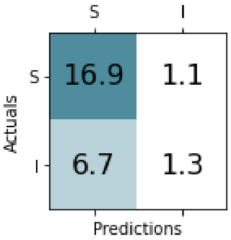
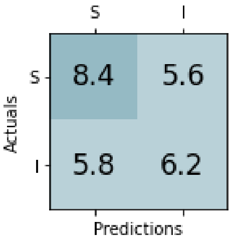
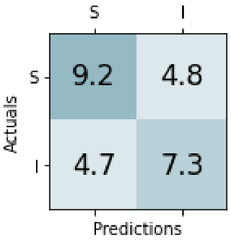
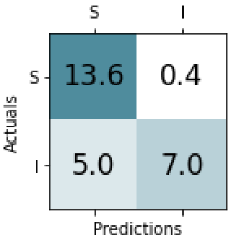
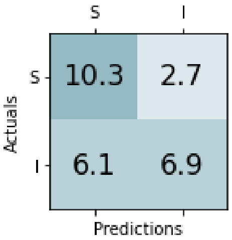
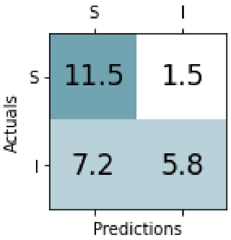
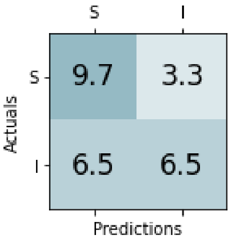
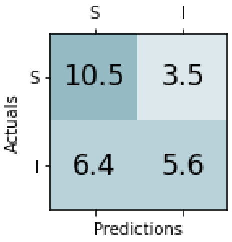
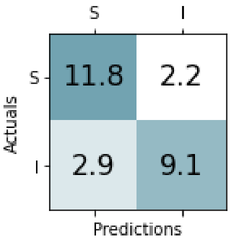
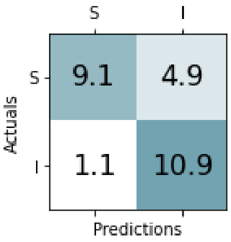
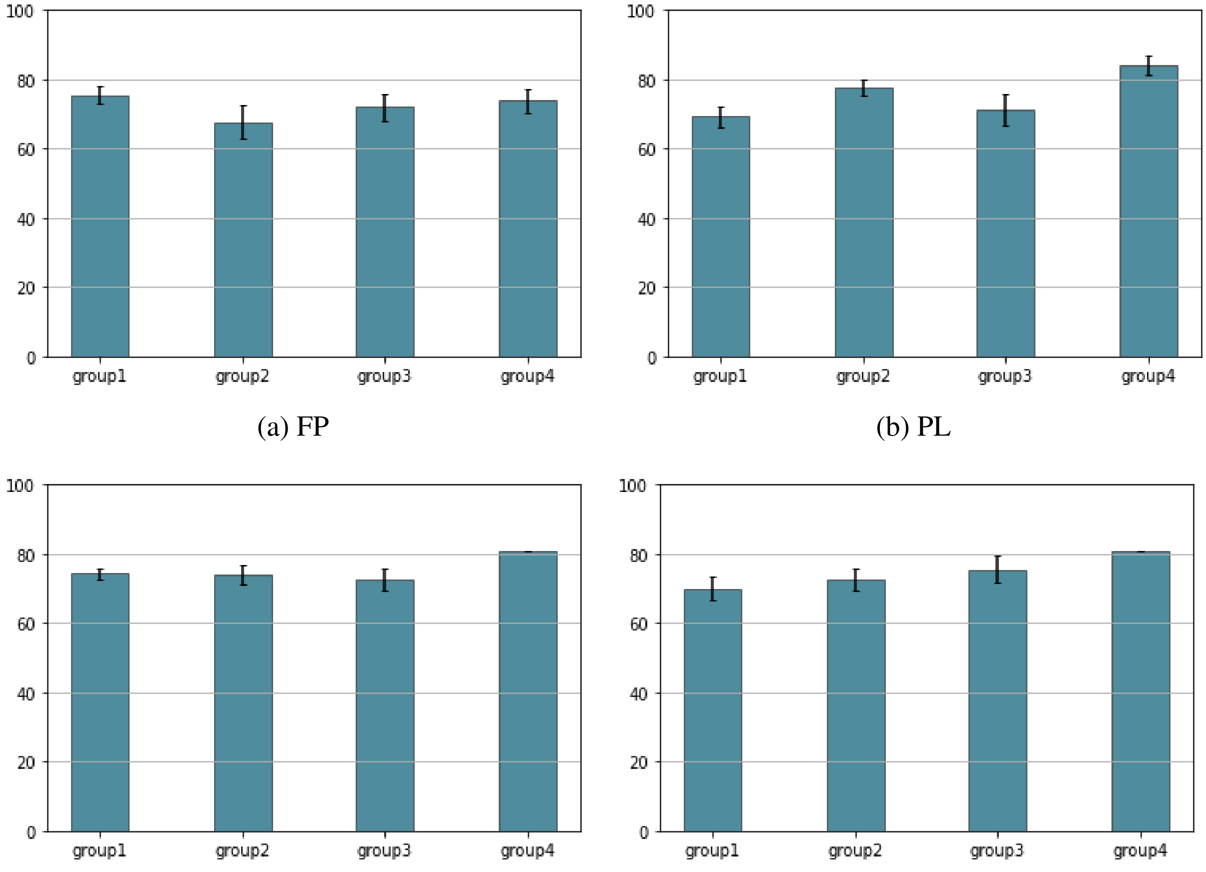
Gr.	Face	Paralanguage	Language
1			
2			
3			
4			

Table 6.2: This table shows the confusion matrices of the 4 groups that were segmented based on the **recordings length in the unimodal approaches**. The numbers denote the averages of 10 repetitions hence it appears in fractions. S: Secure, I: Insecure.

Overall, the results show that there is an effect of the variations of recording length on the performance of our unimodal approaches, especially in paralanguage and language approaches. In particular, both approaches tend to associate longer recordings to the secure class whereas language tends to associate shorter recordings to the insecure class. For face approach, it seems that the length has a slight effect on the ability to identify the secure cases.

Multimodal Approaches

The effect of the recording's length is also analysed for the multimodal approaches and the results are shown in terms of the overall accuracy (see Figure 6.2) and in terms of the corresponding confusion matrices in Table 6.3. Generally, there is more homogeneity in the performance across groups compared to what was observed in the unimodal approaches meaning that length has less effect on these approaches which could be due to the ability of different approaches to provide complementary information to compensate for this effect, for example, while the language approach achieves its worst performance on Group 3, the three multimodal approaches that involve the language component show a significant improvement for this group. Another major finding is that in three approaches (all except FP), the performance in Group 4 is better than other groups, whereas other groups achieve similar performances, especially in FL and FPL. Moreover, a clear pattern appears in FPL where the performance gradually gets better as one goes from Group 1 to Group 4 (i.e., as the recordings get shorter).



(a) FP

(b) PL

(c) FL

(d) FPL

Figure 6.2: The figures show the performance in accuracy depending on the **length of recordings in multimodal approaches**.

In the FP approach, the overall accuracies are approximately similar in all groups except in Group 2 which achieves slightly lower performance. The confusion matrices in Table 6.3 show that the approach is performing approximately similarly in identifying the negative cases (S,

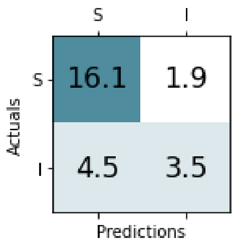
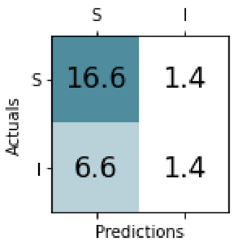
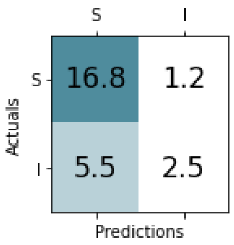
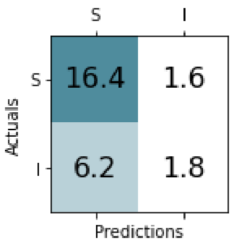
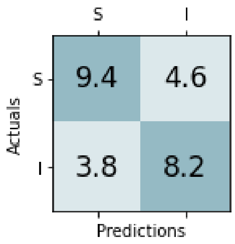
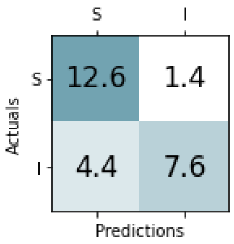
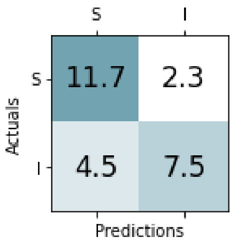
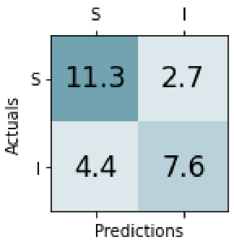
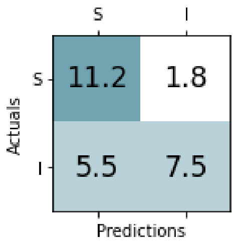
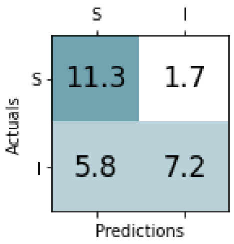
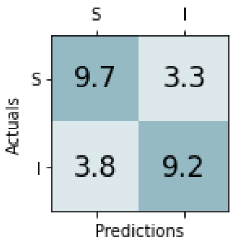
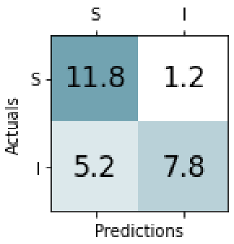
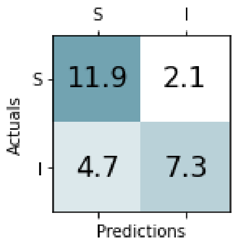
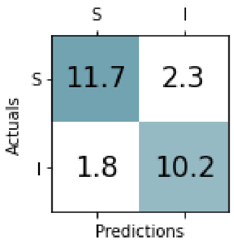
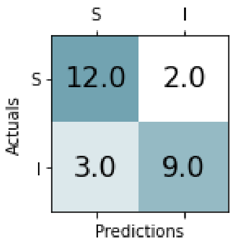
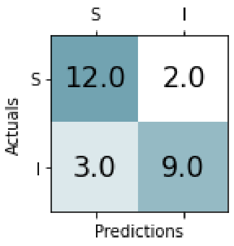
Gr.	FP	PL	FL	FPL
1				
2				
3				
4				

Table 6.3: This table shows the confusion matrices of the 4 groups that were segmented based on the **recording length in the multimodal approaches**. The numbers denote the averages of 10 repetitions hence it appears in fractions. S: Secure, I: Insecure.

secure) in all groups except in Group 2, wherein the approach confuses more negative cases. Regarding the positive case (I, insecure), the approach achieves its best performance in Group 2 and this performance is slightly degraded for lower groups (Group 3 and Group 4). For Group 1, the approach performs poorly in identifying the positive cases. This suggests that the approach can identify most of the negative cases of all groups whereas it is hard to identify the positive cases of Group 1 (long recordings).

In the PL approach, the overall performance in Figure 6.2 is higher for Group 2 and Group 4. By inspecting the confusion matrices in Table 6.3, the approach is able to identify the negative cases similarly well in all groups. However, for positive cases, the approach misclassifies almost all cases over Group 1 and it gets better at identifying these cases for lower groups. Accordingly, in Group 4, the approach achieves high performance in identifying both classes. The result suggests that the length does not have an effect on the ability to recognise negative cases but it has a considerable effect on identifying the positive cases as it seems that the approach is associating the long recordings to the negative class which is not surprising giving that both components (paralanguage and language) are showing this similar effect.

In FL and FPL approaches, as shown in figure 6.2, the overall performance is approximately similar for the first three groups whereas it gets better at Group 4. The confusion matrices show a somewhat similar pattern to the previous approach (PL), as the approaches identified the negative cases in all groups with a high performance (middle groups show a slight decrease in performance compared to PL). For positive cases, the approach confuses most cases in Group 1 and gets better at identifying those cases in lower groups where the least amount of confusion is achieved in Group 4. This suggests that the approaches are also associating the long recordings with the negative class.

Overall, it seems that in all approaches, significant confusion appears in Group 1 of the longest recordings (except FP) resulting from assigning more positive cases to the negative class. This suggests that the effect of associating long recordings to secure class is still present in the multimodal approaches. Another finding is that PL approach has the least amount of confusion at Group 4 compared to other approaches, suggesting that the unimodal components have captured a significant amount of complementary information over the shortest recordings.

6.2.2 Amount of Speech

Another main observation about the dataset is that some participants were actively engaged in completing the stories and describing how the child/mother doll feels while others remained mostly silent and hardly spoke throughout the session. As in the previous section, this intuitively leads one to expect that the amount of speech in the recordings may have an effect on the performance i.e., the more the child speaks the more likely to be classified correctly because the approaches are more likely to capture cues related to the attachment types.

To analyse this effect, the amount of speech was first measured from each recording by utilising the *Pyannote* library [140], an open source tool for speaker diarization which detects the regions of speech activity in the audio signals, and by accumulating the length of all speech regions, the total amount of speech per a recording is obtained. The part of the recordings that corresponds to the story's induction phase was excluded because we are interested in the amount of speech uttered by participants during the session, however, this should not make any difference as this part is exactly similar for all participants. As in the previous section, the participants were sorted in a descending order with respect to the amount of speech and then segmented into 4 disjoint groups where Group 1 contains the participants with the highest amount of speech and Group 4 contains the participants with the least amount. Subsequently, the performance was evaluated for each group in terms of accuracy and confusion matrices.

The Pyannote library performs a speech activity detection task using pre-trained neural models which reach the state-of-the-art performance on several datasets. This library takes the audio files as inputs and returns a list of segments each specifying the beginning and the end timestamps of a detected speech region.

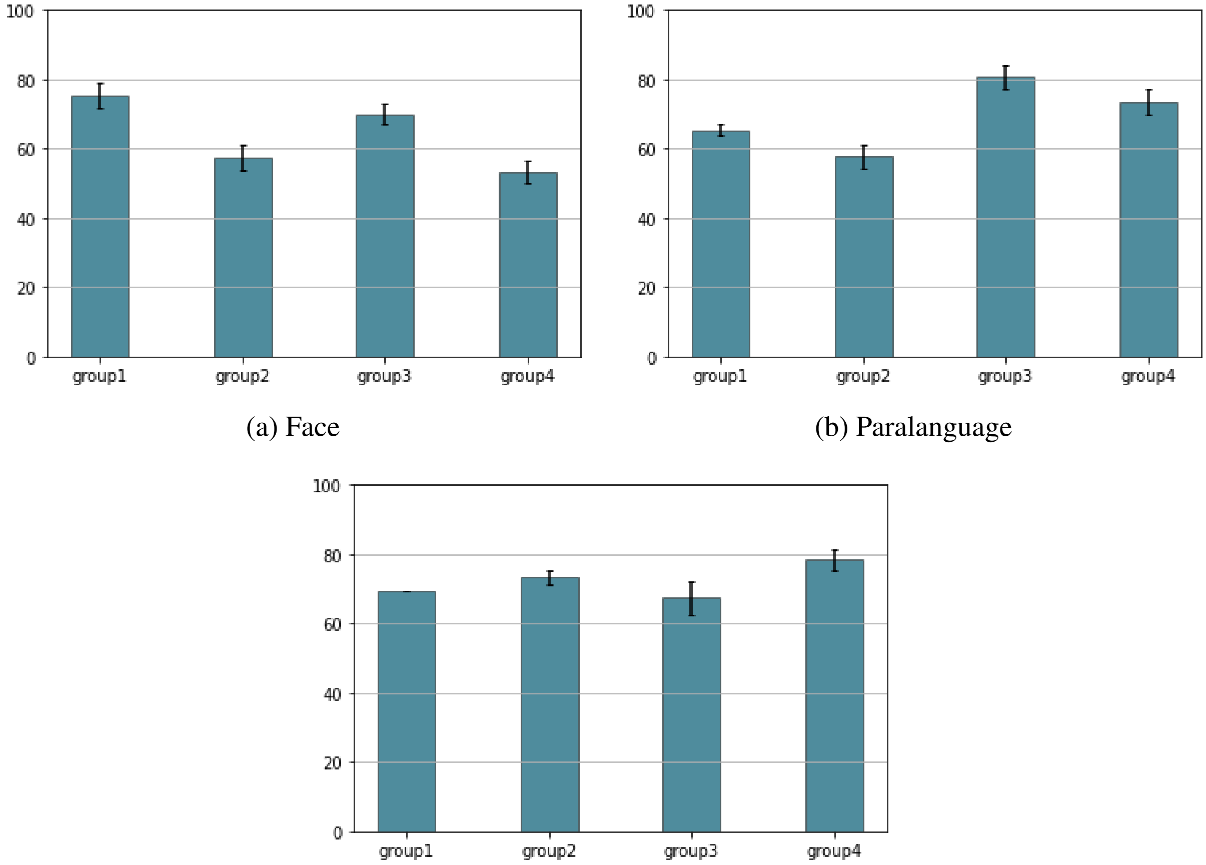
The distribution of the participants with respect to the total amount of speech is shown in Table 6.4. As the table shows, The first three groups have more secure children with Group 1 having the biggest difference between classes. In Group 4, there are more insecure children compared to other groups. Moreover, the table shows the average length of speech regions for each group.

	Group 1	Group 2	Group 3	Group 4
Secure	19	16	16	8
Insecure	7	10	10	18
Random Accuracy	53.1	51.6	51.6	47.4
Amount of Speech	410 ± 211.4	182.0 ± 17.9	124.6 ± 17.2	76.3 ± 17.5

Table 6.4: **Participants Distribution with respect to the Variations of Amount of Speech.** The table also provides the random guess accuracy and the average amount of speech measured in seconds for each group.

Unimodal Approaches

Figure 6.3 and Table 6.5 present the performance in terms of overall accuracy and confusion matrices in all unimodal approaches with respect to each one of the segmented groups. The overall accuracy shows that there are variations in the performance between different groups in all approaches and the pattern of variations is different among the approaches. In the face approach, the best accuracy was achieved in Group 1 and Group 3 and the lowest was achieved in Group 2 and Group 4. The confusion matrices show that the approach identified most negative cases in Group 1 whereas it confused a higher number of negative cases in the lower groups (relative to the overall number of negative cases in the respective group). For positive cases, the approach confuses almost half of the cases in each group except in Group 3, where it identifies the positive cases to a better extent. This result suggests that the approach is more likely to identify the secure children with a high amount of speech whereas it seems there is little effect on the ability to identify the positive cases as shown in Group 3.



(a) Face

(b) Paralanguage

(c) Language

Figure 6.3: The figures show the performance in accuracy depending on the **amount of speech in unimodal approaches**.

In the paralanguage approach, the accuracy is higher in Group 3 and Group 4 compared to the higher groups suggesting that the paralanguage approach is achieving better performance at the lower amount of speech. The corresponding confusion matrices show that the approach

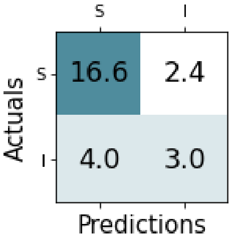
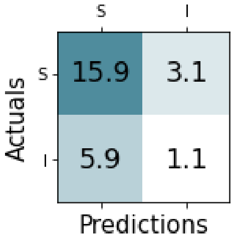
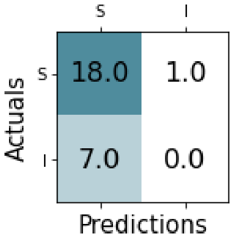
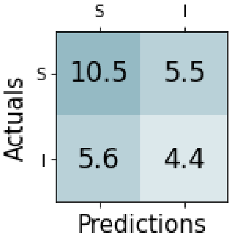
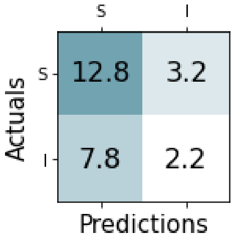
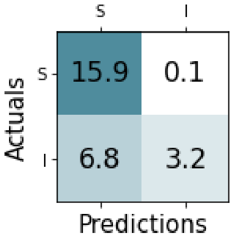
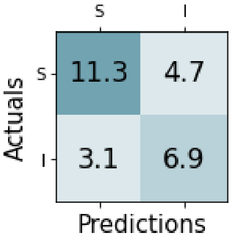
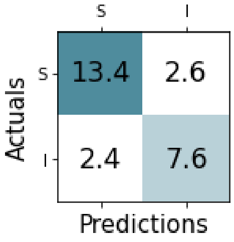
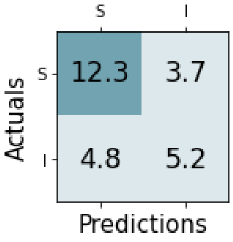
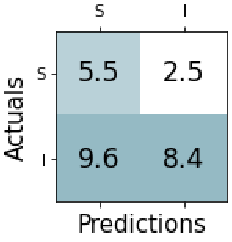
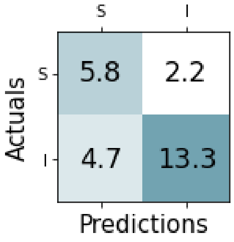
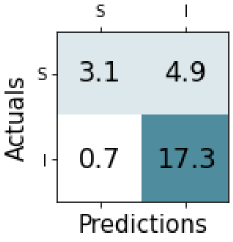
Gr.	Face	Paralanguage	Language
1			
2			
3			
4			

Table 6.5: This table shows the confusion matrices of the 4 groups that were segmented based on the **amount of speech in the unimodal approaches**. The numbers denote the averages of 10 repetitions hence it appears in fractions. S: Secure, I: Insecure.

assigns most cases in Group 1 and Group 2 to the negative class. For lower groups, the approach performs better at identifying the positive cases leading to the least amount of confusion in Group 3 compared to other groups. This result suggests that the approach associates high amount of speech to the secure class as shown in the higher two groups (only 19% is correctly classified in the positive cases whereas it is 82% for the negative cases). On the other hand, for the lower two groups, the approach performs better for both classes in which it correctly classifies 75% of the positive cases and 81% of the negatives.

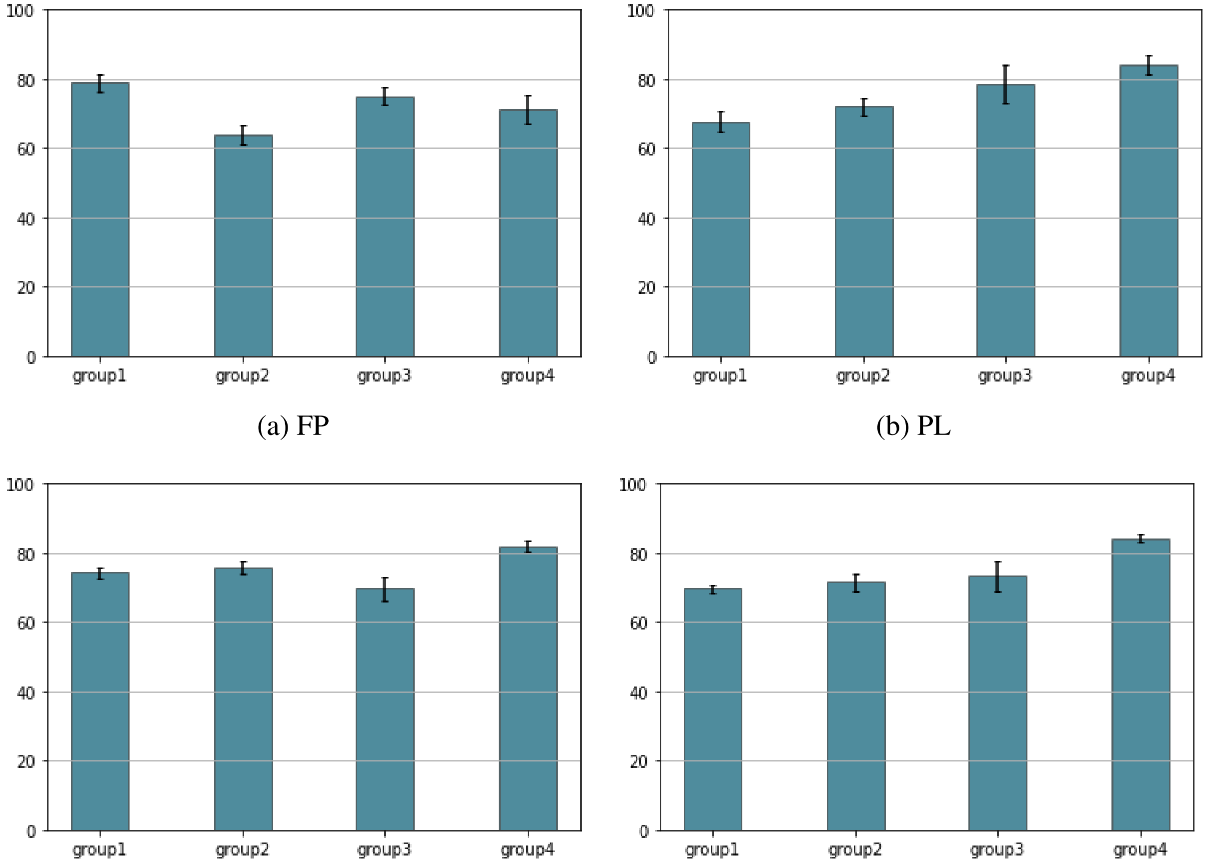
The language approach shows less variation in the overall performance among groups compared to other approaches suggesting that the variations in the amount of speech have less impact on this approach. The best accuracy was achieved in Group 4, while both Group 1 and Group 3 achieved lower accuracy. The confusion matrices show that the approach assigned almost all cases of Group 1 to the negative class whereas in Group 2 the approach gets slightly better at identifying the positive cases. In Group 3, the approach identified more than half of the positive cases while confusing more negative cases compared to higher groups. In the lowest group, Group 4, almost all positives are identified while more than half of the negatives are confused. The result suggests that the approach is associating the high amount of speech to the secure class whereas it is associating the low amount to the insecure class.

Overall these results confirm that the variation in the amount of speech has an effect on the performance, however, the nature of this effect is against what we expected it to be. In particular, both paralanguage and language associate the high amount of speech to the secure class. Moreover, paralanguage achieves higher performance for both classes at low amount of speech. In addition, language approach associates the low amount of speech to the insecure class. On the other hand, face approach is more likely to identify secure children with high amount of speech.

Multimodal approaches

This section presents the results of the effects of the amount of speech with respect to the various multimodal approaches. The expectation is that there is an effect of the amount of speech on various multimodal approaches as there was an effect on the unimodal components and another effect may arise as a result of the combination. Figure 6.4 and Table 6.6 show the performance in accuracy and confusion matrices for each multimodal approach. Overall, the variation seems less pronounced than what was observed in the unimodal approaches, and in three approaches, the best performance was achieved at the lowest amount of speech (Group 4).

In the case of the FP approach, the best performance was achieved in Group 1 whereas the worst performance was achieved in Group 2. The remaining groups show a gradual decrease compared to Group 1. The corresponding confusion matrices show that the approach identifies most negative cases at Group 1 whereas for lower groups the approach confuses more negative cases. For positive cases, the approach confuses more than half of the cases in Group 1 and gets better at identifying these cases in lower groups. This suggests that this approach also associates the high amount of speech to the negative class and it better identifies the positive cases at the



(a) FP

(b) PL

(c) FL

(d) FPL

Figure 6.4: The figures show the performance in accuracy depending on the **amount of speech in multimodal approaches**.

lower amount of speech.

For the remaining approaches, the overall accuracy is higher for Group 4, with PL showing a gradual increase in the performance from Group 1 to Group 4. In confusion matrices, these three approaches exhibit similar patterns. In particular, almost all cases of Group 1 were assigned to the negative class, and at lower groups, the approaches identified positive cases to better extents while confusing more negative cases. In Group 4, all approaches achieve the least amount of confusion suggesting that these approaches achieve their best performance at the lower amount of speech. The results suggest that these approaches too tend to associate a high amount of speech with negative class and achieve their best performances over the least amount of speech.

Overall, the results suggested that the multimodal approaches (except FP) achieve their best performance at low amounts of speech. On the other hand, all approaches perform poorly at identifying insecure cases at high amounts of speech. Another interesting finding is that the

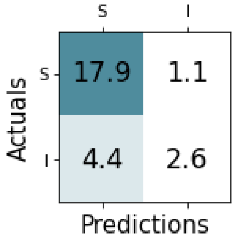
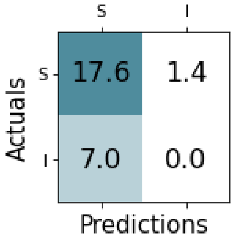
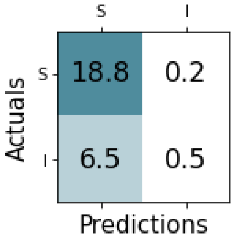
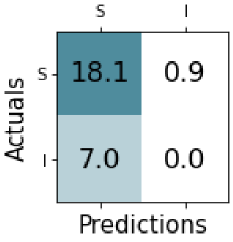
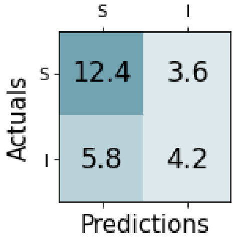
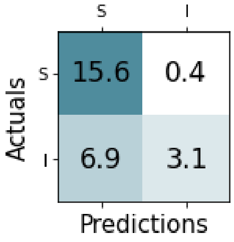
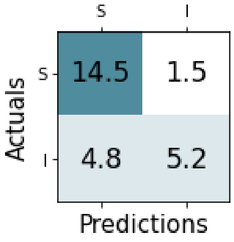
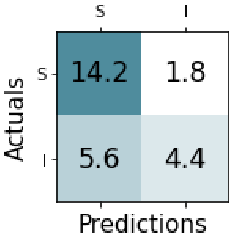
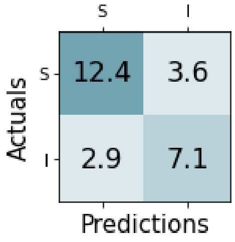
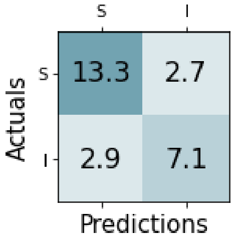
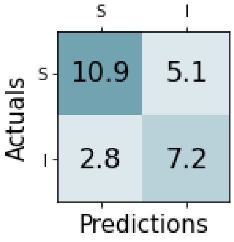
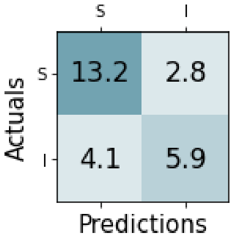
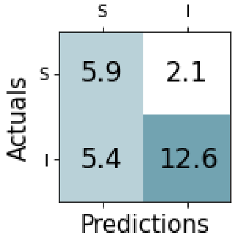
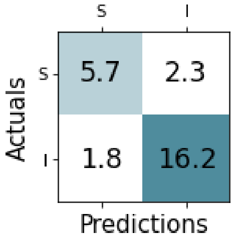
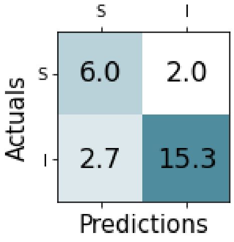
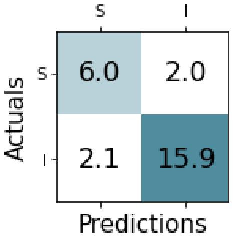
Gr.	FP	PL	FL	FPL
1	 17.9 1.1 4.4 2.6	 17.6 1.4 7.0 0.0	 18.8 0.2 6.5 0.5	 18.1 0.9 7.0 0.0
2	 12.4 3.6 5.8 4.2	 15.6 0.4 6.9 3.1	 14.5 1.5 4.8 5.2	 14.2 1.8 5.6 4.4
3	 12.4 3.6 2.9 7.1	 13.3 2.7 2.9 7.1	 10.9 5.1 2.8 7.2	 13.2 2.8 4.1 5.9
4	 5.9 2.1 5.4 12.6	 5.7 2.3 1.8 16.2	 6.0 2.0 2.7 15.3	 6.0 2.0 2.1 15.9

Table 6.6: This table shows the confusion matrices of the 4 groups that were segmented based on the **Amount of Speech in the multimodal approaches**. The numbers denote the averages of 10 repetitions hence it appears in fractions. S: Secure, I: Insecure.

effect of associating the low amount of speech to insecure class which appears in the language approach is mitigated in the multimodal approaches.

6.2.3 Discussion

The above analysis factors (recordings' length and amount of speech) can be considered as signs of the level of engagement during the assessment sessions. The results suggest that most of our approaches tend to associate higher engagement groups with the secure class whereas language approach associates the lowest engagement group with the insecure class. As mentioned above, the intuitive expectation was that a higher amount of these factors could yield better performance, however, the results suggested otherwise. While the results seem counterintuitive, it is in fact consistent with the MCAST [11] which, based on human ratings, concluded that the level of engagement is one of the strong predictors of the attachment types i.e., the higher level of engagement is associated with secure attachment. This means, although undesirable, that it is a reasonable effect that the approaches capture these factors as predictors of the attachment types. However, the absence of other cues that were easily recognised by human raters could be the reason for this undesirable effect. Therefore, investigating other behavioural cues that may serve as better signs of the level of engagement may attenuate this effect i.e., by providing better discrimination ability between attachment types at high levels of the above analysis factors. Other possible behavioural cues include: head pose, eye gaze, and hand movements as explored in [80].

6.3 Incorporating a Confidence Measure

This section investigates the feasibility of incorporating a confidence measure into the predictions of our recognition approaches. A confidence measure provides the ability to quantify the amount of certainty in the predictions. Incorporating such a measure can increase the usefulness and applicability of the approaches because this way one can set a criterion for accepting or rejecting the system decisions based on the level of its confidence. For example, in health diagnosis systems, health practitioners can utilise this information to aid their decision on whether they can rely on the system diagnosis or if a further manual assessment is needed. In order for such a measure to be useful, confidence estimates should correlate with the correctness of the predictions, i.e., the approaches should possess the property of assigning high confidence when making the correct predictions.

Confidence measures have been explored widely in the literature especially in the field of safety-critical systems when making wrong predictions can be risky, such as autonomous driving and health diagnosis systems. Ranges of confidence measures have been explored in the literature. The most commonly used confidence measure is based on the posterior probabilities that are attributed to input examples by several probabilistic classifiers. Other studies proposed other approaches to estimate the confidence, for instance in [141], the authors introduced the true class probability (TCP) in which the confidence score is learned via an auxiliary neural network model given the input data and the true labels from the primary classification network. For multi-class classification problems, a measure that takes into account the distance between the estimated probabilities of the predicted and the remaining classes is proposed in [142], meaning that if the distance is higher, the classifier should be more confident in its decision. Moreover, a trust score is derived in [143], where a set of labelled data is reduced to a high-density set per class. Then, for a testing example, the trust score is estimated by taking the ratio of the distance between the test example and nearest class[1] to the distance from the predicted class.

In this study, a confidence measure based on the entropy of the predictions is utilised. In particular, our recognition approaches estimate the posterior probabilities of a given test sample to belong to one of the classes and the entropy measures the amount of certainty on this prediction. For face approach, these probabilities are estimated by the logistic regression model. In paralanguage approach, the posterior probability is estimated as the fraction of recording segments assigned to the predicted class. In language approach, the posterior is estimated by the softmax function as a final layer of the CNN model. The multimodal approaches estimated the posterior by applying the SR method on the probabilities that were estimated by various unimodal components.

Having a certainty value associated with each prediction, a threshold value θ can be estimated beforehand to mark a certainty level over the test set, above this θ, predictions are guaranteed to reach a performance level K that is considered to be sufficiently acceptable depending on the problem. In other words, predictions that are made with a certainty value that is higher than θ can be accepted whereas predictions with lower values should be rejected. Therefore, by choosing a suitable K, this θ value can be used as a decision criterion during inference to only accept predictions that are made with a level of confidence that is higher than θ. Ideally, K should be chosen so that the amount of accepted wrong predictions is minimal while the amount of rejected predictions is reasonable to be handled manually.

[1] Which is different than the predicted class.

6.3.1 Confidence with Entropy

In information theory, Shannon entropy or entropy $H(X)$ is a measure of the amount of information contained in a random variable X. The amount of information corresponds to the amount of uncertainty in the possible outcomes of X, meaning that, if an outcome is almost certain to occur among other possible outcomes, then the amount of information that is conveyed by X is minimal whereas if the outcomes are uncertain or equally likely to occur, then the amount of information that is conveyed by this variable will be maximal i.e., observing the value that X takes is uninformative if its occurrence is certain. Initially, entropy was developed in communication theory to measure the minimum number of bits to encode a message generated by a message source and it can be computed from equation 6.1 where p_i denotes the probability of the occurrence of i_{th} outcome and n is the number of possible outcomes.

$$H = \frac{-\sum_{i=1}^{n} p_i \log_2 p_i}{\log_2(n)} \tag{6.1}$$

For binary classification problems (such as the one studied in this research), probabilistic classification models estimate the probability of an input sample belonging to one of the classes as $p(x)$ and thus the probability of belonging to the other class will be $1-p(x)$. In this case, entropy can also be used to measure the amount of uncertainty a classification model has in its predictions where the above equation can be rewritten as in equation 6.2. If an input sample is assigned to a class C_1 with high probability $p(x)$, and therefore $1-p(x)$ will be low, the quantity H will be low, and thus $1-H$ will be high reflecting the high certainty of the model in this prediction. This relationship between $p(x)$ (probability of assigning x to C_1) and certainty $(1-H)$ is plotted in Figure 6.5.

$$H = -(p(x)\log_2 p(x)) + (1-p(x)\log_2 1-p(x)) \tag{6.2}$$

6.3.2 Approach

To examine whether the use of this confidence measure can increase the usefulness of the recognition approaches, the certainty level $(1-H)$ was calculated at each prediction, by taking the posterior probabilities that the approaches assigned to each attachment class using equation 6.2.

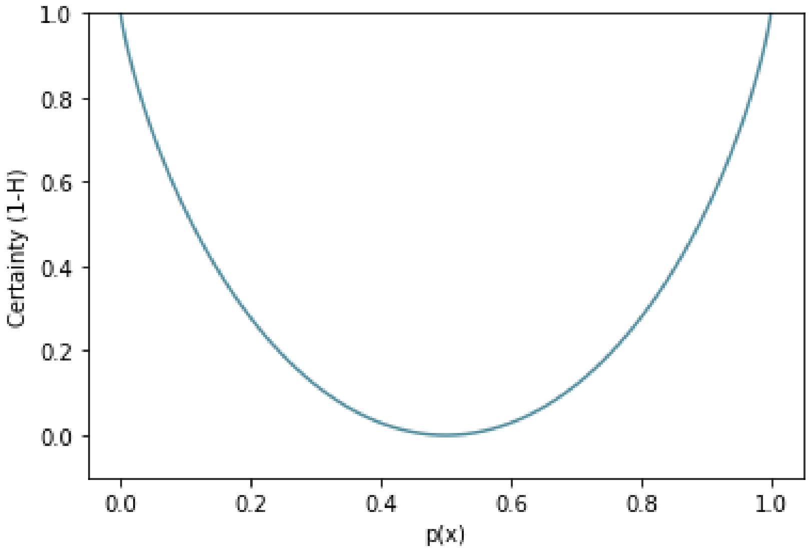


Figure 6.5: **Certainty** $(1-H)$ **as a Function of** $p(x)$**.** The plot shows the relationship between $p(x)$ and certainty in the case of binary classification problems. When $p(x)$ is near zero or one, the certainty will be near its maximum value because the model is certain about assigning x to either class. Whereas, when $p(x)$ is around 0.5, the certainty level will be at its lowest value because the model is assigning almost equivalent probabilities to each one of the classes.

Our approaches involve aggregating multiple predictions (5 predictions for the Unimodal approaches and 10 to 15 for the Multimodal approaches) using SR method to make the participant-level prediction, therefore this can results in values that are greater than 1. In order for the aggregated values to be re-expressed as a valid probability value, the ratio of the aggregated posterior probabilities of one class to the sum of the aggregated posterior probabilities of both classes is calculated. More formally, the probability of assigning the input sample x to the negative class, $p(y_0|x)$ can be computed from:

$$p(y_0|x) = \frac{\sum_{i=1}^{R} p(y_0|x_i)}{\sum_{j=0}^{1}\sum_{i=1}^{R} p(y_j|x_i)} \tag{6.3}$$

where $p(y_j|x_i)$ is the i^{th} story-level posterior probability corresponding to class j and R is the number of predictions per approach.

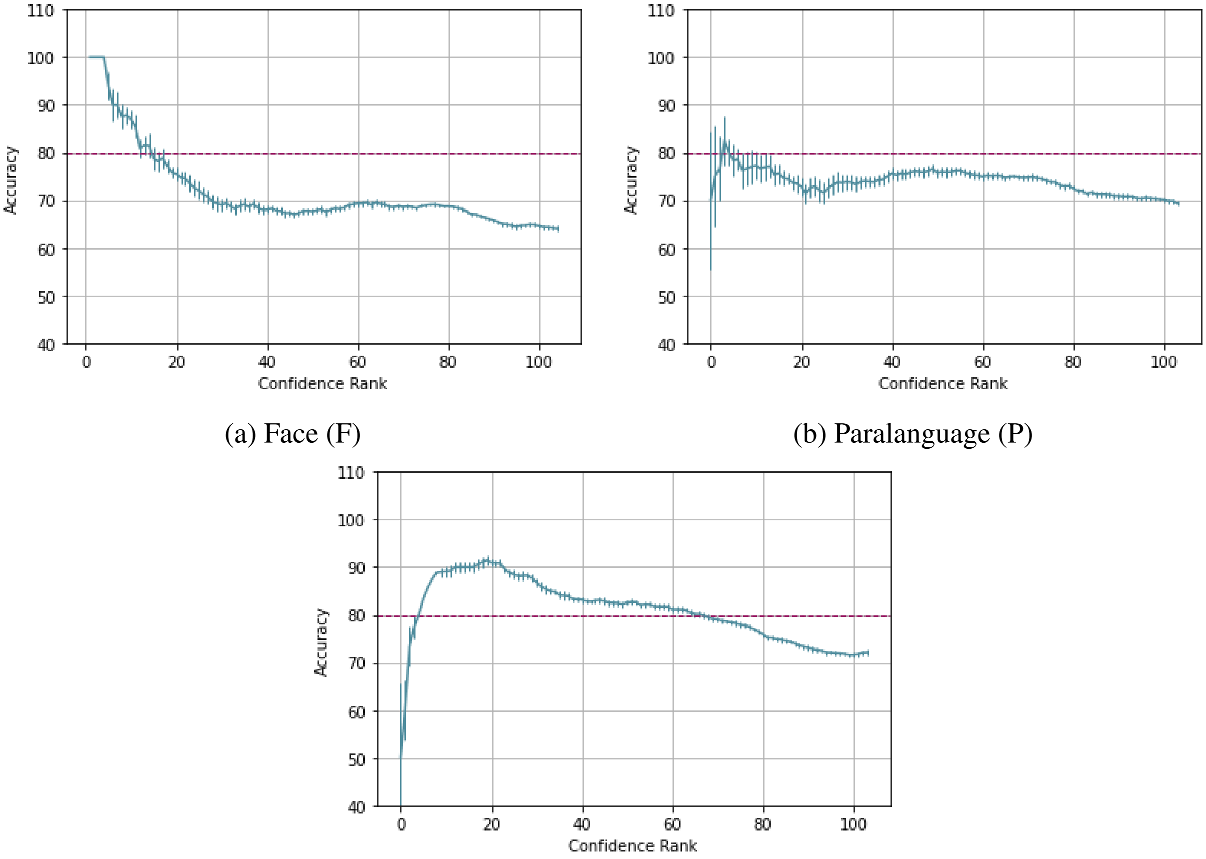
Next, each participant was assigned a confidence rank by sorting the participants in a descending order depending on the calculated $(1-H)$ i.e., the participant with the highest certainty will be in the first rank, and the participant with the least certainty will be in the last rank. Subsequently, the performance in accuracy is evaluated as a function of the certainty rank, meaning that the accuracy at rank 10 is evaluated over the top 10 ranked participants. As mentioned above, for this confidence measure to be useful, the approaches should have the tendency to correctly classify the high-ranked participants. This allows one to decide on a θ value above which the approaches can have an acceptable performance and the decisions of the system can be trusted.

6.3.3 Results

Figure 6.6 presents the performance measured in accuracy and plotted as a function of the confidence rank in all of our unimodal approaches. As shown in the figures, the downward curve that especially appears in the Face and Language approaches suggests that the approaches generally yield better performance for the highest-ranked participants, meaning that these approaches generally tend to be more confident when making the right decision. In paralanguage, however, this tendency is much less apparent, as the performance of the highest-ranked participants is slightly higher than the overall performance (69.3% for paralanguage, as shown by the last point of the curve) meaning that the confidence measure that was estimated in paralanguage approach is not correlated with the correctness of the predictions and this possibly results from the way that the posterior probabilities are estimated in this approach (the fraction of the recording segments that are assigned to the predicted class).

Another major finding is that both paralanguage and language show a very low performance at the highest-ranked examples while the face approach shows 100% and 90% accuracies for the highest 4 and 12 participants respectively. This implies that the face approach, despite achieving the lowest overall performance, has the highest tendency to be more confident when making the right decision. On the other hand, the language approach achieves an accuracy above 90% for the top 20 participants, however, it also misclassifies one participant with very high confidence (the highest ranked participant in most repetitions of language approach experiments) which suggests that the confidence measure that is estimated in the language approach might not be correlated with the correctness at least for some special cases (this participant belongs to the insecure class and corresponds to the highest amount of speech in the dataset).

The main goal of incorporating a confidence measure is to decide on a confidence level such that predictions above this level are accepted otherwise the predictions will be rejected or reported to be handled manually. For instance, by setting the K level to be equal to 80% (shown by the



(a) Face (F)

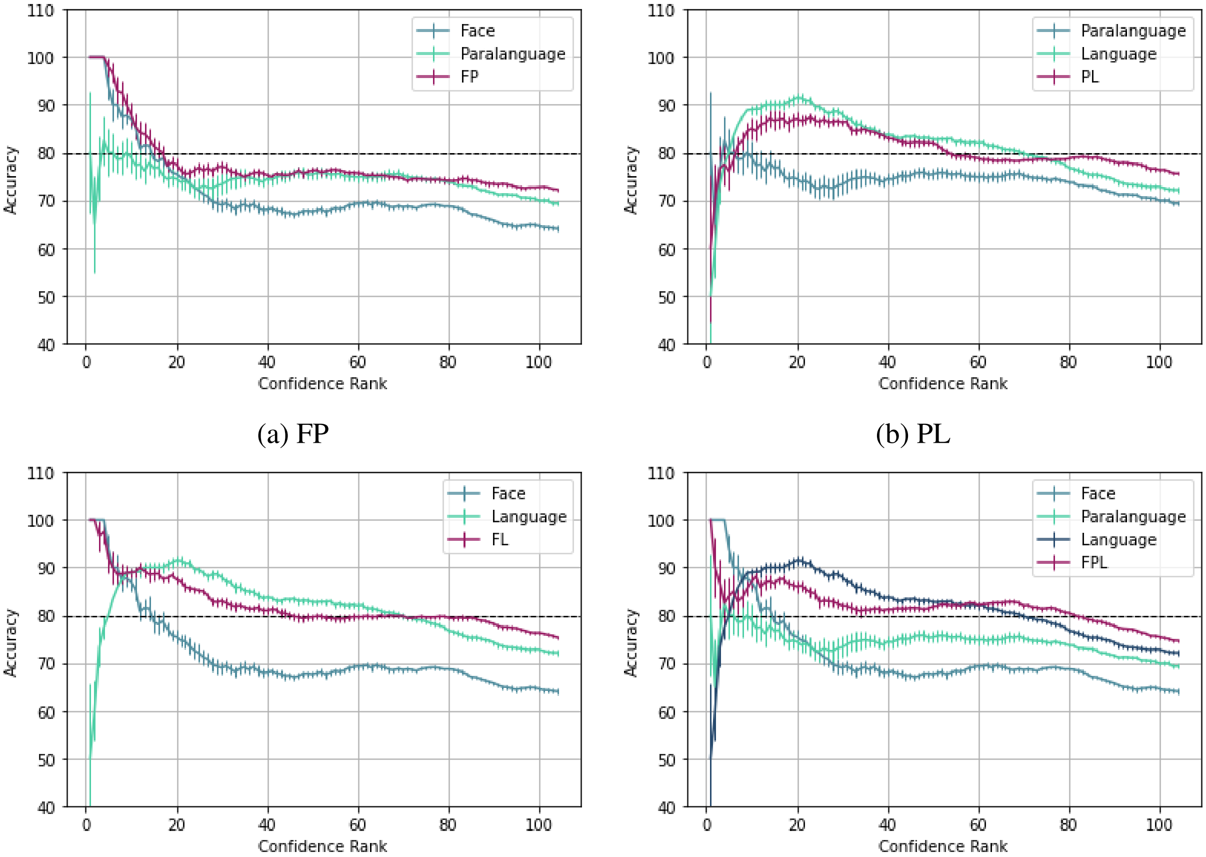
(b) Paralanguage (P)

(c) Language (L)

Figure 6.6: **Confidence Measure in Unimodal Approaches.** The figures plot the performance in accuracy as a function of confidence rank. The performance is presented in terms of averages and standard errors over 10 repetitions. The dashed line marks the 80% accuracy. Each subfigure presents the performance of one individual unimodal approach.

magenta dashed line in the figures), only the top 12 ranked participants in the face approach can be accepted while more than 90 participants will be rejected, this means that 85% of the decisions will be discarded as being uncertain. As discussed above, the paralanguage approach does not seem to correlate confidence with correctness as indicated by the flatness of the curve and the highest standard errors for the highest-ranked participants therefore, incorporating a confidence measure does not seem to enhance the practicality of the approach. Moreover, the performance hardly exceeds our *K* level and in this case, no predictions will be accepted. For the language approach, the top 64% ranked participants are classified with an accuracy of 80%, this is useful because if we consider this K level, less than half of the predictions will be reported for manual assessment i.e., the assessment workload can be reduced significantly.

The same procedure was repeated over the four multimodal approaches and the results are presented in Figure 6.7. Each approach is shown along with the curves that correspond to the unimodal components. For visibility, The curves that correspond to the multimodal approaches are shown in magenta in each subgraph. In general, as in the case of unimodal approaches, the curves are higher at high-ranked participants compared to lower ranks meaning that approaches are more confident when making the right decision which is not surprising given that two unimodal components possess this property to varying degrees.



(a) FP

(b) PL

(c) FL

(d) FPL

Figure 6.7: **Confidence Measure in Multimodal Approaches.** The figures show the performance in accuracy as a function of the confidence rank over the multimodal approaches. Each subgraph presents one multimodal approach along with its unimodal components. The performance is presented in terms of averages and standard errors over 10 repetitions. The black dashed line marks the 80% accuracy.

One interesting finding that emerges from the figures is how different unimodal components contribute to enhance the usefulness of incorporating the confidence measure. This is especially noted in the FL and FPL approaches (two bottom graphs). For example, the FL and FPL approaches mitigated the issue that appears in the language approach which always misclassifies a participant with very high confidence (the two approaches assign lower confidence when predicting this participant).

By considering our K performance level, it is shown that there are no major improvements yielded in the three first approaches FP, PL, and FL as they do not enhance the performances of their best component, in particular, FP only accepts around 16% of predictions whereas it is around 50% in PL approach and 64% for FL. On the other hand, FPL has improved remarkably over the components i.e., a higher number of predictions can be accepted in this approach. In particular, more than 77% of the predictions can be accepted for the FPL approach given our K acceptance level.

6.3.4 Discussion

The results show that the incorporation of a simple confidence measure as the one proposed above, has a significant enhancement on the usefulness of our approaches. The best overall performance of the approaches over the full dataset was 75% achieved by PL and FL approaches. Our confidence measure gives us a better idea of the predictions that are more likely to be misclassified and thus may require a further manual assessment. It is shown above that by considering a θ level that corresponds to the performance level of 80% accuracy, major enhancement can be achieved, in particular, the FPL approach can accept 77% of the predictions and the remaining 23% will be referred for manual assessment.

Ideally, the K performance level should be set to 100% accuracy, thus, in inference, one can be almost certain that a recognition approach only accepts the correct predictions. On the other hand, our K level (80%) has a major significance in this recognition problem. According to [12], human assessors of the SSP task are required to follow a rigorous training process until an agreement of 80% performance is reached. This suggests that if an automatic assessment approach can perform at an accuracy of 80%, it can be regarded to reach human-level accuracy and therefore one can conclude that FPL performs at human-level accuracy on 77% of the cases. In other words, the approach will reject less than a quarter of the predictions which reduces the workload of health practitioners considerably i.e. the manual assessment is now only required for less than a quarter of the cases.

The equivalent rating accuracy K is not known for the MCAST tool, possibly because it has not yet been clinically applicable, and therefore our K level may be underestimated i.e., human-level performance could be higher than K for MCAST. However, there is still vast room for improvement in the recognition process itself and in estimating a confidence measure that better correlates with the correctness which may allow to increase this K level, i.e., one can be more certain that the approach accepts only the correct predictions.

6.4 Chapter Summary

This chapter analysed the performance of our recognition approaches from two broad viewpoints. First, it investigates the effects of two aspects of variations found in the dataset: recordings' length and amount of speech. The results suggest that most of our approaches tend to associate a high amount of these aspects to the secure attachment class meaning that a considerable amount of misclassifications comes from failing to detect insecure children of the high amount of these aspects. The second analysis viewpoint addresses the usefulness of incorporating a confidence measure to allow for setting a criterion to accept or reject predictions. The results show that an approach that is based on fusing all three behavioural modalities is capable of performing comparably to trained human assessors in 77% of the predictions. This suggests that the workload of the assessment process can be reduced significantly.

Chapter 7

Conclusion

7.1 Overview

This thesis aimed at detecting the attachment styles automatically in school age children to limit time and costs needed to identify insecure children. The thesis attempts to detect the two main categories of attachment styles (secure and insecure), by analysing the three main behavioural channels: facial expressions, paralanguage, and language. Moreover, an attempt was made to fuse these modalities using a late fusion scheme in order to exploit the complementarity information that might be learned by different modalities. Overall, the results show that it is possible to identify the main categories of attachment styles in children with an accuracy of 75% by fusing either face and language modalities, or paralanguage and language. This performance corresponds to an F1-score of 68.8% which is achieved by fusing face and language modalities.

The effect of gender and age variation on the performance was explored and valuable conclusions were drawn regarding these two factors. Additionally, two major aspects of the dataset were analysed to explore whether there is an association between these aspects and the outcomes of the approaches. These aspects are, recordings length and the amount of speech. Finally, incorporating a confidence measure was suggested for the various approaches in an attempt to increase the usefulness of the work. The main outcomes of this research are summarised in section 7.3.

7.2 Thesis Statement Revisited

The thesis statement asserts that the differences in attachment styles in school age children leave detectable traces in their behaviours which then can be detected automatically by means of pattern recognition methodologies. Our results show that, using the three main behavioural channels, it is possible to capture attachment differences in children. In particular, it is shown that recognition approaches based on facial expressions, paralanguage and language can detect attachment with performance that is significantly higher than chance i.e., the approaches were able to capture relevant patterns.

7.3 Main Outcomes

The following points summarise the main outcomes of this PhD research:

- An attempt was made to detect the two main attachment styles by analysing the three main behavioural channels. Promising performance was achieved in all explored channels, meaning that attachment styles leave machine-detectable traces. The best result was achieved by means of the language channel with an accuracy of 72.1% and F1-score of 63.9%.

- Multimodal approaches were developed by fusing the above unimodal approaches using a late fusion scheme which leads to significant improvements in the recognition performances. The best result was achieved with accuracy of 75.6% and F1-score of 68.8% by fusing paralanguage and language (PL), and face and language (FL). These experiments confirm that all unimodal approaches are capturing complementary information.

- The age based performance analysis revealed that there are consistent differences between younger children (of age 5-7 years) and older children (of age 7-9 years). In particular, the results suggest that older children are more likely to express their attachment conditions through facial expressions while younger children are likely to express it better through language and paralanguage.

- Similarly, the gender based performance analysis revealed that there are differences among gender groups. Two important differences were noted: firstly, the results suggest that female children are more likely to express their attachment condition through facial expressions while male tend to express it more through language. Secondly: multimodal approaches outperform unimodal components for female children while this is not the case for male children, suggesting that unimodal approaches are capturing complementary information only for female children.

- Analysis of the effects of variation on the dataset with respect to two aspects: recording length and amount of speech. This analysis reveals important findings, in particular, the results show that most of the approaches tend to associate the higher amounts of these factors to the secure class whereas some approaches tend to associate lower amounts to insecure class. In other words, it seems that the approaches are capturing these variations as predictors of attachment styles.

- Incorporating a confidence measure has increased the applicability of the approaches. In particular, the results show that the confidence measure tends to be higher for correct classifications. This adds a valuable enhancement to the performance as it allows one to set a criterion to accept predictions only when the approach is confident. Our results show that an approach based on fusing all modalities can accept 77% of the predictions while performing comparably to a trained human assessor.

7.4 Limitations and Challenges

Several challenges have been encountered while conducting this research, overcoming these challenges may help improve the recognition performance which can be summarised in the following points:

- An attempt to learn a joint representation of the paralanguage and language has lead to a major difficulty. In paralanguage approach we had to process the recordings in segments of around 4.2 seconds long to handle the vanishing and exploding gradients problem. As such, it was essential to align these segments with the corresponding transcripts, unfortunately, it was hard to obtain these segments for the transcripts as the transcription tool provides timestamps only when the speaker changes or when the speech were interrupted by long pauses. This made the task harder to implement for all modalities although it might be straightforward to implement for face and paralanguage approaches[1]. Learning a joint representation may help exploit the mutual information to a better extent compared to the late fusion scheme.

- One issue about the facial expressions features AUs, is that *openface* was trained on a corpus where the minimum age of the participants was 18 years. Therefore, it is unknown how well is this tool at extracting facial AUs from children faces. This could be one of the main reasons behind the poor performance of the facial expressions approach and therefore it is worth investigating other possible ways or tools to extract AUs of children's faces.

7.5 Future Work

Possible directions for future work are:

- Learning joint representations of the various modalities may outperform the late fusion schemes. Given the above limitation it might be worth exploring methods to handle the vanishing gradients problem while allowing for learning a fixed length representation of the full recording i.e., without resorting to the segmentation process as adopted in this research. This can be further augmented with an attention mechanism to enhance the ability of capturing complementary information.

- It is worth exploring whether extracting deep features on facial expression approach can outperform the hand crafted feature that were extracted by *openface*, it may needs to be coupled with a mechanism to deal with the limited number of training data such as transfer learning that allows to train the deep model first on a richer training dataset.

- The results suggest that there might be other behavioural cues which can help discriminate between attachment types to a better extent. In particular, the results of section 6.2 suggests that it is worth exploring whether other behavioural cues that can be considered as signs of the level of engagement can improve the recognition performance. Some examples of such cues are eye gaze, head pose, and hand movements.

[1] This is because we have set the frame size of extracting the speech features to be equivalent to one video frame.

7.6 Closing Remarks

The goal of this thesis was to automate the attachment assessment process and to allow for large-scale population screenings to aid health practitioners such that they can direct their efforts to areas where human intelligence is unequalled, such as the treatment process. The results showed that our recognition approaches were able to correctly predict more than 77% of the cases with a performance comparable to a trained human assessor, providing a tremendous enhancement over the existing manual tools. Given that this work is one of the first attempts to address this problem, these results are promising, and potential improvements have been identified which may lead to better performance.

Appendix A

Normality Tests

Normality tests were conducted to ensure that the distribution of a given sample does not deviate significantly from the normal distribution, thus ensuring that the normality assumption of the t-test is not violated. In this respect, the Shapiro-Wilk statistical test was performed in which the null hypothesis states that the sample does not deviate significantly from the normal distribution and therefore, a p-value less than the significance level α indicates that the null hypothesis should be rejected.

Table A.1 shows the p-values of all approaches for both story-level and aggregation based on the accuracy results. Given that there are multiple hypothesis tests, the significance level is adjusted to account for type 1 errors according to Benferroni correction, thus $\alpha = 0.05/45 \approx 0.001$. As shown in the table, the null hypothesis cannot be rejected in any case.

	Face	Paralanguage	Language	FP	PL	FL	FPL
Breakfast (BF)	0.12	0.18	0.32	0.33	0.39	0.18	0.11
Nightmare (NM)	0.27	0.51	0.26	0.94	0.17	0.18	0.37
Tummyache (TA)	0.11	0.91	0.81	0.70	0.52	0.02	0.43
Hopscotch (HS)	0.13	0.03	0.68	0.34	0.29	0.76	0.44
Shop. Mall (SM)	0.28	0.81	0.07	0.69	0.09	0.55	0.50
All (MV)	0.49	0.24	059	-	-	-	-
All (SR)	0.68	0.87	0.02	0.03	0.09	0.24	0.14

Table A.1: **P-values of the Shapiro-Wilk Normality Test.** This table corresponds to the results that were obtained in accuracy for each approach.

Bibliography

[1] J. Bowlby, *Attachment and Loss*, vol. 1 of *Attachment*. New York: Basic Books, 1969.

[2] J. Bowlby, *Attachment and Loss*, vol. 2 of *Separation: Anxiety and anger*. New York: Basic Books, 1973.

[3] J. Bowlby, *Attachment and Loss*, vol. 3 of *Loss, sadness and depression*. New York: Basic Books, 1980.

[4] M. D. S. Ainsworth, M. C. Blehar, E. Waters, and S. Wall, “Strange Situation Procedure,” *Clinical Child Psychology and Psychiatry*, 1978.

[5] M. Dong, W. H. Giles, V. J. Felitti, S. R. Dube, J. E. Williams, D. P. Chapman, and R. F. Anda, “Insights into causal pathways for ischemic heart disease: adverse childhood experiences study,” *Circulation*, vol. 110, no. 13, pp. 1761–1766, 2004.

[6] C. J. Packard, V. Bezlyak, J. S. McLean, G. D. Batty, I. Ford, H. Burns, J. Cavanagh, K. A. Deans, M. Henderson, A. McGinty, *et al.*, “Early life socioeconomic adversity is associated in adult life with chronic inflammation, carotid atherosclerosis, poorer lung function and decreased cognitive performance: a cross-sectional, population-based study,” *BMC public health*, vol. 11, no. 1, pp. 1–16, 2011.

[7] P. Wilson, P. Bradshaw, S. Tipping, M. Henderson, G. Der, and H. Minnis, “What predicts persistent early conduct problems? evidence from the growing up in scotland cohort,” *J Epidemiol Community Health*, vol. 67, no. 1, pp. 76–80, 2013.

[8] M. H. Van IJzendoorn, “Attachment, emergent morality, and aggression: Toward a developmental socioemotional model of antisocial behaviour,” *International Journal of Be-*

havioral Development, vol. 21, no. 4, pp. 703–728, 1997.

[9] C. Hazan and P. Shaver, "Romantic love conceptualized as an attachment process," *Journal of Personality and Social Psychology*, vol. 52, no. 3, pp. 511–524, 1987.

[10] D. Oppenheim, "The attachment doll-play interview for preschoolers," *International Journal of Behavioral Development*, vol. 20, no. 4, pp. 681–697, 1997.

[11] J. Green, C. Stanley, V. Smith, and R. Goldwyn, "A new method of evaluating attachment representations in young school-age children: The Manchester Child Attachment Story Task," *Attachment & Human Development*, vol. 2, no. 1, pp. 48–70, 2000.

[12] M. Rooksby, S. Di Folco, M. Tayarani, D.-B. Vo, R. Huan, A. Vinciarelli, S. A. Brewster, and H. Minnis, "The school attachment monitor—a novel computational tool for assessment of attachment in middle childhood," *Plos One*, vol. 16, no. 7, p. e0240277, 2021.

[13] M. Esposito, L. Parisi, B. Gallai, R. Marotta, A. Di Dona, S. M. Lavano, M. Roccella, and M. Carotenuto, "Attachment styles in children affected by migraine without aura," *Neuropsychiatric Disease and Treatment*, vol. 9, p. 1513, 2013.

[14] E. Moss, C. Cyr, and K. Dubois-Comtois, "Attachment at early school age and developmental risk: examining family contexts and behavior problems of controlling-caregiving, controlling-punitive, and behaviorally disorganized children.," *Developmental psychology*, vol. 40, no. 4, p. 519, 2004.

[15] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 262–268, 2021.

[16] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 542–552, 2020.

[17] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective*

Computing, vol. 9, no. 4, pp. 578–584, 2018.

[18] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, “Depression level prediction using deep spatiotemporal features and multilayer Bi-LTSM,” *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 864–870, 2022.

[19] R. Flores, M. Tlachac, A. Shrestha, and E. Rundensteiner, “Temporal facial features for depression screening,” in *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and the ACM International Symposium on Wearable Computers*, pp. 488–493, 2022.

[20] N. Cummins, V. Sethu, J. Epps, J. R. Williamson, T. F. Quatieri, and J. Krajewski, “Generalized two-stage rank regression framework for depression score prediction from speech,” *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 272–283, 2020.

[21] Z. Zhao, Z. Bao, Z. Zhang, J. Deng, N. Cummins, H. Wang, J. Tao, and B. Schuller, “Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, 2019.

[22] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, “Affective and content analysis of online depression communities,” *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

[23] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, “Diagnosis of depression by behavioural signals: a multimodal approach,” in *Proceedings of the ACM International Workshop on Audio/visual Emotion Challenge*, pp. 11–20, 2013.

[24] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, “Automatic depression scale prediction using facial expression dynamics and regression,” in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, pp. 73–80, 2014.

[25] X. Ma, D. Huang, Y. Wang, and Y. Wang, “Cost-sensitive two-stage depression prediction using dynamic visual clues,” in *Asian Conference on Computer Vision*, pp. 338–351, Springer, 2016.

[26] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, “Artificial intelligent system for automatic depression level analysis through visual and vocal expressions,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2017.

[27] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, “Multimodal spatiotemporal representation for automatic depression level detection,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 294–307, 2023.

[28] L. Yang, D. Jiang, and H. Sahli, “Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures,” *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 239–253, 2021.

[29] Z. Huang, J. Epps, and D. Joachim, “Investigation of speech landmark patterns for depression detection,” *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 666–679, 2019.

[30] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, “Dynamic multimodal measurement of depression severity using deep autoencoding,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2018.

[31] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, “Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors,” *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, 2018.

[32] Y. Yang, C. Fairbairn, and J. F. Cohn, “Detecting depression severity from vocal prosody,” *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, 2013.

[33] H. Cai, X. Zhang, Y. Zhang, Z. Wang, and B. Hu, “A case-based reasoning model for depression based on three-electrode EEG data,” *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 383–392, 2020.

[34] J. Shen, X. Zhang, G. Wang, Z. Ding, and B. Hu, “An improved empirical mode decomposition of electroencephalogram signals for depression detection,” *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 262–271, 2022.

[35] B. Li, S. Mehta, D. Aneja, C. Foster, P. Ventola, F. Shic, and L. Shapiro, "A facial affect analysis system for autism spectrum disorder," in *2019 IEEE International Conference on Image Processing*, pp. 4549–4553, 2019.

[36] T. Guha, Z. Yang, R. B. Grossman, and S. S. Narayanan, "A computational study of expressive facial dynamics in children with autism," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 14–20, 2018.

[37] M. Leo, M. Del Coco, P. Carcagni, C. Distante, M. Bernava, G. Pioggia, and G. Palestra, "Automatic emotion recognition in robot-children interaction for ASD treatment," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 145–153, 2015.

[38] V. Yaneva, L. A. Ha, S. Eraslan, Y. Yesilada, and R. Mitkov, "Detecting autism based on eye-tracking data from web searching tasks," in *Proceedings of the International Web for All Conference*, pp. 1–10, 2018.

[39] K. Sun, L. Li, L. Li, N. He, and J. Zhu, "Spatial attentional bilinear 3d convolutional network for video-based autism spectrum disorder detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3387–3391, IEEE, 2020.

[40] A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, and V. Murino, "Video gesture analysis for autism spectrum disorder detection," in *Proceedings of the International Conference on Pattern Recognition*, pp. 3421–3426, 2018.

[41] M. D. M. J. Bovery, G. Dawson, J. Hashemi, and G. Sapiro, "A scalable off-the-shelf framework for measuring patterns of attention in young children and its application in autism spectrum disorder," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 722–731, 2021.

[42] J. Hashemi, G. Dawson, K. L. Carpenter, K. Campbell, Q. Qiu, S. Espinosa, S. Marsan, J. P. Baker, H. L. Egger, and G. Sapiro, "Computer vision analysis for quantification

of autism risk behaviors," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 215–226, 2021.

[43] P. R. K. Babu, P. Oza, and U. Lahiri, "Gaze-sensitive virtual reality based social communication platform for individuals with autism," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 450–462, 2018.

[44] Y. Tian, X. Min, G. Zhai, and Z. Gao, "Video-based early ASD detection via temporal pyramid networks," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 272–277, 2019.

[45] J. Gao, X. Jiang, Y. Yang, D. Li, and L. Qiu, "Unsupervised video anomaly detection for stereotypical behaviours in autism," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

[46] K. Campbell, K. L. Carpenter, J. Hashemi, S. Espinosa, S. Marsan, J. S. Borg, Z. Chang, Q. Qiu, S. Vermeer, E. Adler, *et al.*, "Computer vision analysis captures atypical attention in toddlers with autism," *Autism*, vol. 23, no. 3, pp. 619–628, 2019.

[47] D. Haputhanthri, G. Brihadiswaran, S. Gunathilaka, D. Meedeniya, Y. Jayawardena, S. Jayarathna, and M. Jaime, "An EEG based channel optimized classification approach for autism spectrum disorder," in *Proceedings of the Moratuwa Engineering Research Conference*, pp. 123–128, 2019.

[48] T. Eslami and F. Saeed, "Auto-ASD-network: a technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fMRI data," in *Proceedings of the ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 646–651, 2019.

[49] W. Zheng, T. Eilamstock, T. Wu, A. Spagna, C. Chen, B. Hu, and J. Fan, "Multi-feature based network revealing the structural abnormalities in autism spectrum disorder," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 732–742, 2021.

[50] S. Mostafa, L. Tang, and F.-X. Wu, "Diagnosis of autism spectrum disorder based on eigenvalues of brain networks," *IEEE Access*, vol. 7, pp. 128474–128486, 2019.

[51] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 2, pp. 1–26, 2018.

[52] W.-H. Wen, G.-Y. Liu, Z.-H. Mao, W.-J. Huang, X. Zhang, H. Hu, J. Yang, and W. Jia, "Toward constructing a real-time social anxiety evaluation system: Exploring effective heart rate feature," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 100–110, 2020.

[53] Z. Li, X. Wu, X. Xu, H. Wang, Z. Guo, Z. Zhan, and L. Yao, "The recognition of multiple anxiety levels based on electroencephalograph," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 519–529, 2022.

[54] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2016.

[55] Q. Chang, C. Li, Q. Tian, Q. Bo, J. Zhang, Y. Xiong, and C. Wang, "Classification of first-episode schizophrenia, chronic schizophrenia and healthy control based on brain network of mismatch negativity by graph neural network," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1784–1794, 2021.

[56] K.-Y. Huang, C.-H. Wu, M.-H. Su, and Y.-T. Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 393–404, 2020.

[57] D. C. Herath, C. Kroos, C. Stevens, and D. Burnham, "Adopt-a-robot: A story of attachment," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 135–136, 2013.

[58] H. Ishihara, Y. Yoshikawa, and M. Asada, "Realistic child robot "affetto" for understanding the caregiver-child attachment relationship that guides the child development," in *Proceedings of the IEEE International Conference on Development and Learning*, vol. 2, pp. 1–5, IEEE, 2011.

[59] A. Hiolle, K. A. Bard, and L. Canamero, "Assessing human reactions to different robot attachment profiles," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pp. 251–256, Sep. 2009.

[60] A. Meschtscherjakov, D. Wilfinger, and M. Tscheligi, "Mobile attachment causes and consequences for emotional bonding with mobile phones," in *Proceedings of CHI*, pp. 2317–2326, 2014.

[61] W. Odom, J. Pierce, E. Stolterman, and E. Blevis, "Understanding why we preserve some things and discard others in the context of interaction design," in *Proceedings of CHI*, pp. 1053–1062, 2009.

[62] T. Law, M. Chita-Tegmark, N. Rabb, and M. Scheutz, "Examining attachment to robots: Benefits, challenges, and alternatives," *ACM Transactions on Human-Robot Interaction*, vol. 11, no. 4, pp. 1–18, 2022.

[63] N. Freed, J. Qi, A. Setapen, C. Breazeal, L. Buechley, and H. Raffle, "Sticking together: handcrafting personalized communication interfaces," in *Proceedings of the International Conference on Interaction Design and Children*, pp. 238–241, 2011.

[64] S. Yarosh, "Designing technology to empower children to communicate with non-residential parents," *International Journal of Child-Computer Interaction*, vol. 3, pp. 1–13, 2015.

[65] C. Harbig, M. Burton, M. Melkumyan, L. Zhang, and J. Choi, "Signbright: A storytelling application to connect deaf children and hearing parents," in *CHI Extended Abstracts*, pp. 977–982, 2011.

[66] A. Edalat and F. Mancinelli, "Strong attractors of hopfield neural networks to model attachment types and behavioural patterns," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 1–10, 2013.

[67] R. Tucker, A. Edalat, and A. Faisal, "Stochastic hopfield networks to model secure and insecure attachment types with noise," *B. Eng Thesis, Imperial College London, Department of Computing*, 2013.

[68] M. Oyama-Higa, J. Tsujino, and M. Tanabiki, "Does a mother's attachment to her child affect biological information provided by the child?-chaos analysis of fingertip pulse waves of children," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 3, pp. 2030–2034, 2006.

[69] J. Tsujino, M. Oyama-Higa, and M. Tanabiki, "Measurement of ear pulse waves in children: Effect of facing another child and relationship to an action index," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, pp. 2972–2976, 2008.

[70] C. E. Izard, S. W. Porges, R. F. Simons, O. M. Haynes, C. Hyde, M. Parisi, and B. Cohen, "Infant cardiac activity: Developmental changes and relations with attachment.," *Developmental Psychology*, vol. 27, no. 3, p. 432, 1991.

[71] C. Remy, S. Gegenbauer, and E. M. Huang, "Bridging the theory-practice gap: Lessons and challenges of applying the attachment framework for sustainable hci design," in *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems*, pp. 1305–1314, 2015.

[72] D. B. Vo, S. Brewster, and A. Vinciarelli, "Did the children behave? investigating the relationship between attachment condition and child computer interaction," in *Proceedings of the International Conference on Multimodal Interaction*, pp. 88–96, 2020.

[73] A. Edalat, "Introduction to self-attachment and its neural basis," in *Proceedings of International Joint Conference on Neural Networks*, pp. 1–8, 2015.

[74] G. Spangler and K. E. Grossmann, "Biobehavioral organization in securely and insecurely attached infants," *Child Development*, vol. 64, no. 5, pp. 1439–1450, 1993.

[75] M. D. Salter Ainsworth and S. M. Bell, "Attachment, exploration, and separation: Illustrated by the behavior of one-year-olds in a strange situation," in *The Life Cycle: Readings in Human Development*, pp. 57–71, Columbia University Press, 1981.

[76] M. Dozier and R. R. Kobak, "Psychophysiology in attachment interviews: Converging evidence for deactivating strategies," *Child Development*, vol. 63, no. 6, pp. 1473–1480, 1992.

[77] C. George, N. Kaplan, M. Main, *et al.*, “Adult attachment interview,” *Interpersona: An International Journal on Personal Relationships*, 1996.

[78] H. Li, J. Cui, L. Wang, and H. Zha, “Infant attachment prediction using vision and audio features in mother-infant interaction,” in *Proceedings of the Asian Conference on Pattern Recognition*, pp. 489–502, 2020.

[79] E. B. Prince, A. Ciptadi, Y. Tao, A. Rozga, K. B. Martin, J. Rehg, and D. S. Messinger, “Continuous measurement of attachment behavior: A multimodal view of the strange situation procedure,” *Infant Behavior and Development*, vol. 63, 2021.

[80] G. Roffo, D.-B. Vo, M. Tayarani, M. Rooksby, A. Sorrentino, S. Di Folco, H. Minnis, S. Brewster, and A. Vinciarelli, “Automating the administration and analysis of psychiatric tests: The case of attachment in school age children,” in *Proceedings of CHI*, pp. 1–12, 2019.

[81] F. Parra, S. Scherer, Y. Benezeth, P. Tsvetanova, and S. Tereno, “Development and cross-cultural evaluation of a scoring algorithm for the biometric attachment test: Overcoming the challenges of multimodal fusion with “small data”,” *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 211–225, 2022.

[82] D.-B. Vo, M. Tayarani, M. Rooksby, R. Huan, A. Vinciarelli, H. Minnis, and S. A. Brewster, “The School Attachment Monitor,” in *Proceedings of the ACM International Conference on Multimodal Interaction*, pp. 497–498, 2017.

[83] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[84] P. J. Werbos, “Backpropagation through time: what it does and how to do it,” *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[85] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[86] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *International Conference on Machine Learning*, pp. 1310–1318, 2013.

[87] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[88] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[89] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[90] H. Jaeger, “The “Echo State” approach to analysing and training recurrent neural networks,” *German National Research Center for Information Technology*, vol. 148, no. 34, p. 13, 2001.

[91] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[92] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, “Learning precise timing with LSTM recurrent networks,” *Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2002.

[93] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, 2012.

[94] J. Schmidhuber, “Learning complex, extended sequences using the principle of history compression,” *Neural Computation*, vol. 4, no. 2, pp. 234–242, 1992.

[95] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to construct deep recurrent neural networks,” *arXiv preprint arXiv:1312.6026*, 2013.

[96] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[97] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[98] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, “A survey on text classification: From traditional to deep learning,” *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, 2022.

[99] Y. Kim, “Convolutional neural networks for sentence classification,” *arXiv preprint arXiv:1408.5882*, 2014.

[100] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” *arXiv preprint arXiv:1606.01781*, 2016.

[101] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” *arXiv preprint arXiv:1404.2188*, 2014.

[102] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[103] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[104] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[105] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: A review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[106] R. Ranawana and V. Palade, “Multi-classifier systems: Review and a roadmap for developers,” *International journal of Hybrid Intelligent Systems*, vol. 3, no. 1, pp. 35–61, 2006.

[107] F. Camastra and A. Vinciarelli, *Machine learning for audio, image and video analysis: theory and applications*. Springer, 2015.

[108] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, “On combining classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[109] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.

[110] J. Solomon, C. George, and A. De Jong, "Children classified as controlling at age six: Evidence of disorganized representational strategies and aggression at home and at school," *Development and Psychopathology*, vol. 7, no. 3, pp. 447–463, 1995.

[111] R. Goldwyn, C. Stanley, V. Smith, and J. Green, "The manchester child attachment story task: relationship with parental aai, sat and child behaviour," *Attachment & Human Development*, vol. 2, no. 1, pp. 71–84, 2000.

[112] H. Minnis, W. Read, B. Connolly, A. Burston, T.-S. Schumm, S. Putter-Lareman, and J. Green, "The computerised manchester child attachment story task: a novel medium for assessing attachment patterns," *International Journal of Methods in Psychiatric Research*, vol. 19, no. 4, pp. 233–242, 2010.

[113] H. Alsofyani and A. Vinciarelli, "Stacked recurrent neural networks for speech-based inference of attachment condition in school age children.," in *Proceedings of Interspeech*, pp. 2491–2495, 2021.

[114] H. Alsofyani and A. Vinciarelli, "Attachment recognition in school age children based on automatic analysis of facial expressions and nonverbal vocal behaviour," in *Proceedings of the International Conference on Multimodal Interaction*, pp. 221–228, 2021.

[115] H. Alsofyani and A. Vinciarelli, "Attachment recognition in school-age children: A multimodal approach based on language and paralanguage analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8172–8176, IEEE, 2022.

[116] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[117] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[118] V. Bettadapura, "Face expression recognition and analysis: the state of the art," *arXiv preprint arXiv:1203.6722*, 2012.

[119] I. S. Pandzic, "Mpeg-4 facial animation framework for the web and mobile applications," *MPEG-4 Facial Animation: The Standard, Implementation And Applications*, pp. 65–79, 2002.

[120] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[121] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, vol. 1, pp. 56–75, 1976.

[122] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. 1454–1462, 2014.

[123] Ekman and Friesen, "Facs - facial action coding system [online]." Available: https://www.cs.cmu.edu/~face/facs.htm, 1978. Last Checked: 2023-05-26.

[124] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 325–347, 2019.

[125] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

[126] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 59–66, 2018.

[127] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, vol. 6, pp. 1–6, 2015.

[128] P. Ekman and H. Oster, "Facial expressions of emotion," *Annual Review of Psychology*, vol. 30, no. 1, pp. 527–554, 1979.

[129] U. Dimberg and L.-O. Lundquist, "Gender differences in facial reactions to facial expressions," *Biological Psychology*, vol. 30, no. 2, pp. 151–159, 1990.

[130] G. E. Schwartz, S.-L. Brown, and G. L. Ahern, "Facial muscle patterning and subjective experience during affective imagery: Sex differences," *Psychophysiology*, vol. 17, no. 1, pp. 75–82, 1980.

[131] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[132] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[133] D. P. Ellis, "An introduction to signal processing for speech," *The Handbook of Phonetic Sciences*, pp. 755–780, 2010.

[134] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

[135] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the ACM International Conference on Multimedia*, pp. 835–838, 2013.

[136] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proceedings of Interspeech*, pp. 312–315, 2009.

[137] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the ACM Symposium on Applied Computing*, pp. 1556–1560, 2008.

[138] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the International Workshop on Audio/Visual Emotion Challenge*, pp. 3–8, 2015.

[139] A. Kapoor and R. W. Picard, “Multimodal affect recognition in learning environments,” in *Proceedings of the Annual ACM International Conference on Multimedia*, pp. 677–682, 2005.

[140] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, “Pyannote. audio: neural building blocks for speaker diarization,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7124–7128, 2020.

[141] C. Corbiere, N. Thome, A. Saporta, T.-H. Vu, M. Cord, and P. Perez, “Confidence estimation via auxiliary models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6043–6055, 2021.

[142] A. A. Taha, L. Hennig, and P. Knoth, “Confidence estimation of classification based on the distribution of the neural network output layer,” *arXiv preprint arXiv:2210.07745*, 2022.

[143] H. Jiang, B. Kim, M. Guan, and M. Gupta, “To trust or not to trust a classifier,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.