Automated Speech-based Depression Severity Assessment Using ALMD and Dual-Stream Transformer

Author Ishana Jabbar

BSc (Hons.) Computer Science
Year 4 Dissertation

Supervised by Dr. MD. Azher Uddin

Heriot-Watt University
School of Mathematical and Computer Sciences

March 2025

Declaration

I, Author Ishana Jabbar, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g., ideas, equations, figures, text, tables, programs) are properly acknowledged at any point of their use. A list of the references employed is included.

Signed:

Date:
March 27, 2025

Abstract

Depression is considered the largest cause of disability in the world, severely affecting the well-being of an individual. Early detection and treatment are considerably difficult to carry out, creating an immediate need for scalable and accessible diagnostic approaches. In recent years, automated depression severity assessment has shown impressive enhancements. Although speech is a rich medium capturing subtle change and varied vocal patterns, we see that speech-based automated depression severity estimation is less common and achieves lower performance when compared to video and multi-modal approaches.

This dissertation proposes a new framework for speech-based depression severity assessment by capturing the progression of features across frames more effectively. This is achieved by extracting motion-aware texture information from Mel spectrograms through a modified version of Adaptive Local Motion Descriptor (ALMD), namely RGB-ALMD. Further, the extracted features are then fused and processed by a Dual-stream transformer model predicting the depression severity score (BDI-II). Achieving state-of-the-art performance with RMSE and MAE, this study aims to improve the performance and robustness of speech-based depression assessment by capturing subtle indicators of emotional and psychological states, providing a foundation for scalable diagnostic tools and improving early detection accessibility.

Keywords:

Depression detection, Speech-based Depression Severity Estimation, ALMD, Adaptive Local Motion Descriptor, RGB-ALMD, Dual-stream Transformer, BDI-II, dynamic texture descriptors, hand-crafted dynamic descriptors

Acknowledgements

I am truly grateful for my supervisor, Dr. MD. Azher Uddin, whose guidance and close monitoring of my work provided me with meaningful and constant feedback throughout the project. His mentorship has been invaluable not only academically but also personally, significantly contributing to my growth as an individual. Dr. Uddin's unwavering enthusiasm and support made the dissertation process genuinely enjoyable and inspiring. His willingness to assist at all hours, including responding to my queries even as late as 1 AM, highlights his extraordinary dedication and care. He consistently encouraged me and provided practical advice on tackling challenges, teaching me valuable skills in problem-solving and project execution. I genuinely believe that completing this dissertation would have been impossible without his exceptional mentorship, encouragement, and continuous assistance.

Table of Contents

List of Figures

List of Tables

Abbreviations

ADTP Audio Delta Ternary Patterns. 7
AI Artificial Intelligence. 1, 50
ALMD Adaptive Local Motion Descriptor. iii, vii, ix, xi, xiii, xv, 0–3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27–31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51
AVEC Audio-Visual Emotion Challenge and Workshop. 8, 24
AVEC-2013 3rd Audio-Visual Emotion recognition Challenge. 4–6, 9, 10, 24
AVEC-2014 4th Audio-Visual Emotion recognition Challenge. xi, 2, 4, 6, 7, 9, 10, 12, 24, 43, 48, 50
AWS Amazon Web Services. 48

BCS British Computing Society. 50
BDI Beck Depression Inventory. 1
BDI-II Beck Depression Inventory-II. iii, 1, 2, 4, 6, 10, 13, 25, 26, 43
Bi-LSTM Bi-Long Short-Term Memory. 9, 10

CNN Convolutional Neural Network. 6, 8–11
CPU Processor. 22

DCNN Deep Convolutional Neural Network. 5, 6, 8
DCNN-LSTM Deep Convolutional Neural Network with Long Short-Term Memory. 7
DNN Deep Neural Network. 6
DSC Depression Recognition Sub-Challenge. 24
DSM-IV Diagnostic and Statistical manual of Mental disorders. 1

FDHH Feature Dynamic History Histograms. 8, 10
FFT Fast Fourier Transform. 7
FVCM Feature Variation Coordination Measurement. 7

G-SR Gaussian Staircase Regression. 4, 5
GAP Global Average Pooling. 30
GDPR General Data Protection Regulation. 50
GFN Graph Fusion Networks. 10
GPU Graphics Card. 22

HAM-D Hamilton Rating Scale for Depression. 1

HMHN Hybrid Multi-Head Cross Attention Network. 8

LASSO Least Absolute Shrinkage and Selection Operator. 6
LLDs low-level descriptors. 4–6
LPQ Local Phase Quantization. 8
LSCAformer Long and Short-term Cross-Attention-aware transFormer. 9
LSTM Long Short-Term Memory. 6, 7, 10, 11

MADRS Montgomery–Åsberg Depression Rating Scale. 1
MAE Mean Absolute Error. iii, vii, 2, 3, 7, 21, 24, 26
MAFF Multi-modal Attention Feature Fusion. 10
MFCCs Mel-frequency Cepstral Coefficients. 4, 6, 7, 15
MFM-Att Multi-level Attention mechanism. 10
MHH Motion History Histograms. 8, 10
ML Deep Learning. 31
ML Machine Learning. 1, 31
MRELBP Median Robust Extended Local Binary Patterns. 5
MRLBP-TOP Median Robust Local Binary Patterns from Three Orthogonal Planes. 8
MSN Multi-scale Spatio-temporal Network. 8

parallel-CNN Parallel-Convolutional Neural Network. 8
PHQ Patient Health Questionnaire. 1
PLS Partial Least Squares. 8, 10
PRA-Net Part-and-Relation Attention Network. 8

RAM Memory. 22
RGB Red-Green-Blue. 3
RGB-ALMD Red-Green-Blue Adaptive Local Motion Descriptor. 37
RMSE Root Mean Square Error. iii, vii, 2–5, 7, 21, 24, 26
RNN Recurrent Neural Network. 8, 11
rPPG Remote Photoplethysmographic. 10
RVM Relevance Vector Machine. 4
RVM-SR Relevance Vector Machine Staircase Regression. 4, 5

SAN Self-Attention Networks. 6
SE Squeeze-and-Excitation. 30
SER Speech Emotion Recognition. 6, 7
SM-RR Speaker Marginalization Rank Regression. 4, 5
SR Speaker Recognition. 6, 7
STA Spatio-Temporal Attention. 10

SVR Support Vector Regressor. 4, 6

TAP Temporal Attentive Pooling. 10
TMFE Transformer-based Multi-modal Feature Enhancement network. 10

VGG Visual Geometry Group. 28
VLDN Volume Local Directional Number. 9
VLDSP Volume Local Directional Structural Pattern. 9

WG-SR Weighted Gaussian Staircase Regression. 4, 5
WHO World Health Organisation. 1

1 Introduction

Depression, regarded as the single largest contributor to global disability [WHO 2020], is a highly common mental health disorder. Affecting millions globally, it causes significant negative consequences for individuals, and society as a whole. It often results in constant sadness, a loss of interest in activities, loneliness, disturbances during sleep, and impaired concentration. In some cases, depression has also led to severe physical conditions, including heart disease, Parkinson's, cancer, diabetes, and more [Aswal et al. 2018]. The World Health Organisation (WHO) reported a 25% increase of anxiety and depression was seen globally since COVID-19 [WHO 2022], with depression affecting approximately 280 million people. Not surprisingly, depression happens to be the most common cause of suicide [Nz 2014], tragically claiming over 700,000 lives each year as reported by WHO [2020, 2023]. Despite known effective treatment for mental disorders, only a small percentage of affected individuals receive the needed care [WHO 2020, 2023]. The percentage of those who receive this care is unfortunately fewer than 10% in many countries [Aswal et al. 2018; WHO 2020]. This, along with the rising prevalence of depression, raises a desperate need for innovative approaches to detect and intervene depression during the early stages. Fortunately, the evolution in Artificial Intelligence (AI) and Machine Learning (ML), has unfolded several new approaches for automated depression detection that could complement traditional diagnostic methods.

There have been multiple approaches to detect and estimate the severity of depression. Score prediction, a subset of severity estimation has emerged to be one of the most effective and reliable methods for the same by assigning a continuous value that reflects the severity of symptoms. This is highly encouraged in clinical settings, where distinguishing between the intensity and severity of depression can significantly influence what treatment the patient should receive. Several such rating scales used in depression research are Hamilton Rating Scale for Depression (HAM-D), Montgomery–Åsberg Depression Rating Scale (MADRS) and Beck Depression Inventory (BDI), Patient Health Questionnaire (PHQ) [Beck et al. 1996, 1961; Demyttenaere and De Fruyt 2003; Hamilton 1959; Maust et al. 2012; Montgomery and Åsberg 1979; Spitzer et al. 1999; Svanborg and Åsberg 2001]. One of the most widely used and reliable [Faraci and Tirrito 2013; Ginting et al. 2013; Gottfried et al. 2024; Hailu Gebrie 2018; McElroy et al. 2018] self-report scale to evaluate the severity of depressive symptoms is Beck Depression Inventory-II (BDI-II) (a revised version of BDI corresponding with the updated Diagnostic and Statistical manual of Mental disorders (DSM-IV) [Association et al. 2000] criteria for depression) [Demyttenaere and De Fruyt 2003]. This would be used to predict the depression score for the following dissertation. These scores are a categorization of depression into distinct levels as shown in Table 1.

Despite recent advances in AI, research has focused mainly on Multi-modal and Video-based depression severity assessment. Comparatively, speech-based depression severity assessment

Level	Score Range
Minimal depression	0–13
Mild depression	14–19
Moderate depression	20–28
Severe depression	29–63

Table 1. Depression Severity Levels

is less common although indicative states of depression such as the subtle changes in speech patterns and vocal characteristics can be found.

The existing work in speech-based depression severity assessment mostly tend to achieve a higher error rate when compared to those in video as well as multi-modal based. They also focuses on deep networks that are stacked, which can affect the model's performance to handle variations in audio quality [Yin et al. 2023]. Furthermore, to the best of our belief, the studies on speech-based automated depression severity assessment have not explored the use of applied dynamic texture descriptors in speech-based depression detection. Addressing these gaps, a novel hybrid architecture that integrates advanced feature extraction methods such as Adaptive Local Motion Descriptor (ALMD) [Uddin et al. 2017] on 3 channels is proposed, capturing dynamic elements within the audio. These features are fused and evaluated using a dual-stream Transformer model architecture [Vaswani 2017] to measure the BDI-II score. The 4^{th} Audio-Visual Emotion recognition Challenge (AVEC-2014) [Valstar et al. 2014] dataset is used for the training and evaluation of the model.

1.1 Aim

The aim of the dissertation is to improve speech-based depression detection through a framework applying advanced feature extraction methods ensuring dynamic elements are taken into consideration, a machine learning model, and evaluation.

1.2 Objectives

We achieve this aim by:

- Developing an end-to-end framework to assess severity of speech-based depression.
- Segmenting the audio into spectrograms.
- Capturing dynamic information from spectrograms.
- Implementing a dual-stream transformer model to predict depression severity score.
- Training and evaluating the proposed framework using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).
- Comparing the model's error rate against the state-of-the-art models.

In this dissertation, by incorporating RGB-ALMD with a dual-stream transformer (as discussed in Section 3), we achieved state-of-the-art performance with RMSE of 5.60 and MAE of 4.48.

Despite promising results, this study was limited by its dataset size as well as cultural homogeneity, which could affect broader generalizability. Additionally, the computational complexity of the proposed model poses challenges for real-time deployment. Future work should focus on expanding datasets for diverse populations, optimizing model efficiency, and integrating multimodal inputs to further enhance the accuracy and scalability of depression severity assessments.

The structure of the rest of the dissertation is as follows: Chapter 2 focuses on literature related to automated depression severity assessment, highlighting previous methodologies, critically analysing work done as well as find gaps. Chapter 3 presents the proposed methodology, detailing the implementation of ALMD, and the proposed RGB-ALMD alongside the dual-stream transformer architecture. Chapter 4 outlines the project requirements, including functional and non-functional aspects, as well as hardware and software specifications. Following, Chapter 5 discusses the evaluation of the model, including a description of the dataset, the metrics used to evaluate the model, and the interpretation of results. Finally, Chapter 6 provides concluding remarks and discusses future work directions.

2 Literature Review

This chapter looks into the different approaches and models used in the automated severity assessment of depression. The most widely researched modalities are Speech-based, Video-based and Multi-modal based in no specific order. We shall look into the previous research done in these areas in the sub-sections below. Speech-based automated depression severity assessment in Section 2.1 will be examined in detail as it directly aligns with our proposed idea. This will be followed with a critical analysis in Section 2.4 as well as a comparison of related work in Section 2.5.

2.1 Speech-based Automated Depression Severity Assessment

In the 3rd Audio-Visual Emotion recognition Challenge (AVEC-2013), Valstar et al. [2013] utilized crucial audio features from 3-second short segments using the openEAR toolkit [Eyben et al. 2009]. These features such as energy and spectral related low-level descriptors (LLDs), assisted in predicting continuous values for valence and arousal using Support Vector Regressor (SVR). This research was framed as the baseline and deemed depression severity assessment as a regression problem. In the proceeding challenge, 4th Audio-Visual Emotion recognition Challenge (AVEC-2014), Valstar et al. [2014] utilized audio and video features to predict depression severity, with audio features comprising LLDs as done in [Valstar et al. 2013] as well as MFCCs. The focus was on capturing the rhythm and acoustic patterns from tasks such as reading and free-form speech. This challenge established a benchmark for RMSE in predicting BDI-II scores, facilitating comparative research.

Building upon the foundation laid by AVEC-2013 and AVEC-2014 challenge, [Cummins et al. 2015c] considered Relevance Vector Machine (RVM)[1], selected because of their potential advantages over SVR as mentioned in [Tipping 2001] and [Tipping 2003]. In the context of speech depression severity assessment, datasets are often limited in number of speakers and duration. RVMs are well suited to shorten this gap as it performs dimensionality reduction and feature selection on the dataset. Features were extracted through a brute-force approach outputting a wide range of speech features such as pitch variability, formant frequencies, sub-band energy variability and other paralinguistic cues. This was tested on both AVEC-2013 and AVEC-2014, where the results of AVEC-2013 outperformed the other.

Cummins et al. [2017] continued to build upon his previous work, using just AVEC-2013. The authors compared various regression approaches to address the irregularities between the features of audio and the stages of depression. The variations included Gaussian Staircase Regression (G-SR), Weighted Gaussian Staircase Regression (WG-SR), Relevance Vector Machine Staircase Regression (RVM-SR), and Speaker Marginalization Rank Regression (SM-RR)

[1] a Bayesian regression approach that has become widely popular for various speech-based regression tasks

[Cummins et al. 2015a; Kaya et al. 2014; Valstar et al. 2013; Williamson et al. 2013], where each was designed to address distinct feature spaces and specific conditional ranking functions. The authors have also tested out 3 combinations as shown in Table 2, out of which the fusion of G-SR, WG-SR and SM-RR gave an RMSE of 8.16 which was the lowest known RMSE for the dataset AVEC-2013.

System	**RMSE**
Baseline [Valstar et al. 2013]	14.12
Brute-Force & Decision-tree [Kaya et al. 2014]	9.78
G-SR [Williamson et al. 2013]	8.50
WG-SR [Cummins et al. 2015a]	9.75
RVM-SR [Cummins et al. 2017]	9.86
SM-RR [Cummins et al. 2017]	9.64

(a) Results compared in [Cummins et al. 2017]

Fused Systems	**RMSE**
WG-SR + RVM-SR + SM-RR	9.26
WG-SR + RVM-SR + SM-RR + G-SR	8.27
WG-SR + SM-RR + G-SR	8.16

(b) Results achieved by fused systems Cummins et al. [2017]

Table 2. Comparison of systems and fused systems for depression detection

Deep-learned features derived from neural networks outperform hand-crafted features across various domains. This was proven by He and Cao [2018] who proposed a novel approach combining the two features to predict depression severity from speech signals. Over 2 thousand baseline audio features were extracted, Median Robust Extended Local Binary Patterns (MRELBP) was applied to spectrograms generated for each audio clip. Deep-learned features were extracted from two Deep Convolutional Neural Network (DCNN), one taking the raw speech as input and the other using spectrograms. 20 seconds was found to be the optimal length for the LLDs. The two DCNNs were then joined to boost performance. Among hand-crafted features and deep-learned features, the later achieved better results. This showed that deep-learned model can help predict depression better and the spectrogram DCNN represents the characteristics of depression well.

Niu et al. [2019] introduced a hybrid network, extracting MFCCs segments of speech through Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Deep Neural Network (DNN). This combination captured depression-related information in various representations including spatial and temporal changes as well as a discriminative representation. Using p-norm pooling combined with Least Absolute Shrinkage and Selection Operator (LASSO), utterance-level features are created from segment-level features. Classification is done using SVR, to predict the BDI-II scores. Results had outperformed previous approaches, largely due to the optimization of the pooling parameter and the effectively captured high-level features relating to depression.

Integrating Self-Attention Networks (SAN) with DCNN, Zhao et al. [2020] proposed Figure 1, a hybrid feature extraction model for depression severity assessment from audio data. The hybrid network made use of LLDs and 3D log-Mel spectrograms to capture long-term dependencies and local temporal structures respectively through SAN and DCNN. Segment-level complementary features are formed by combining the outputs from the independently trained models. These features are then fed into a SVR for BDI-II score prediction. On both datasets, AVEC-2013 and AVEC-2014, results outperformed baselines and other models.

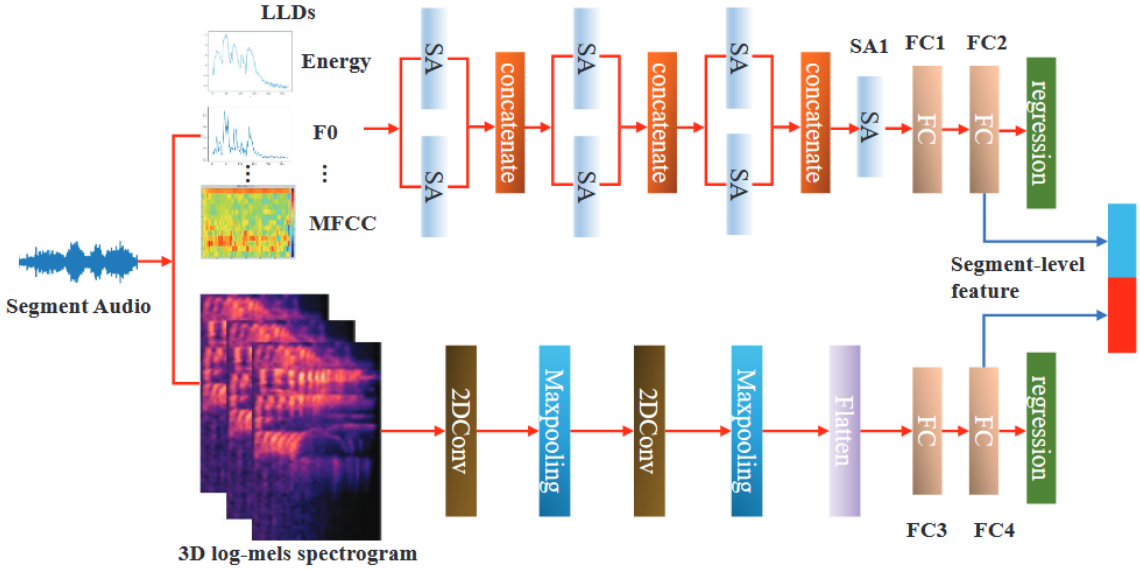


Fig. 1. Hybrid Network for extracting segment-level complementary features [Zhao et al. 2020]

A hierarchical model introduced by Dong and Yang [2021], integrates deep Speaker Recognition (SR) and Speech Emotion Recognition (SER) features to identify vocal and emotional cues

in speech. Features are extracted by extracting SR and SER features from spectrograms using a pre-trained ResNet-50. A Feature Variation Coordination Measurement (FVCM) is used to further analyze temporal patterns and correlations on the obtained feature matrices. The first layer of the hierarchical model, predicts the depression severity regression interval for the recordings determined by training many fuzzy classifiers. In the second layer, a regressor is trained using the deep speech coordination features that was learned from the first layer's classifiers. The approach achieved RMSE value of 8.82 on and MAE value of 6.79, surpassing other speech-based models that existed. These results were also quite competitive to video and multi-modal systems.

Fu et al. [2022] used the AVEC-2014 dataset to focus on spectral and temporal dynamics. Traditional audio features (extracted using MFCCs and Fast Fourier Transform (FFT) spectrograms) along with Audio Delta Ternary Patterns (ADTP), proposed by the author, captures temporal movements in speech frequencies. The MFCCs and FFTs images, are then provided to a Deep Convolutional Neural Network with Long Short-Term Memory (DCNN-LSTM), extracting high-level features, comprising two LSTMs layers and three fully connected layers in a joint tuning configuration to integrate ADTP, MFCCs, and FFT deep features. This model is depicted in Figure 2.

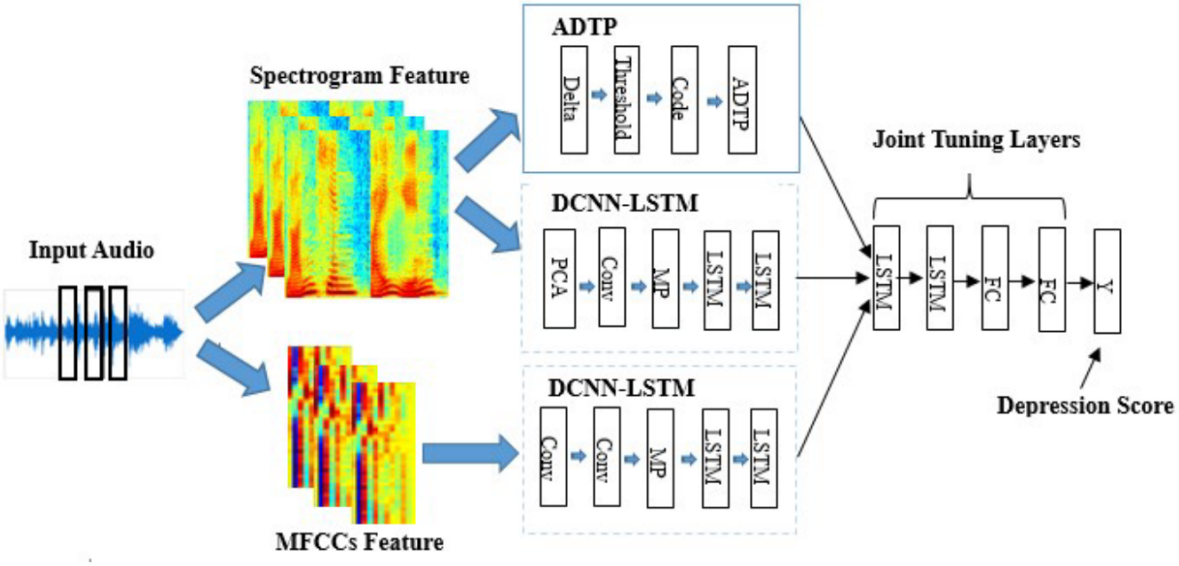


Fig. 2. Model proposed by Fu et al. [2022] to capture temporal motion features of depression

Yin et al. [2023] addresses limitations in traditional deep-learning methods, which rely on single-stream stacked networks. In this case, capturing the entire range of depression indicators in audio data might not be possible. To overcome this limitation, they integrate a Parallel-Convolutional Neural Network (parallel-CNN) with a transformer to capture both local and temporal features in speech.

2.2 Video-based Automated Depression Severity Assessment

Here, we see the evolution from simple CNNs to complex architectures that integrate spatial and temporal features together with attention mechanisms. Early models like the ones implemented by Meng et al. [2013] and Valstar et al. [2013] focused on hand-crafted features, such as Motion History Histograms (MHH) and Local Phase Quantization (LPQ), applying Partial Least Squares (PLS) regression on the AVEC datasets.

With the adoption of CNNs, Zhu et al. [2017] introduced a two-stream CNN architecture capturing both facial appearance and motion, and applying the appearance and dynamics DCNN, while Jan et al. [2017] combined CNNs with their model, Feature Dynamic History Histograms (FDHH) to enhance temporal feature extraction. Building on CNNs, He et al. [2018] proposed Median Robust Local Binary Patterns from Three Orthogonal Planes (MRLBP-TOP), a novel dynamic feature descriptor extracting dynamic features from face.

The recent emergence of advanced architectures, utilizes both spatial and temporal dimensions of facial data. Researchers, Al Jazaery and Guo [2018]; Zhou et al. [2020] used 3D-CNNs and Recurrent Neural Network (RNN)s to improve spatio-temporal feature extraction. Addressing the limited dynamic encoding of traditional 2D CNNs and the dependence of 3D CNNs on temporal information from a single range, De Melo et al. [2020] introduced a 3D Multi-scale Spatio-temporal Network (MSN), combining 3D CNNs for capturing facial dynamics in depression with an exploration of different temporal ranges. This outperformed simpler CNNs in depression detection.

Recent works have explored attention mechanisms extensively. While researchers like He et al. [2021]; Jianwen and Xiao [2023] incorporated attention mechanism to improve spatial focus and emphasize on local and global relevant facial areas, Li et al. [2023] developed a Hybrid Multi-Head Cross Attention Network (HMHN) capturing complex relationships among depression-related features from various key facial areas, reducing the error rate of depression assessment. Liu et al. [2023] proposed Part-and-Relation Attention Network (PRA-Net) enhancing depression features by separating feature maps into different parts of representations while applying self-attention and relation attention mechanisms. This approach strengthens the model's ability to identify specific facial regions which display the most depressive symptoms, allowing the model achieve state-of-the-art performance.

Advanced frameworks using Bi-LSTM and Transformer models have further improved temporal analysis. This is seen from Uddin et al. [2022], who introduced Volume Local Directional Structural Pattern (VLDSP), addressing limitations in extracting finer facial motion details in Volume Local Directional Number (VLDN) [Uddin et al. 2020], while mentioning the importance of facial dynamics [He et al. 2021], and reducing the computational complexity observed in previous methods such as [De Melo et al. 2020]. They also utilized the Inception-ResNet-v2 network [Szegedy et al. 2016], extracting visual spatial features. These extracted features were then fed into a CNN and a Bi-LSTM model for temporal analysis as depicted in Figure 3.

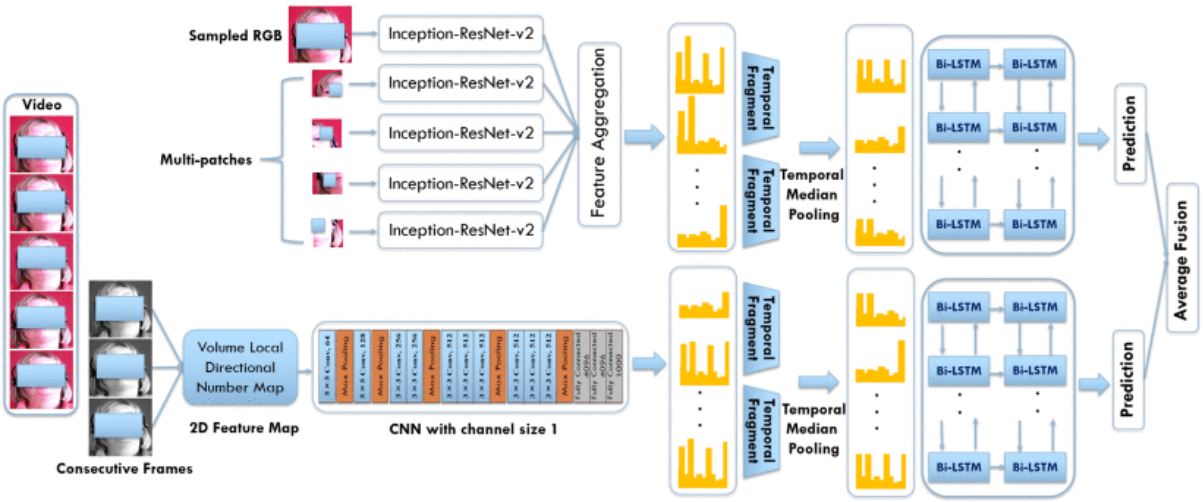


Fig. 3. The model proposed by Uddin et al. [2022] involving advanced frameworks

He et al. [2024] used Long and Short-term Cross-Attention-aware transFormer (LSCAformer), employing a dual-branch system capturing long and short term temporal features and combined them using cross-attention mechanisms. This structure helped capture the full range of depressive facial cues.

Finally, Lage Cañellas et al. [2023] proved the importance of pre-processing and scheduling techniques by using ResNet-50 with optimized pre-processing and achieving comparable results to more complex architectures on both AVEC-2013 and AVEC-2014 datasets.

2.3 Multi-Modal Automated Depression Severity Assessment

As seen in Section 2.2, Meng et al. [2013] extracted dynamic features basing on MHH and applied PLS regression to capture the relationship between the depression label and the feature for facial features. This was also applied on a combination of features of changing facial and vocal expressions, tested upon using AVEC-2013. The model was among the first ones that used both facial and vocal expressions in a dynamic context. Jan et al. [2017] extracted features from visual as spoken in the previous sub-section which was then fused with CNN for facial feature extraction and FDHH for audio processing. Their model achieved a good performance on the AVEC-2014 dataset.

Focusing on facial expressions for recognizing depression, Niu et al. [2020] employed a STA network in combination with Multi-modal Attention Feature Fusion (MAFF). Audio and video frames were segmented to extract relevant features using MAFF. This approach is applied to both spatial and temporal data to improve depression predictions.

Uddin et al. [2022] developed a comprehensive multi-modal framework using both audio and video data, applying MAFF pooling along with spatio-temporal networks and Temporal Attentive Pooling (TAP). TAP focused on segment level as well as temporal features achieving a reliable estimation of the BDI-II depression scores.

Fang et al. [2023] proposed a multi-modal Fusion model with a Multi-level Attention mechanism (MFM-Att) depicted in Figure 4. This was designed to capture intra-modal and inter-modal attention using LSTMs, Bi-LSTM and attention mechanisms across audio, visual, and text data. The model utilized various features from various modalities to enhance performance. This was followed by Fan et al. [2024], who introduced a Transformer-based Multi-modal Feature Enhancement network (TMFE) combining visuals, speech, and Remote Photoplethysmographic (rPPG) signals. They also integrated inter-modal and intra-modal Transformers to improve feature extraction. Further, Graph Fusion Networks (GFN) was employed and deep CNNs were then applied to extract the audio and video abstract features. State-of-art results was achieved on both AVEC-2013 and AVEC-2014.

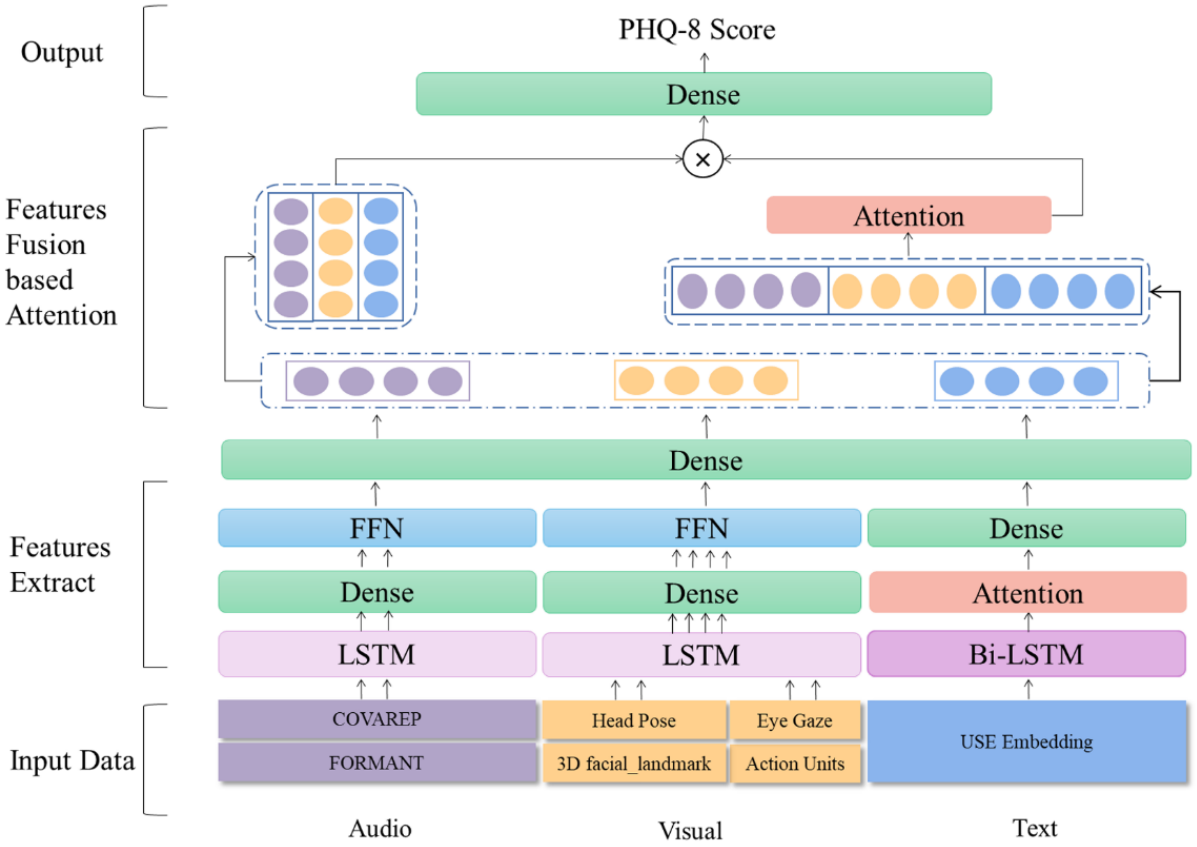


Fig. 4. A tri-modal method proposed by Fang et al. [2023] capturing Audio, Visual and Text based features

2.4 Critical Analysis on Related Work

With the recent advancements in depression detection models, there has been a large shift to deep learning-based models. While speech-based methods have good potential, they result in a higher error rate when compared to video-based or multi-modal approaches. Nevertheless, audio remains a promising modality if the features were considered in detail, capturing small fluctuations and hesitations in speech.

Most speech-based models use conventional deep learning models like CNNs and LSTMs for feature extraction and classification [Cummins et al. 2017; He and Cao 2018; Niu et al. 2019; Yin et al. 2023]. These models, although standardized, they might miss subtle, unique audio features that could enhance the depression detection prediction. Uddin et al. [2022]; Zhao et al. [2020] relied on RNNs but found it difficult to capture long-range dependencies in audio data. This in turn, limited their effectiveness for lengthy recordings. He and Cao [2018];

Niu et al. [2019] applied hand-crafted static feature extraction approaches but this requires domain expertise, making the approach less adaptable and more labor-intensive compared to automated deep learned features. Furthermore, only a limited number of studies incorporate hand-crafted dynamic descriptors such as Fu et al. [2022] who employed hand-crafted dynamic descriptors to extract temporal features from spectrograms. Those spectrograms were then used as input for 3D convolutional models but failed to sufficiently integrate complementary temporal descriptors, critical for capturing the progression of temporal patterns in acoustic features.

Till current date, no research in speech-based depression has implemented dynamic texture descriptors to capture the missing variations of vocal characteristics in spectrograms. These gaps underscore the need for models using both dynamic texture descriptors along with hand-crafted dynamic descriptors, motivating the idea of our proposed method in Section 3.

2.5 Comparison of Related Work

As discussed earlier, video-based and multi-modal models currently surpass audio-only methods in terms of performance. Hence, there exists a scope for improvement in speech-based depression detection. Table 3 compares the various automated speech-based depression approaches that use the AVEC-2014 seen in Section 2.1.

Study	**RMSE (Test)**	**MAE (Test)**
Baseline [Valstar et al. 2014]	12.57	10.03
Cummins et al. [2015c]	10.99	N/A
He and Cao [2018]	9.99	8.19
Niu et al. [2019]	9.66	8.02
Zhao et al. [2020]	9.57	7.94
Dong and Yang [2021]	8.82	6.79
Fu et al. [2022]	9.27	7.26
Uddin et al. [2022]	8.46	6.95

Table 3. Comparison of various speech-based depression studies tested on the AVEC-2014

In Section 3 we shall take a detailed look into our methodology which will be followed by the requirements necessary for the project in Section 4.

3 Proposed Methodology

As shown in Figure 5, the proposed model is an end-to-end framework for automated depression severity assessment. The processing starts with the segmentation of audio into equal segments from each raw audio. Each audio segment is then transformed into a spectrogram to visualize the audio, to extract the frequency and intensity over time. Adaptive Local Motion Descriptor (ALMD)[2] [Uddin et al. 2017] would be utilized on RGB spectrograms to capture the dynamic and temporal aspects. These features will be used to predict the depression severity scores using BDI-II scores through a dual-stream Transformer [Vaswani 2017]. This section provides an in-depth explanation of the data pre-processing pipeline (applied to the AVEC2014 dataset), the feature extractors, including their individual workings, the justification for their inclusion, and how they collectively contribute to the effectiveness of the system.

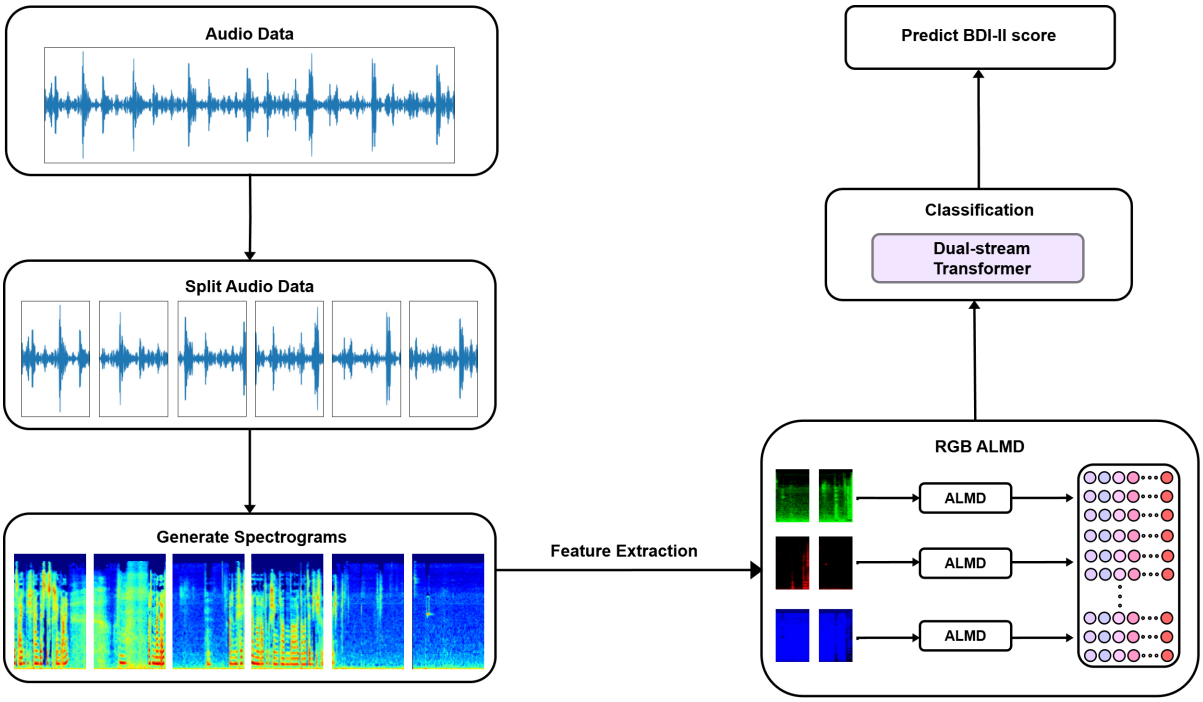


Fig. 5. Our proposed method

[2]ALMD, inspired by Histogram of Optical flow [Perš et al. 2010] and Local Ternary Pattern [Tan and Triggs 2010]

3.1 Pre-processing

Appropriate audio pre-processing is a crucial step in emotion and depression recognition tasks because it has a direct influence on the quality of input data and thereby on the performance of downstream analytical models. Poorly pre-processed audio can lead to incorrect feature extraction and deteriorate the model's ability to successfully capture fine-grained emotional cues.

Initially, the audio data was pre-processed through silence trimming to remove low-energy parts and resampling to ensure consistency. The speech signals were split into equal-length intervals, and a strategic division of n frames (n = 10 chosen empirically) was employed for the purpose of achieving a trade-off between context and detailed analysis.

As the performance of emotion and depression recognition systems is highly dependent on the quality of input data [Venkataramanan and Rajamohan 2019], these audio segments were then converted to Mel Spectrograms with 128 Mel frequency bands between 20 Hz and 8000 Hz. All spectrograms were on the decibel (dB) scale, with red representing high-energy frequencies, green and yellow representing medium-energy frequencies, and blue representing low-energy frequencies. Per-channel normalization was done for all spectrograms. Figure 6 displays the mel spectrograms generated from an audio.

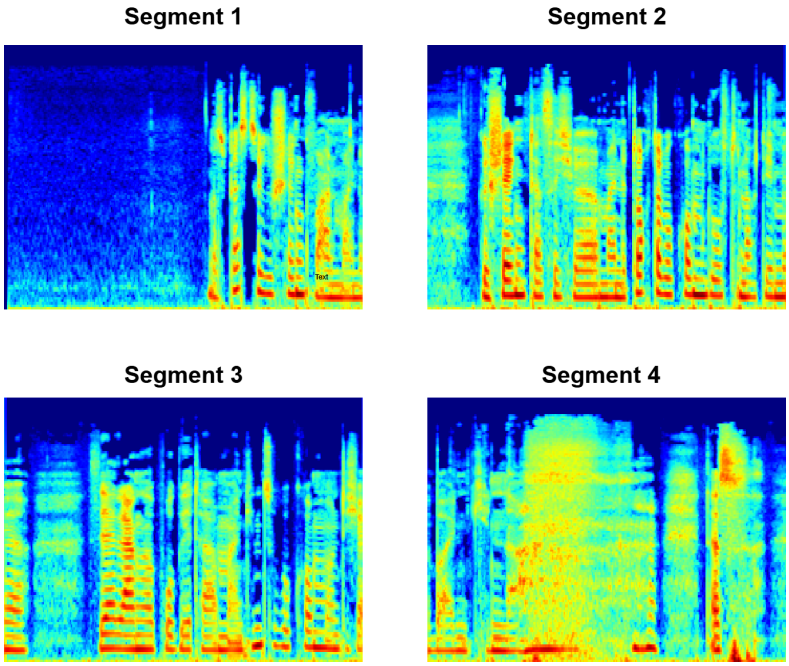


Fig. 6. Four consecutive Mel spectrograms from a segmented audio of 10 frames

Mel Spectrograms were specifically selected due to their proven efficacy in depicting subtle speech frequency changes that are crucial for effective emotion and depression identification, as supported by recent research in the field [Das and Naskar 2024; Kadam et al. 2024a,b; Meng et al. 2019].

The spectrograms obtained were used in the feature extraction model developed, which utilized the handcrafted descriptor Adaptive Local Motion Descriptor (ALMD) in estimating the severity of depression. The approach successfully takes advantage of the powerful audio features depicted through Mel Spectrograms.

3.2 RGB-ALMD

As discussed in Section 2.4, most previous works have overlooked the importance of handcrafted dynamic descriptors in speech-based depression assessment. The integration of the Adaptive Local Motion Descriptor (ALMD) addresses this limitation by capturing long-term and consistent spectral transitions that are frequently missed by static descriptors such as MFCCs [Fu et al. 2022; Niu et al. 2019]. As opposed to static approaches, ALMD leverages temporal dynamics and can represent subtle changes in energy, frequency, and spectral content between consecutive spectrogram frames. This renders ALMD very effective in capturing depression-related voice modulations, such as monotonic speech patterns, slow tempo, and energy distribution changes, that might not be modeled by deep learning models by themselves.

The ALMD algorithm generalizes local pattern descriptors such as Local Binary Patterns (LBP) [Ojala et al. 2002] and Local Ternary Patterns (LTP) [Tan and Triggs 2010] to the analysis of spectrograms of audio signals. Such descriptors have been popularly applied in the computer vision community for texture classification, facial expression recognition, and dynamic event detection in video. In audio signal processing, ALMD generalizes such spatial-temporal descriptors to describe spectrograms of audio signals as if they were image sequences. Essentially, ALMD considers two temporally adjacent spectrogram frames as two consecutive 'image frames' and calculates differences of corresponding pixels or frequency bins between them. This pixel-wise comparison brings out binary patterns of local motion or change in spectral characteristics, which are then compiled into histograms, encoding dynamic behaviors over the audio spectrogram.

Formally, ALMD computation involves two consecutive frames—the previous and next frame—and is mathematically defined as follows:

$$\mathrm{ALMD}_{n,r}^{\mathrm{upper}}(x_c, y_c) = \sum_{k=1}^{n-1} \Big[U_g\big(g_{p1} - g_{nc}\big) \;\oplus\; U_g\big(g_{p1} - g_{nc}\big) \Big] 2^k \tag{1}$$

$$\mathrm{ALMD}_{n,r}^{\mathrm{lower}}(x_c, y_c) = \sum_{k=1}^{n-1} \Big[L_g\big(g_{p1} - g_{nc}\big) \ \oplus \ L_g\big(g_{p1} - g_{nc}\big) \Big] \, 2^k \tag{2}$$

where

$$U_g(a) = \begin{cases} 1, & \text{if } a \geq 0, \\ 0, & \text{otherwise,} \end{cases} \qquad L_g(a) = \begin{cases} 1, & \text{if } a < 0, \\ 0, & \text{otherwise.} \end{cases}$$

Here, n and r represent the number and radius of neighboring pixels, respectively. The terms g_{p1} and g_{nc} denote the neighbor pixels of the previous and next frames, while g_{nc} is the center pixel of the next frame. The threshold is adaptively chosen as the average of the medians computed from the absolute differences between the center pixel and its neighboring pixels across the next frames. For example, the center location (x, y) can be calculated as

$$x = \frac{x_p + x_n}{2}, \quad y = \frac{y_p + y_n}{2},$$

where (x_p, y_p) and (x_n, y_n) are corresponding positions in the previous and next frames, respectively.

Initially, ALMD was employed on single-channel (greyscale) images, in our case spectrograms. In this configuration, the descriptor identified and encoded dynamic patterns based solely on intensity variations. While effective, this single-channel approach inherently limited the ability of ALMD to distinguish distinct frequency bands since different frequencies often carry crucial and differing cues related to depression. Low-frequency bands tend to convey fundamental frequency and pitch information, mid-frequency bands contain important formant structure and voice quality information, and high-frequency bands convey transient and articulatory information. A greyscale presentation, which did not have the discrete spectral channels, effectively blurred these informative spectral differences, potentially reducing the discriminative power of the resulting features.

To address this crucial limitation, we introduced a three-channel (RGB) version of the ALMD called RGB-ALMD, specially tailored to the needs of speech depression analysis. RGB-ALMD subsequently computes ALMD separately on successive frames for each of the three channels. This extension allows RGB-ALMD to follow frequency-specific transitions by exploiting the color dimension to encode fine spectral motion patterns that would otherwise be inseparable in a greyscale image.

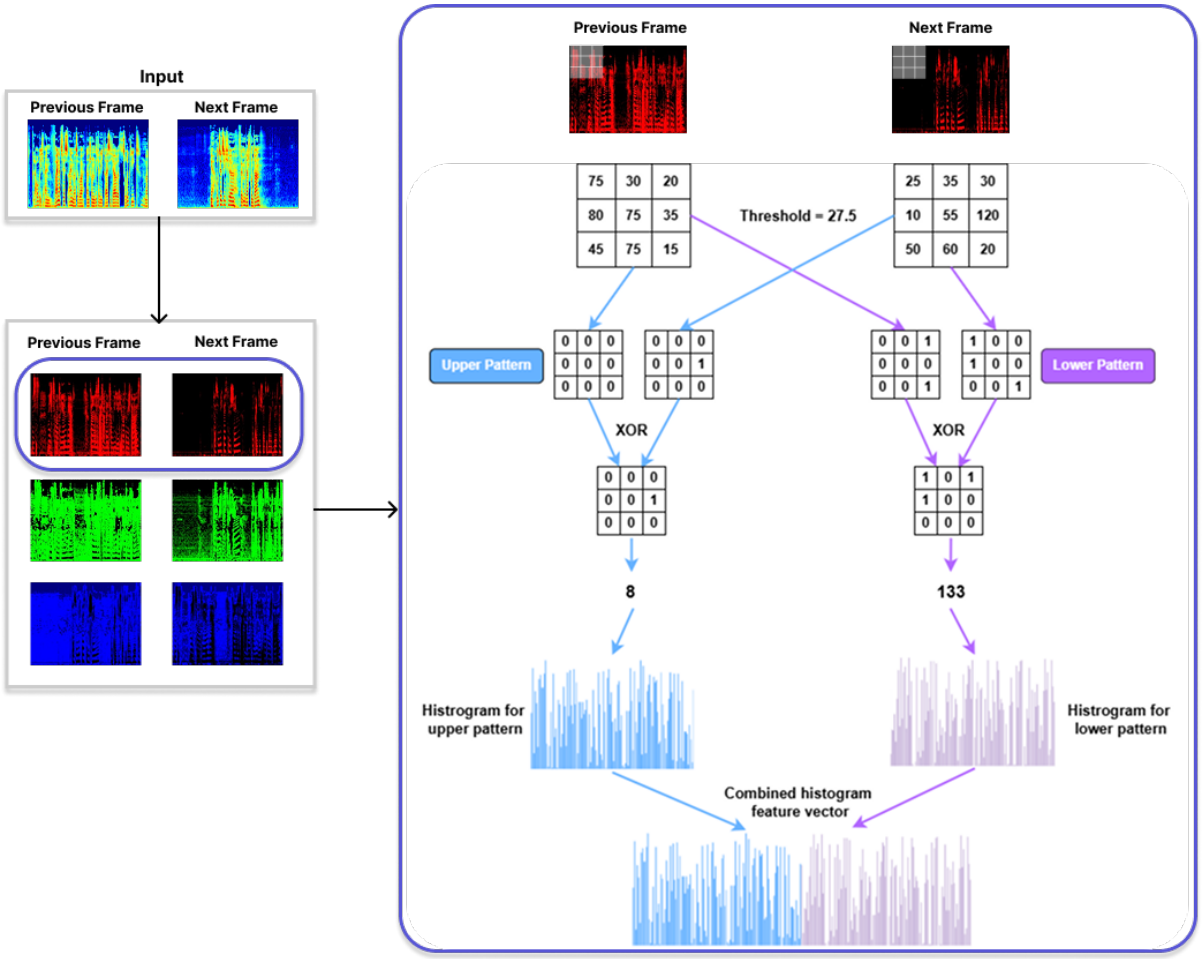


Fig. 7. RGB-ALMD performed on a spectrogram

By effectively capturing and encoding these multi-channel spectral transitions, our RGB-ALMD is able to provide a far richer and more subtle feature representation. This representation is more capable of differentiating depression-relevant speech attributes that are subtle, including prosody changes, spectral energy distribution, and rhythm. This improvement also bridges handcrafted feature engineering and modern deep learning methods quite well, taking advantage of both to obtain improved predictive performance.

3.3 Dual-Stream Transformer

In the procedure of capturing long-term temporal dependencies in speech, a transformer model is crucial, significantly enhances the system ability for depression severity estimation with high accuracy. Unlike traditional sequential models such as recurrent neural networks, Transformers rely on self-attention mechanisms to model interactions between features at various time steps efficiently. The next section elaborates on the fine-grained structure of our proposed Transformer, its layers, specialized modules, and the rationale behind the design in light of recent literature.

Our model employs a dual-stream architecture where features of two disparate speech tasks (Freeform and Northwind) are processed individually. This is inspired by modality-specific processing demonstrated in the ViT model by Dosovitskiy et al. [2020] and LXMERT by Tan and Bansal [2019]. By not fusing the streams until late-stage fusion, each stream learns task-specific embedding features, thereby effectively avoiding early fusion noise and cross-modal interference. This is necessary to capture modality-specific nuances critical to the regression-based depression prediction task.

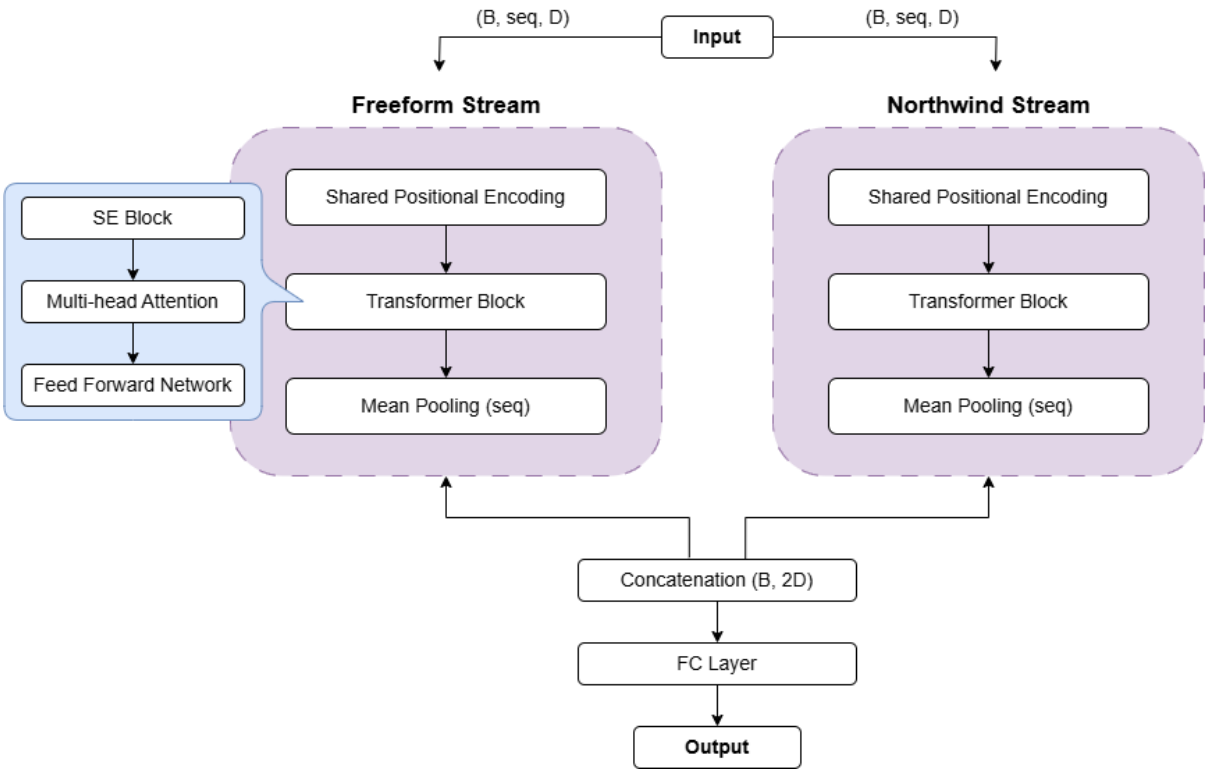


Fig. 8. Dual-Stream Transformer approach

3.3.1 Modified Positional Encoding

The Transformer employs positional encoding to maintain the order of sequence data. Sinusoidal positional encoding as presented in [Vaswani 2017] is modified in our model such that a flexible logarithmic base depending on the feature size is used, tailored to handle the diverse sequence lengths of the AVEC2014 dataset. This adjustment is also in accordance with the results of Transformer-XL [Dai et al. 2019], where flexible positional encoding prevents saturation and guarantees stable gradient flow, which is particularly useful for time sequences of varying lengths.

3.3.2 SEBlock Integration for Channel-Wise Attention

A Squeeze-and-Excitation (SE) Block, as employed by Hu et al. [2018], is inserted into each Transformer block to perform channel-wise attention gating prior to self-attention computation. The SE mechanism dynamically re-weights feature channels based on their importance:

$$Scale = \sigma(W_2 \cdot GELU(W_1 \cdot GAP(x)))$$

where GAP stands for Global Average Pooling.

SE blocks significantly enhance transformer performance, particularly on small datasets like AVEC2014, by emphasizing salient emotional and prosodic features and inhibiting redundant or noisy feature dimensions.

3.3.3 Multi-Head Self-Attention

Following the SEBlock, our architecture employs multi-head self-attention layers. This allows the model to simultaneously focus on information from different representation subspaces, effectively capturing diverse temporal patterns in the speech data. Multi-head attention facilitates the modeling of both local and global temporal dependencies, crucial for assessing speech-based depression severity.

3.3.4 Feedforward Network

Each Transformer block concludes with a position-wise feedforward network, which consists of two linear transformations with a GELU activation. This component introduces nonlinearity capacity to the Transformer for modeling complex, high-dimensional speech feature interactions, to further enhance the temporal embedding representation.

3.3.5 Mean Pooling Strategy

Instead of introducing additional complexity in the form of recurrent layers or learned tokens, we apply a simple yet effective mean pooling operation along the sequence axis

(seq_len=number_of_segments). Inspired by the results of Reimers and Gurevych [2019], mean pooling enables stable aggregation of temporal information, summarizing nicely contextual speech patterns without introducing too many parameters, thereby avoiding overfitting.

3.3.6 Fully Connected Regression Head

Following concatenation of the independently learned Freeform and Northwind stream embeddings, the combined representation is passed through a fully connected regression head. This head is comprised of a linear transformation followed by ReLU activation, dropout (with probability p=0.3), and a final linear layer mapping to a single scalar depression severity score. This design choice aligns with best practices for regression on small datasets by Ba et al. [2016], encouraging robust training and generalization by preventing overfitting.

The effectiveness of the proposed dual-stream transformer architecture was enhanced through its deliberate design choices.

First, the features are extracted through independent dual streams separately for the 2 tasks, enabling the preservation of task-specific characteristics. This approach is supported by prior work such as Dosovitskiy et al. [2020]; Tan and Bansal [2019], who specify the benefits of separating streams in multi-modal learning contexts.

Second, channel-wise attention enhancement is incorporated via Squeeze-and-Excitation (SE) blocks, which have been validated by Hu et al. [2018] to amplify discriminative features—an especially crucial factor in affective computing tasks. Third, we enhanced temporal modeling by using a modified positional encoding which was inspired by Dai et al. [2019]; Vaswani [2017]. Through this, our model was able to better capture dependencies in the audio sequences.

Furthermore, effective fusion of the tasks and regularization techniques are employed through mean pooling along with a carefully crafted regression head using dropout combined with Layer Normalization to avoid overfitting. These design decisions are best practice as attested in a number of works such as Devlin et al. [2019]; Reimers and Gurevych [2019].

Together, these design decisions—including dual-stream processing, SE attention, adaptive positional encoding, and judicious regularization—constitute a very robust architecture for precise modeling of complex depression-related vocal patterns.

4 Project Requirements

This section provides a structured and comprehensive overview of the different requirements needed for the success of the project.

4.1 Functional Requirements

Table 4 below displays the functional requirements needed for the implementation of the model.

ID	Requirement Description	Priority	Status
FR1	Generate spectrograms for different audio segments	MUST	Complete
FR2	Split the raw audio into different segment counts (e.g., 5, 10, 15)	MUST	Complete
FR3	Organize the generated spectrograms into separate folders for each segment count	MUST	Complete
FR4	Store the pre-processed data in the same folder to reduce redundant processing	SHOULD	Complete
FR5	Extract dynamic and temporal features from each spectrogram using ALMD	MUST	Complete
FR6	Implement the dual-stream Transformer model	MUST	Complete
FR7	Validate the model on the test split	MUST	Complete
FR8	Evaluate the model's error rate using RMSE and MAE	MUST	Complete
FR9	Experiment with the hyper-parameters for optimal performance	MUST	Complete
FR10	Determine the best-performing model and hyper-parameters	MUST	Complete
FR11	Save the trained model features for reuse	SHOULD	Complete
FR12	Provide a configuration file for optimal parameters	COULD	Partial
FR13	Create a script to run the model with customizable parameters	WOULD	Incomplete

Table 4. Functional Requirements

4.2 Non-Functional Requirements

In this model, the non-functional requirements necessary are outlined by Table 5.

ID	Requirement Description	Priority	Status
NFR1	Store data securely, ensuring no personal identification from audio data	MUST	Complete
NFR2	Maintain stability and consistency in model outputs across runs	SHOULD	Complete
NFR3	Ensure localized data processing to prevent data loss	MUST	Complete
NFR4	Aim for superior performance compared to benchmark models	SHOULD	Complete
NFR5	Well documented code-base and models for ease of functionality extension	SHOULD	Complete

Table 5. Non-Functional Requirements

To ensure efficient training, testing, and inference of the proposed Dual-Stream Transformer model with RGB-ALMD, the system must meet the following minimum hardware and software requirements. The specifications below reflect the environment used during the experimental evaluation.

4.3 Hardware Requirements

The minimum hardware configuration necessary for smooth execution and training is summarized in Table 6. This setup ensures compatibility with GPU-accelerated frameworks and sufficient computational capacity to handle high-resolution spectrograms and large batch sizes.

ID	Component	Requirement Description
HR1	Processor (CPU)	Intel Core i7-13620H
HR2	Graphics Card (GPU)	NVIDIA GeForce RTX 4060 GPU
HR3	Memory (RAM)	16 GB RAM (recommended)
HR4	Storage	1 TB SSD

Table 6. Minimum Hardware Requirements

4.4 Software Requirements

The software environment includes multiple open-source libraries and frameworks necessary for audio pre-processing, spectrogram generation, model training, and evaluation. The following tools and dependencies were used:

- **Python 3.9**
- **PyTorch-GPU (1.12)**
- **TensorFlow-GPU** and **Keras**
- **OpenCV**
- **Scikit-learn**
- **Matplotlib / Seaborn**
- **Optuna**

These tools were installed and managed in an Ubuntu 20.04 LTS environment using Windows Subsystem for Linux (WSL2), ensuring compatibility with GPU acceleration and ease of reproducibility.

In the following section, we shall take a further look into the evaluation aspect of the model.

5 Evaluation

In this section, we shall take a look into the dataset being used to evaluated the proposed model (Figure 5) as well as the evaluation metrics that shall be used to test the performance and error rates of the model.

5.1 Dataset

The AVEC-2014 dataset is widely preferred for depression severity assessment, due to its multi-modal data (audio and video) enabling standardized comparisons using metrics like RMSE and MAE while promoting advancements in audio-based, video-based and multi-modal approaches. As for speech-related research, AVEC-2014 provides diverse low-level and dynamic features, making it particularly effective for research.

In this study, we utilize a smaller part of the AVEC-2013 audio-visual depression corpus [Valstar et al. 2013], the AVEC-2014[3] dataset [Valstar et al. 2014] shown in Figure 9 which was developed as part of the Audio-Visual Emotion Challenge and Workshop (AVEC) series. This challenge has two sub-challenges, we shall be looking at the **Depression Recognition Sub-Challenge (DSC)** which requires to predict the level of self-reported depression. AVEC-2014 is well used and recognized, designed for research in affective computing and mental health analysis, supporting both emotion recognition and depression severity estimation tasks.

Fig. 9. Samples of the AVEC2014 dataset

[3]Only audio files of the dataset would be leveraged for this study

The dataset contains a total of 300 audio-visual recordings of 84 German-speaking participants, each engaging in two tasks:

(1) **Northwind Task:** The participants read aloud an excerpt from the German fable "The North Wind and the Sun".
(2) **Freeform Task:** The participants answer one open-ended personal questions about topics such as childhood memories, favorite dish, or best gift.

The audio for each recording was resampled to a uniform bitrate of 128kbps, while the video was standardized to a resolution of 640 x 480 pixels with a frame rate of 30 fps.

The tasks discussed above were then split into three equal segments (50 recordings each of the Northwind task and the Freeform task in each of the 3 segments), while ensuring there is a balanced distribution in terms of age, gender, and depression levels. The labels are present for the Training and the Development set.

The depression score for each recording session is predicted using the BDI-II scores which serve as the ground truth for model evaluation, representing various levels of depression severity [Beck et al. 1996]. (Refer to Table 1) These scores are used to assess the model's capability to estimate depression severity. The models are trained and tested on BDI-II scores which are seen as the target labels. The evaluation metrics used for the models shall be further discussed in Section 5.2.

5.2 Evaluation Metrics

The model's error rate is tested using the two extensively used prediction accuracy metrics used for regression, namely Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) and shall be further discussed below. Here, errors are the differences between the BDI-II score predicted by the regression model and the actual BDI-II values.

5.2.1 Root Mean Square Error (RMSE)

RMSE is a standard metric to evaluate the accuracy of predictions. It measures the average magnitude of prediction errors, giving more weight to large errors. RMSE is given by the formula:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where y_i and $\hat{y}_i$ represents the actual and predicted BDI-II scores, respectively.

Higher model performance is achieved by lower RMSE values, indicating that there are fewer large errors in the model's predictions. As RMSE is the sum of the squared errors, it could be highly affected by outliers. This could lead to a worst overall RMSE just with a few wrong predictions [Campana and Delmastro 2017].

5.2.2 Mean Absolute Error (MAE)

MAE calculates the average absolute difference between model's predicted BDI-II score and the actual BDI-II scores. Unlike RMSE, MAE treats all of the errors equally, providing a simpler measure of model performance. It is given by the formula:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

where y_i and $\hat{y}_i$ represents the actual and predicted BDI-II scores, respectively.

Here, better model performance is achieved by lower MAE, indicating that there errors and significantly minimized.

These metrics aim to provide reliable predictions to develop the models aligning with clinical standards, ensuring a reliable assessment of our model's performance in estimating speech-based depression severity.

6 Experiments

To assess the effectiveness of our proposed dual-stream Transformer architecture for audio-based depression severity prediction, a set of experiments were carried out using the AVEC2014 dataset. The evaluation was designed to validate the model's capability in capturing subtle temporal and spectral cues associated with depressive speech patterns, and to benchmark its performance against standard regression metrics (RMSE and MAE).

6.1 Experimental Setup

The experiments of our proposed dual-stream Transformer model were conducted on a system with an Intel Core i7 processor, 16GB RAM, and an Nvidia RTX 4060 GPU. The entire implementation was carried out using Python with PyTorch as the deep learning framework, running on an Ubuntu environment via Windows Subsystem for Linux (WSL).

Prior to training, the dataset was split according to the official AVEC2014 partitions, where the original training and validation subsets were combined for creating a training set. The test set remained unchanged and was used only evaluating on our final model to ensure consistency and comparability of results.

Training was conducted with 200 epochs, while employing Adam optimizer due to its adaptive optimization capabilities. Model performance was measured using standard regression metrics: RMSE and MAE.

The hyperparameters for the model were tuned using Optuna across 200 optimization trials. A diverse set of parameters was explored to identify the most effective configuration for performance on the depression prediction task. These included the number of attention heads (`num_heads`) with values of 2, 4, and 8; feedforward dimensions (`ff_dim`) ranging from 250 to 500 in increments of 50; and the number of dense layer units (`dense_units`) from 100 to 500 in steps of 20. Dropout rates (`dropout_rate`) were tested between 0.1 and 0.4, incremented by 0.1, while learning rates (`learning_rate`) were sampled from 1e-3, 1e-4, and 1e-5. Several loss functions were also evaluated, including Huber loss, Log-Cosh, Mean Squared Error (MSE), Weighted MSE, a combined MSE-MAE loss, and RMSE.

After tuning, the optimal hyperparameters identified were: `num_heads` = 2, `ff_dim` = 400, `dense_units` = 460, `dropout_rate` = 0.3, `learning_rate` = 0.001, and the selected loss function was Huber loss.

7 Experiment and Result Analysis

This section looks into an evaluation of our proposed approach on the AVEC2014 depression dataset. In order to systematically evaluate the impact of different design decisions on our framework, we conducted a series of ablation experiments depicted in Section 7.1. We began by establishing baseline results using a few machine learning models, and then progressed to more complex deep learning architectures. Finally, we focused on Transformer-based models, introducing several enhancements and comparing different feature extraction strategies.

7.1 Ablation Study

We examine the progression from simpler or partial configurations to our final, best-performing model. We begin by comparing different versions of ALMD including greyscale, channel-wise, and RGB variants to highlight the benefits of multi-channel motion features. We then explore how incremental modifications to the Transformer architecture (positional encoding, Squeeze-and-Excitation blocks, and dual-stream fusion) affect performance.

Subsequently, we investigate the effectiveness of different spatial feature extraction strategies by contrasting VGG-16, ResNet-101, and the proposed RGB-ALMD. We also compare classical machine learning methods (ANN, SVR, Random Forest) against the Dual-Stream Transformer, and then contrast several deep learning baselines (1D CNN, LSTM, BiLSTM, GRU) with our final approach. Further analysis is carried out to find the influence of segment granularity (10, 25, and 50 frames per segment) on model accuracy. Finally, we present a scatter plot showing the correlation between actual and predicted depression severity scores for the best model, demonstrating how closely our system's predictions align with ground-truth BDI-II labels.

The comparison in Figure 10, shows that greyscale ALMD variant exhibits one of the highest prediction error, whereas the multi-channel RGB-ALMD achieves the lowest error. Each color channel (red, green, blue) highlighted distinct aspects of temporal dynamics. In terms of individual channels, the Red and Blue channels outperformed greyscale ALMD (Red: 6.01 MAE / 7.77 RMSE, Blue: 6.10 MAE / 7.74 RMSE vs. greyscale: 6.48 MAE / 8.37 RMSE). This motivated us to create an enhanced RGB-ALMD leveraging full RGB information to be developed (refer to Section 3.2) to capture complementary dynamics more effectively. The combined RGB-ALMD descriptor performed best, reducing the error to 4.48 MAE and 5.60 RMSE. This substantial improvement confirms that incorporating multi-channel (color) dynamic texture information captures more complementary depression-related cues, outperforming any single-channel or greyscale approach. Consequently, the RGB-ALMD features were deemed the most effective representation for subsequent experiments.

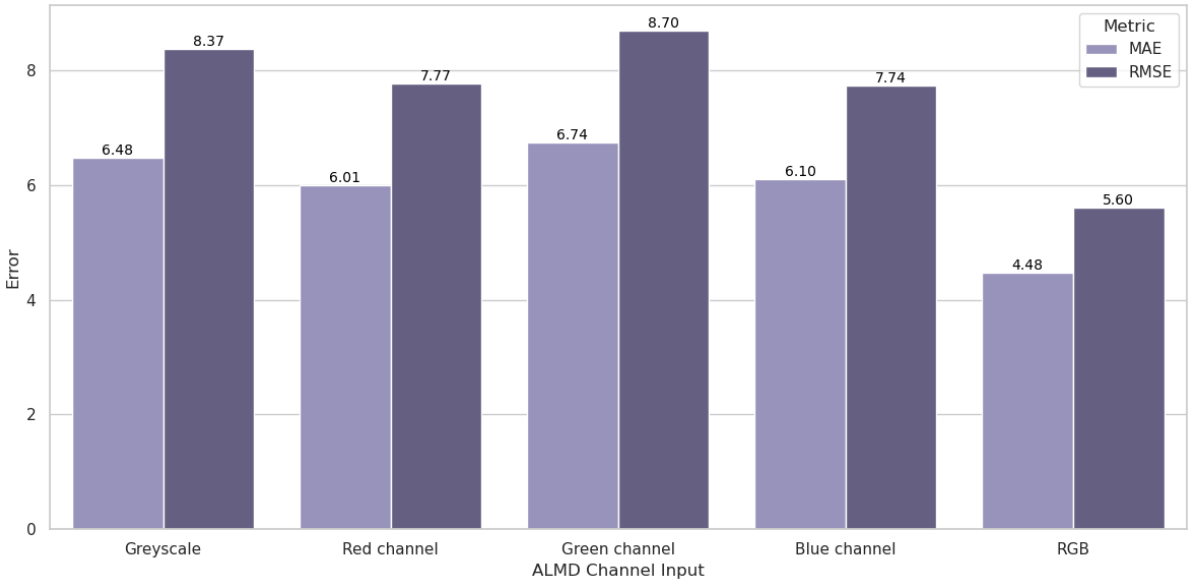


Fig. 10. Comparison of ALMD vs RGB-ALMD on dual-stream transformer

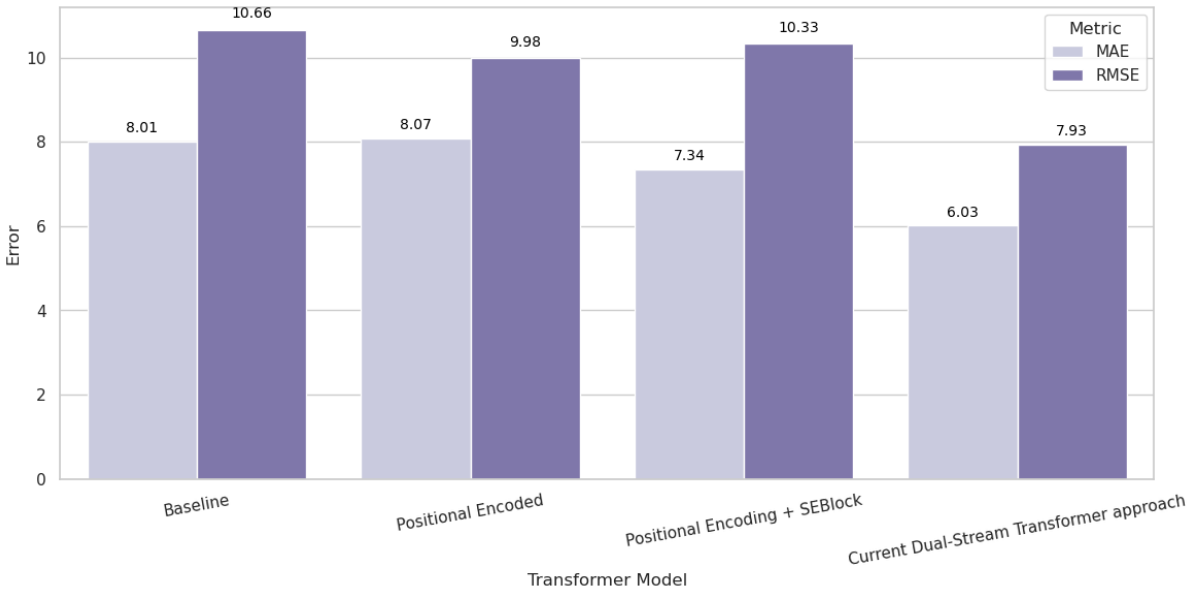


Fig. 11. Different transformer models and their results on ResNet-101

We further validated our dual-stream Transformer, by comparing it with other transformer models and enhancements as illustrated in Figure 11, run on ResNet-101. Positional encoding

significantly improved model accuracy by explicitly encoding temporal order, crucial for sequential data. Adding a Squeeze-and-Excitation (SE) block further improved performance by adaptively emphasizing informative features. Finally, the Dual-Stream Transformer (combining Freeform and Northwind) achieved the lowest error metrics with 6.03 MAE and 7.93 RMSE, confirming that both tasks have features that should be checked separately, as it greatly enhances depression severity predictions.

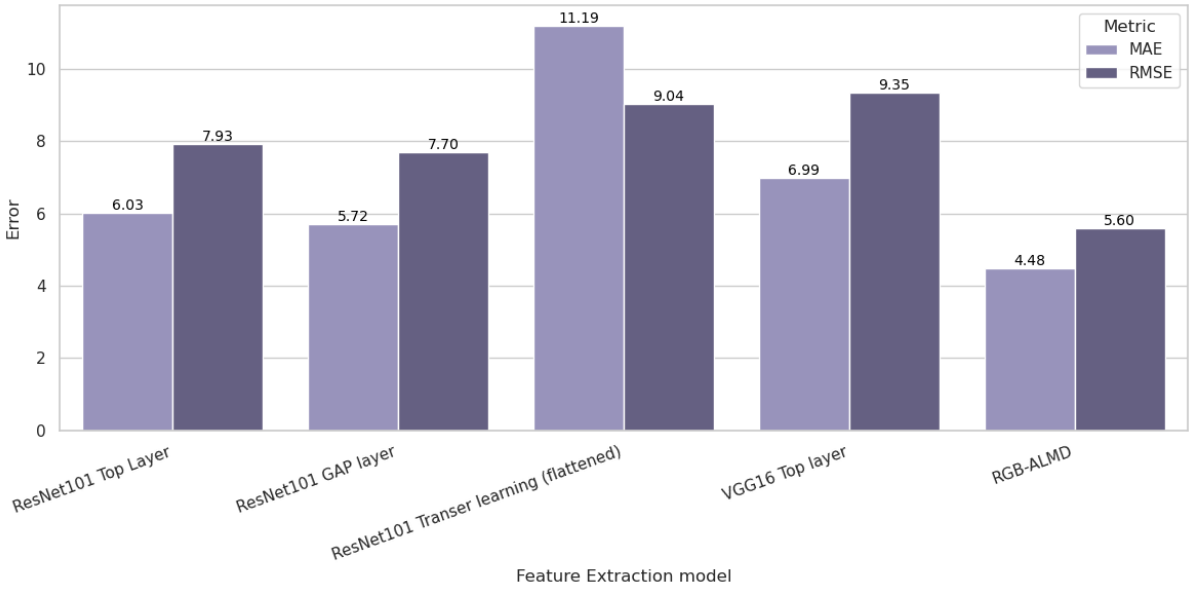


Fig. 12. Comparison of different feature extraction results

To further validate our chosen hand-crafted descriptor, we compare it with other CNN feature extraction models (see Figure 12). Due to RGB-ALMDs temporal feature extraction, we see the results are substantially better than those of spatial feature extraction models. ResNet-101 substantially outperformed VGG-16 due to its deeper architecture and residual connections, providing more discriminative features crucial for capturing detailed spectral patterns. When comparing the models individually, VGG-16 had underperformed with 6.99 MAE and 9.35 RMSE. We perform transfer learning on ResNet-101 but unfortunately this scores the worst. However, ResNet-101 with GAP layer achieved an impressive MAE of 5.72 and RMSE of 7.70 when compared to the other spatial descriptors. This could be due to the Average pooling layer being able to display more important features. Nevertheless, it was not able to beat the performance shown by RGB-ALMD.

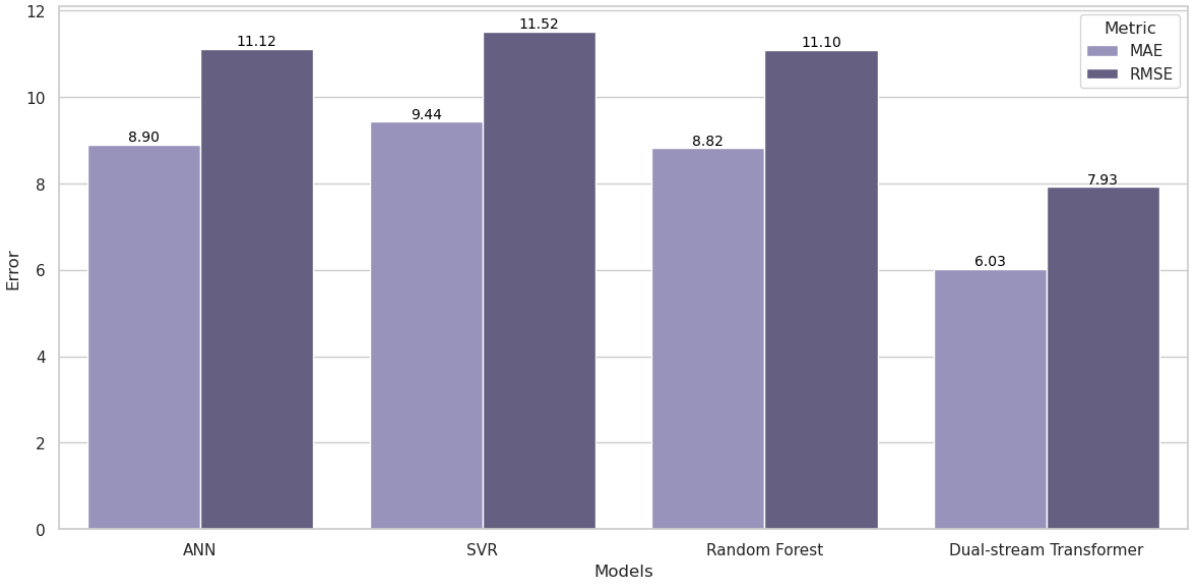


Fig. 13. Machine Learning model's performance on ResNet-101

We also compared our dual-stream transformer to three classical machine learning regressors: Support Vector Regression (SVR), a Artificial Neural Network (ANN), and Random Forest Classifier. Figure 13 summarizes their performance on the depression severity prediction task. Overall, these methods inadequately modelled the temporal dynamics inherent in audio data, underscoring the need for more powerful deep learning methods. Although the models generally yielded suboptimal results, the dual-stream transformer model emerged as the most promising performer. Specifically, Random Forest recorded an MAE of 8.82 and RMSE of 11.10, while SVR showed 9.44 MAE and 11.52 RMSE. We see that Random Forest showed slightly better robustness yet still lacked on overall prediction accuracy. The ANN achieved 8.90 MAE and 11.12 RMSE, further confirming the inadequacy of these classical regressors for this task.

With the previously shown results, where our chosen deep learning architecture performed the best, we were motivated to compare our model with other such ML models. These included RNNs and convolutional models, to capture the sequential nature of speech data better. Specifically, we implemented an Conv1D, 1DCNN, LSTM, BiLSTM, GRU, and an initial Transformer model. Figure 14 illustrates their performance. The 1DCNN performed the worst among the models. Among the RNNs, BiLSTM achieved an MAE of 8.81 and RMSE of 10.70, while LSTM and GRU recorded similar performances at 8.55 / 10.85 and 8.50 / 10.63, respectively. This improved accuracy (compared to ML models) confirmed the benefit of modelling temporal sequences. BiLSTM marginally outperformed the other RNNs, proving that forward and backward contexts enhance depression-related feature extraction. The Conv1D model achieved

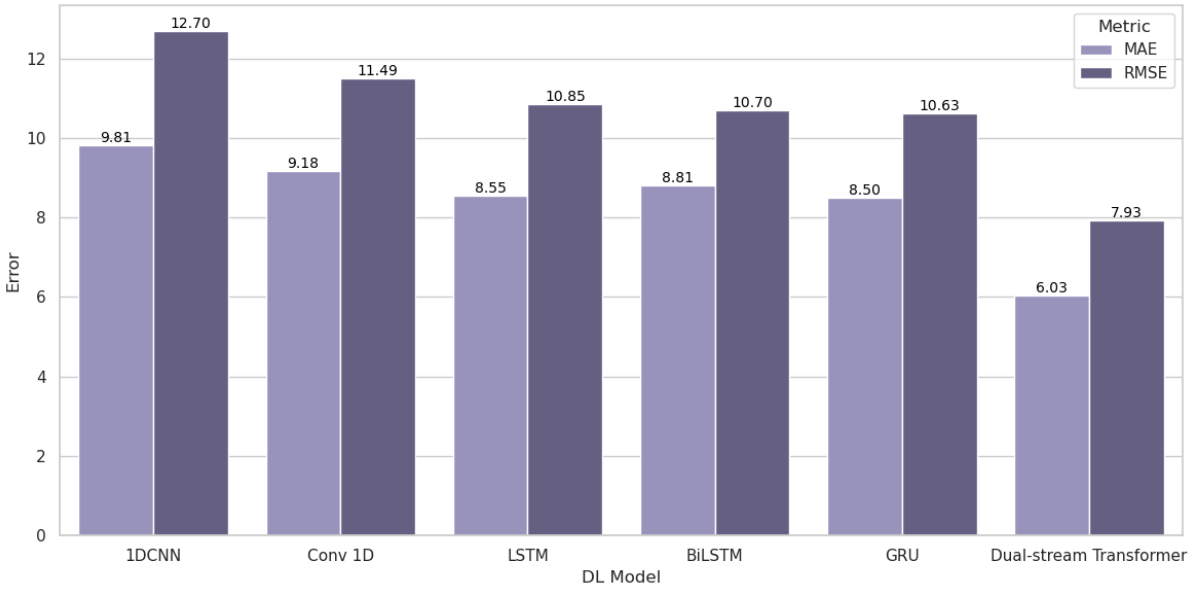


Fig. 14. Deep Learning model's performance on ResNet-101

competitive results, suggesting convolutional filters could effectively extract local patterns. The dual-stream transformer on the other hand, outperformed all others while producing MAE of 6.03 and RMSE of 7.93 due to its superior capacity in modelling complex temporal dependencies using self-attention.

The influence of segmentation of the audio on model performance is analyzed in Figure 15. Across the segment sizes, we see that RGB-ALMD maintained superior performance with an average MAE below 5.2 and RMSE below 6.5. Interestingly, certain segmentation settings showed minimal additional gain from static feature fusion, suggesting ALMD's standalone robustness at optimal segment lengths. With an optimal 10 segments of the audio, RGB-ALMD performed with the best result: 4.48 in MAE and 5.60 in RMSE.

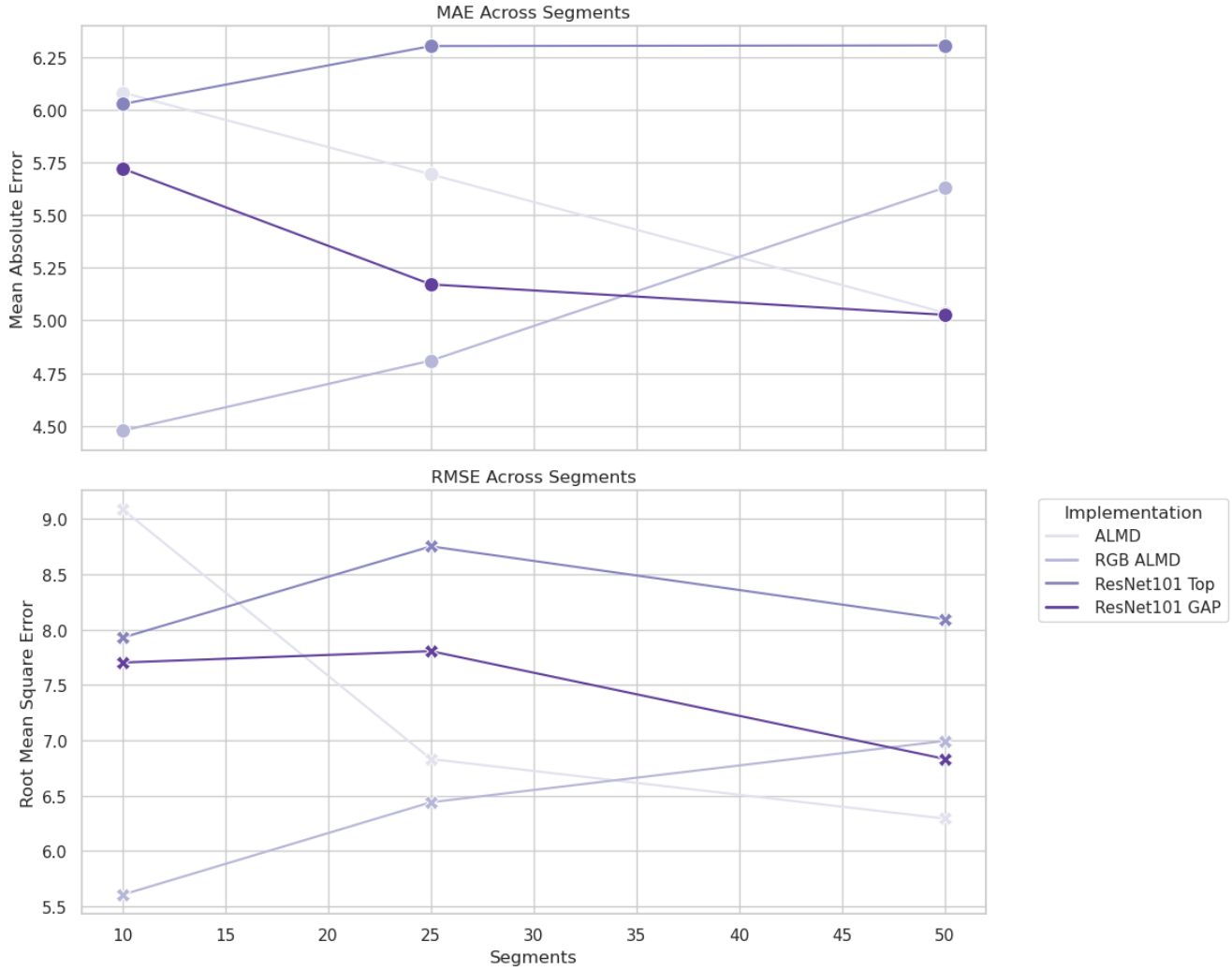


Fig. 15. Performance of Extracted features on 10, 25 and 50 segments of audio

Hence an optimal segment length of 10 frames with RGB-ALMD performed on dual-stream transformer was chosen which yielded the lowest errors by balancing temporal resolution and context.

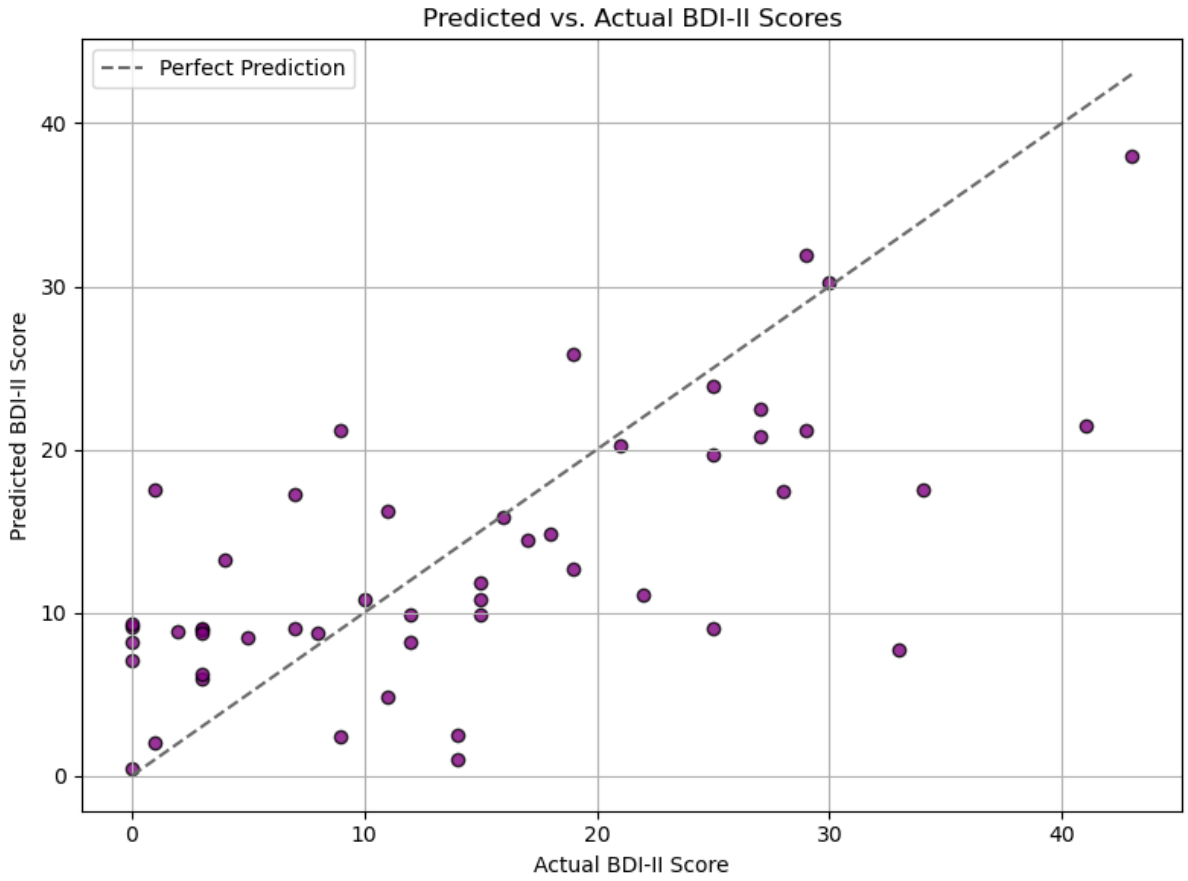


Fig. 16. Scatter plot showing the actual and predicted values distribution from our best performing model

Figure 16 above plots the predicted BDI-II scores versus actual scores of the best performing 10 segmented RGB-ALMD on Dual-Stream Transformer (see Figure 15). The graph shows strong positive correlation and close clustering around the diagonal, indicating accurate predictions with minor deviations. Outliers exist, reflecting slight regression effects, but overall predictive capability remains comparably reliable.

7.2 Comparison with State of the Art

The proposed Dual-Stream Transformer combined with our RGB-ALMD achieves lower error rates than all prior works on AVEC2014 in the context of audio. The model's prediction error is the smallest in both metrics, with **RMSE of 5.60** and **MAE of 4.48**, outperforming the nearest competitors by a significant margin.

Taking the most notable recent researches, Uddin et al. [2022] reported an RMSE of 8.46 and MAE of 6.95, while Dong and Yang [2021] achieved 8.82 RMSE and 6.79 MAE. Our method reduces these errors by roughly ~**35%** in both RMSE and MAE. This improvement across both error metrics establishes our approach as the new state-of-the-art for speech-based depression severity prediction on AVEC2014. The reduction in these error rates also indicates a closer alignment with actual BDI-II scores as shown previously, highlighting the effectiveness of combining hand-crafted dynamic texture descriptors with a Dual-Stream Transformer.

Study	**RMSE (Test)**	**MAE (Test)**
Baseline [Valstar et al. 2014]	12.57	10.03
Cummins et al. [2015b]	10.99	N/A
He and Cao [2018]	9.99	8.19
Niu et al. [2019]	9.66	8.02
Zhao et al. [2020]	9.57	7.94
Dong and Yang [2021]	8.82	6.79
Fu et al. [2022]	9.27	7.26
Uddin et al. [2022]	8.46	6.95
Proposed Framework	**5.60**	**4.48**

Table 7. RMSE and MAE comparison of various speech-based depression studies tested on the AVEC-2014

In summary, by capturing long-term spectral dynamics using RGB-ALMD and modeling them with a robust dual-stream attention architecture, our model sets a new performance benchmark, surpassing all contemporary studies by a considerable margin.

8 Conclusion

In this concluding chapter, we provide an overview of the study's contributions, discuss its main limitations, and outline potential directions for future research.

8.1 Summary

This dissertation displayed an **end-to-end Dual-Stream Transformer framework incorporating RGB-ALMD** for automated speech-based depression severity assessment. The proposed architecture was designed to extract both detailed spectral features and long-term temporal patterns from Mel spectrogram representations of audio signals. In this pipeline, raw audio was segmented, converted to spectrograms, and processed through RGB-ALMD to extract handcrafted temporal motion features, which were then modeled using a Dual-Stream Transformer for final regression.

The RGB-ALMD module could detect frequency energy changes across the Red, Green, and Blue channels, while the transformer layers captured long-range dependencies using its multi-head self attention. This unified integration of handcrafted dynamics and deep sequential modeling enabled more expressive feature learning for BDI-II score estimation.

The results validated that incorporating RGB-ALMD—rather than traditional greyscale descriptors—and coupling it with a task-specific dual-stream Transformer significantly improved prediction performance. The system not only exceeded machine learning and deep learning baselines but also surpassed multiple prior state-of-the-art methods by a considerable margin.

Concluding, our study has successfully met its objective by proposing a novel, accurate, and interpretable system for speech-based depression assessment. The contributions of this work lie in advancing the methodology for modeling dynamic temporal changes in speech, hence offering a stronger foundation for developing future clinical tools for mental health diagnostics.

8.2 Main Limitations of Work

Although the introduced system demonstrates promising performance in automatic speech-based severity assessment of depression, there are certain drawbacks:

First, the study suffers from the relatively limited and non-diversified dataset, with the data being solely German audio recordings. This narrowness may confine the generalizability of the findings across languages, dialects, and cultural settings [Cummins et al. 2015b; Valstar et al. 2014]

Second, the dual-stream transformer model, while providing great performance, suffers from the high computation and large parameter numbers and therefore presents difficulties in real-time deployment and in resource-constrained settings [Sanh et al. 2019; Sun et al. 2020]

Lastly, the proposed model is currently only tested and confined to audio modality. Integrating complementary modalities such as facial expressions or body cues would incorporate contextual information and further tune the severity estimate for depression [Chen et al. 2024].

8.3 Future Work

Although the introduced system establishes a new benchmark in automatic depression severity evaluation via speech, there are several directions for improvement. Future research would be facilitated with the incorporation of alternative spatial feature extractors beyond RGB-ALMD. For example, the employment of advanced architectures such as DenseNet [Huang et al. 2018] or EfficientNet [Tan and Le 2020] would introduce additional spatial information and enable stronger feature fusion and improved generalization to diverse acoustic conditions.

In addition, while the current architecture fuses features at a static level, another direction would be dual-stream fusion in the parallel format. By allowing the streams to learn independently before the fusion step, the system would be capable of maintaining unique stream properties and permitting greater interaction among modalities [Dosovitskiy et al. 2020]. The approach would be able to generate richer feature representations and reduce prediction errors even further.

Re-evaluating the integration with ResNet-101, could be another such approach. Through our experiments, we find ResNet-101 was not the optimal model when coupled with RGB-ALMD and would likely benefit from even sophisticated fusion methods—such as attention-based gating [Hu et al. 2018] or modality-aware transformers [Sanh et al. 2019]—to fuse spatial and temporal information better.

Furthermore, the model could be incorporated with additional modalities and diverse data sets. The incorporation of additional inputs such as video, bio-signal inputs, or textual inputs can provide a richer depression assessment workflow and therefore greater overall prediction accuracy [Chen et al. 2024].

Finally, to further validate our model, we shall be experimenting with the AVEC2013 dataset alongside AVEC2014. We shall also look into larger real-life data sets to determine the robustness in varied acoustic conditions and in various populations to ensure the scalability and real-life applicability of the system in clinical environments [Sun et al. 2020].

References

Mohamad Al Jazaery and Guodong Guo. 2018. Video-based depression level analysis by encoding deep spatiotemporal features. IEEE Transactions on Affective Computing 12, 1 (2018), 262–268.

American Psychiatric Association et al. 2000. Diagnostic and statistical manual of mental disorders. Text revision (2000).

N Aswal, SK Singh, and P Kamarapu. 2018. Study on antidepressant drug to cure depression. J Formul Sci Bioavailab 2, 121 (2018), 2577–0543.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).

Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. 1996. Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. Journal of personality assessment 67, 3 (1996), 588–597.

Aaron T Beck, Calvin H Ward, Mock Mendelson, Jeremiah Mock, and John Erbaugh. 1961. An inventory for measuring depression. Archives of general psychiatry 4, 6 (1961), 561–571.

Mattia G. Campana and Franca Delmastro. 2017. Recommender Systems for Online and Mobile Social Networks: A survey. Online Social Networks and Media 3–4 (Oct. 2017), 75–97. https://doi.org/10.1016/j.osnem.2017.10.005

Jie Chen, Ngan Yin Chan, Chun-Tung Li, Joey WY Chan, Yaping Liu, Shirley Xin Li, Steven WH Chau, Kwong Sak Leung, Pheng-Ann Heng, Tatia MC Lee, et al. 2024. Multimodal digital assessment of depression with actigraphy and app in Hong Kong Chinese. Translational Psychiatry 14, 1 (2024), 150.

Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, and Jarek Krajewski. 2015a. Weighted pairwise Gaussian likelihood regression for depression score prediction. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Brisbane, Australia, 4779–4783.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015b. A review of depression and suicide risk assessment using speech analysis. Speech communication 71 (2015), 10–49.

Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, and Jarek Krajewski. 2015c. Relevance vector machine for depression prediction.. In Interspeech. ISCA, Dresden, Germany, 110–114.

Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, James R Williamson, Thomas F Quatieri, and Jarek Krajewski. 2017. Generalized two-stage rank regression framework for depression score prediction from speech. IEEE Transactions on Affective Computing 11, 2 (2017), 272–283.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860 (2019).

Arnab Kumar Das and Ruchira Naskar. 2024. A deep learning model for depression detection based on MFCC and CNN generated spectrogram features. Biomedical Signal Processing and Control 90 (2024), 105898. https://doi.org/10.1016/j.bspc.2023.105898

Wheidima Carneiro De Melo, Eric Granger, and Abdenour Hadid. 2020. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE transactions on affective computing 13, 3 (2020), 1581–1592.

Koen Demyttenaere and Jürgen De Fruyt. 2003. Getting What You Ask For: On the Selectivity of Depression Rating Scales. Psychotherapy and Psychosomatics 72, 2 (2003), 61–70. https://doi.org/10.1159/000068690

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

Yizhuo Dong and Xinyu Yang. 2021. A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing 441 (2021), 279–290.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2009. OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In 2009 3rd international conference on affective computing and intelligent interaction and workshops. IEEE, IEEE, Amsterdam, Netherlands, 1–6.

Huiting Fan, Xingnan Zhang, Yingying Xu, Jiangxiong Fang, Shiqing Zhang, Xiaoming Zhao, and Jun Yu. 2024. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion 104 (2024), 102161.

Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. Biomedical Signal Processing and Control 82 (2023), 104561.

Palmira Faraci and Angela Tirrito. 2013. Fifty years studying the Beck Depression Inventory (BDI) factorial stability without consistent results: A further contribution. Clinical Neuropsychiatry 10 (Jan. 2013), 274–279.

Xiaoyan Fu, Jinming Li, Honghong Liu, Miaomiao Zhang, and Ge Xin. 2022. Audio signal-based depression level prediction combining temporal and spectral features. In 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, IEEE, Montreal, Canada, 359–365.

Henndy Ginting, Gérard Näring, William M. van der Veld, Wilis Srisayekti, and Eni S. Becker. 2013. Validating the Beck Depression Inventory-II in Indonesia's general population and coronary heart disease patients. International Journal of Clinical and Health Psychology 13, 3 (Sept. 2013), 235–242. https://doi.org/10.1016/S1697-2600(13)70028-0

Jaroslav Gottfried, Edita Chvojka, Adam Klocek, Tomas Kratochvil, Petr Palíšek, and Martin Tancoš. 2024. BDI-II: Self-Report and Interview-based Administration Yield the Same Results in Young Adults. Journal of Psychopathology and Behavioral Assessment 46, 3 (Sept. 2024), 851–856. https://doi.org/10.1007/s10862-024-10154-z

Mignote Hailu Gebrie. 2018. An Analysis of Beck Depression Inventory 2nd Edition (BDI-II). Global Journal of Endocrinological Metabolism 2, 3 (July 2018). https://doi.org/10.31031/GJEM.2018.02.000540

MAX Hamilton. 1959. The assessment of anxiety states by rating. British journal of medical psychology (1959).

Lang He and Cui Cao. 2018. Automated depression analysis using convolutional neural networks from speech. Journal of biomedical informatics 83 (2018), 103–111.

Lang He, Jonathan Cheung-Wai Chan, and Zhongmin Wang. 2021. Automatic depression recognition using CNN with attention mechanism from videos. Neurocomputing 422 (2021), 165–175.

Lang He, Dongmei Jiang, and Hichem Sahli. 2018. Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding. IEEE Transactions on Multimedia 21, 6 (2018), 1476–1486.

Lang He, Zheng Li, Prayag Tiwari, Feng Zhu, and Di Wu. 2024. LSCAformer: Long and short-term cross-attention-aware transformer for depression recognition from video sequences. Biomedical Signal Processing and Control 98 (2024), 106767.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 7132–7141.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV] https://arxiv.org/abs/1608.06993

Asim Jan, Hongying Meng, Yona Falinie Binti A Gaus, and Fan Zhang. 2017. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE Transactions on Cognitive and Developmental Systems 10, 3 (2017), 668–680.

WANG Jianwen and SHA Xiao. 2023. Recognition of Depression from Video Frames by using Convolutional Neural Networks. International Journal of Advanced Computer Science & Applications 14, 11 (2023).

Shubhan Kadam, Jay Jani, Aniket Kudtarkar, and Reeta Koshy. 2024a. Speech Emotion Recognition Using Mel-Frequency Cepstral Coefficients & Convolutional Neural Networks. In 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). 1595–1602. https://doi.org/10.1109/IDCIoT59759.2024.10467837

Shubhan Kadam, Jay Jani, Aniket Kudtarkar, and Reeta Koshy. 2024b. Speech emotion recognition using mel-frequency cepstral coefficients & convolutional neural networks. In 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). IEEE, 1595–1602.

Heysem Kaya, Florian Eyben, Albert Ali Salah, and Björn Schuller. 2014. CCA based feature selection with application to continuous depression recognition from acoustic speech features. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, IEEE, Florence, Italy, 3729–3733.

Manuel Lage Cañellas, Constantino Álvarez Casado, Le Nguyen, and Miguel Bordallo López. 2023. Depression recognition from facial videos: Preprocessing and scheduling choices hide the architectural contributions. Electronics Letters 59, 20 (2023), e12992.

Yutong Li, Zhenyu Liu, Li Zhou, Xiaoyan Yuan, Zixuan Shangguan, Xiping Hu, and Bin Hu. 2023. A facial depression recognition method based on hybrid multi-head cross attention network. Frontiers in Neuroscience 17 (2023), 1188434.

Zhenyu Liu, Xiaoyan Yuan, Yutong Li, Zixuan Shangguan, Li Zhou, and Bin Hu. 2023. PRA-Net: Part-and-Relation Attention Network for depression recognition from facial expression. Computers in Biology and Medicine 157 (2023), 106589.

Donovan Maust, Mario Cristancho, Laurie Gray, Susan Rushing, Chris Tjoa, and Michael E. Thase. 2012. Psychiatric rating scales. Handbook of Clinical Neurology (Jan. 2012), 227–237. https://doi.org/10.1016/b978-0-444-52002-9.00013-9

E. McElroy, P. Casey, G. Adamson, P. Filippopoulos, and M. Shevlin. 2018. A comprehensive analysis of the factor structure of the Beck Depression Inventory-II in a sample of outpatients with adjustment disorder and depressive episode. Irish Journal of Psychological Medicine 35, 1 (March 2018), 53–61. https://doi.org/10.1017/ipm.2017.52

Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed Ai-Shuraifi, and Yunhong Wang. 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. 21–30.

Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. 2019. Speech emotion recognition from 3D log-mel spectrograms with deep learning network. IEEE access 7 (2019), 125868–125881.

Stuart A. Montgomery and Marie Åsberg. 1979. A New Depression Scale Designed to be Sensitive to Change. The British Journal of Psychiatry 134, 4 (April 1979), 382–389. https://doi.org/10.1192/bjp.134.4.382

Mingyue Niu, Jianhua Tao, Bin Liu, and Cunhang Fan. 2019. Automatic depression level detection via lp-norm pooling. Proc. INTERSPEECH, Graz, Austria -, September (2019), 4559–4563.

Mingyue Niu, Jianhua Tao, Bin Liu, Jian Huang, and Zheng Lian. 2020. Multimodal spatiotemporal representation for automatic depression level detection. IEEE transactions on affective computing 14, 1 (2020), 294–307.

Zainal Nz. 2014. Research in Depression. The Malaysian Journal of Psychiatry 23, 2 (2014), 1–2.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence 24, 7 (2002), 971–987.

Janez Perš, Vildana Sulić, Matej Kristan, Matej Perše, Klemen Polanec, and Stanislav Kovačič. 2010. Histograms of optical flow for efficient representation of body motion. Pattern Recognition Letters 31, 11 (2010), 1369–1376.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019).

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, Patient Health Questionnaire Primary Care Study Group, Patient Health Questionnaire Primary Care Study Group, et al. 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Jama 282, 18 (1999), 1737–1744.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984 (2020).

Pär Svanborg and Marie Åsberg. 2001. A comparison between the Beck Depression Inventory (BDI) and the self-rating version of the Montgomery Åsberg Depression Rating Scale (MADRS). Journal of Affective Disorders 64, 2–3 (May 2001), 203–216. https://doi.org/10.1016/s0165-0327(00)00242-1

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-V4, Inception-ResNet and the impact of residual connections on learning. https://arxiv.org/abs/1602.07261v2

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019).

Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG] https://arxiv.org/abs/1905.11946

Xiaoyang Tan and Bill Triggs. 2010. Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE transactions on image processing 19, 6 (2010), 1635–1650.

Michael E Tipping. 2001. Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research 1, Jun (2001), 211–244.

Michael E Tipping. 2003. Bayesian inference: An introduction to principles and practice in machine learning. In Summer School on Machine Learning. Springer, Oberammergau, Germany, 41–62.

Md Azher Uddin, Joolekha Bibi Joolee, Aftab Alam, and Young-Koo Lee. 2017. Human Action Recognition Using Adaptive Local Motion Descriptor in Spark. IEEE Access 5 (2017), 21157–21167. https://doi.org/10.1109/ACCESS.2017.2759225

Md Azher Uddin, Joolekha Bibi Joolee, and Young-Koo Lee. 2020. Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. IEEE Transactions on Affective Computing 13, 2 (2020), 864–870.

Md Azher Uddin, Joolekha Bibi Joolee, and Kyung-Ah Sohn. 2022. Deep multi-modal network based automated depression severity estimation. IEEE transactions on affective computing 14, 3 (2022), 2153–2167.

Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. Avec 2014: 3d dimensional affect and depression recognition challenge. In Proceedings of the 4th international workshop on audio/visual emotion challenge. 3–10.

Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, Barcelona, Spain, 3–10.

A Vaswani. 2017. Attention is all you need. Advances in Neural Information Processing Systems (2017).

Kannan Venkataramanan and Haresh Rengaraj Rajamohan. 2019. Emotion recognition from speech. arXiv preprint arXiv:1912.10458 (2019).

World Health Organization: WHO. 2020. Depression. https://www.who.int/india/health-topics/depression

World Health Organization: WHO. 2022. COVID-19 Pandemic Triggers 25% Increase in Prevalence Of Anxiety and Depression Worldwide. https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

World Health Organization: WHO. 2023. Depressive disorder (depression). https://www.who.int/news-room/fact-sheets/detail/depression

James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta. 2013. Vocal biomarkers of depression based on motor incoordination. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, Barcelona, Spain, 41–48.

Faming Yin, Jing Du, Xinzhou Xu, and Li Zhao. 2023. Depression detection in speech using transformer and parallel convolutional neural networks. Electronics 12, 2 (2023), 328.

Ziping Zhao, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, and Björn Schuller. 2020. Interspeech 2020. In Hybrid network feature extraction for depression assessment from speech. ISCA, Shanghai, China, 4956–4960.

Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. 2020. Visually Interpretable Representation Learning for Depression Recognition from Facial Images. IEEE Transactions on Affective Computing 11, 3 (2020), 542–552. https://doi.org/10.1109/TAFFC.2018.2828819

Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. 2017. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. IEEE Transactions on Affective Computing 9, 4 (2017), 578–584.

A Project Management

Organization and planning are important aspects for the management of a project to stay on track with its goals and tasks, to avoid the risk of straying off of deadlines, leading to an incomplete or rushed project. The following section will cover the scope and plan of the project along with the milestones that would guide the project's execution, followed by a set of possible risks that could occur during the making of this project as well as the mitigating strategies related to the risks. It also defines the scope, objectives, and milestones that guide the project's execution.

A.1 Project Scope

The research aims to address current gaps in speech-based depression detection, particularly the limited use of dynamic texture descriptors along with hand-crafted dynamic descriptors. This is achieved by creating an end-to-end novel framework that integrates advanced feature extraction methods such as Adaptive Local Motion Descriptor (ALMD) and ResNet-101. These extracted features are then fused and evaluated through a dual-stream Transformer model architecture to measure the BDI-II score using the AVEC-2014 dataset.

Key tasks with the project scope include:

- Developing a pipeline to segment, pre-process, and analyze audio data.
- Applying the feature extraction methods, specifically ALMD and ResNet-101.
- Employing a dual-stream Transformer model to predict the BDI-II score.
- Evaluating the model's performance using RMSE and MAE metrics.
- Comparing the model against existing models, benchmark models, and state-of-the-art models.
- Exploring the professional, legal, ethical, and social considerations in the use of sensitive data.

A.2 Project Deliverables

The project is divided into 4 crucial parts with definite deadlines:

A.2.1 Deliverable 1 Report

The first deliverable contains a brief overview of the study. This includes:

(1) Deciding a topic.
(2) Researching on the selected topic.
(3) Selecting a dataset to work with.
(4) Creating a detailed literature review.
(5) Finding the gaps in the literature.
(6) Proposing a model that could address the gap(s) found.
(7) Understanding the model and its requirements.
(8) Selecting evaluation strategies to test the performance of the model.
(9) Understanding the professional, legal, ethical and social considerations.
(10) Creating a project plan for the remaining deliverables.
(11) Document all the above into a brief report.

A.2.2 Final Dissertation Report

This deliverable is an expansion and implementation of the initial report which would include the following:

(1) Understand the model chosen more in-depth.
(2) Begin implementing the model.
(3) Create the hyper-parameters.
(4) Use the required pre-trained models.
(5) Build the remaining models if necessary.
(6) Train the dataset on the model created.
(7) Test the dataset using the evaluation metrics.
(8) Find the best hyper-parameters.
(9) Conclude the result by comparing it with existing models.
(10) Document all findings and processes into a comprehensive report.
(11) Find the possible improvements that could be implemented in the future to complement the model.

A.2.3 Code Submission

Once the implementation is completed, the code for the end-to-end framework is required to be submitted along with all its dependencies. This could also include instructions to replicate for future work.

A.2.4 Poster and Mini-Viva

Once the final dissertation has been submitted, summarize the research findings and framework architecture implemented through an oral presentation. Ensure that the aims, objectives as well as the results and conclusion is effectively addressed.

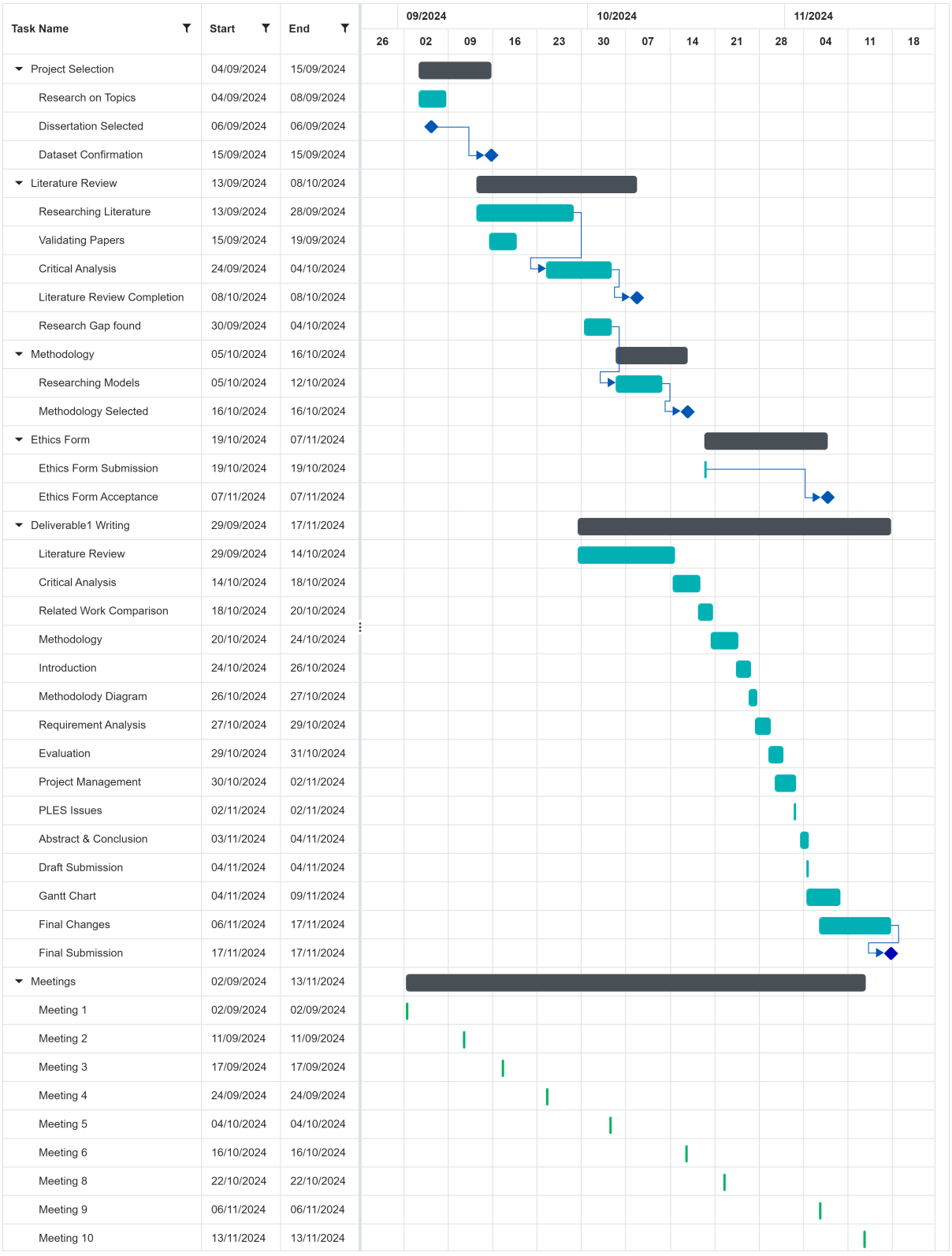
A.3 Project Plan

The project timeline is structured using a Gantt chart to track progress and ensure timely completion of tasks as milestones. The plan is divided into a span of 2 semesters:

(1) Semester 1 which includes Deliverable 1 Report (Appendix A.2.1)
(2) Semester 2 which includes the remaining deliverables consisting of Final Dissertation Report (Appendix A.2.2), Code Submission (Appendix A.2.3) followed by the Poster and Mini-Viva (Appendix A.2.4).

Detailed Gantt charts outlining the progression of this project as described above are outlined is shown below in Figure 17 and Figure 18. This provides a visual representation of the tasks and milestones achieved and to be achieved, ensuring a structured approach to the fulfillment of the project aims and objectives.

Following the gantt charts, we shall look into the possible risks our project could face in Appendix A.4.



Task Name	Start	End
Project Selection	04/09/2024	15/09/2024
Research on Topics	04/09/2024	08/09/2024
Dissertation Selected	06/09/2024	06/09/2024
Dataset Confirmation	15/09/2024	15/09/2024
Literature Review	13/09/2024	08/10/2024
Researching Literature	13/09/2024	28/09/2024
Validating Papers	15/09/2024	19/09/2024
Critical Analysis	24/09/2024	04/10/2024
Literature Review Completion	08/10/2024	08/10/2024
Research Gap found	30/09/2024	04/10/2024
Methodology	05/10/2024	16/10/2024
Researching Models	05/10/2024	12/10/2024
Methodology Selected	16/10/2024	16/10/2024
Ethics Form	19/10/2024	07/11/2024
Ethics Form Submission	19/10/2024	19/10/2024
Ethics Form Acceptance	07/11/2024	07/11/2024
Deliverable1 Writing	29/09/2024	17/11/2024
Literature Review	29/09/2024	14/10/2024
Critical Analysis	14/10/2024	18/10/2024
Related Work Comparison	18/10/2024	20/10/2024
Methodology	20/10/2024	24/10/2024
Introduction	24/10/2024	26/10/2024
Methodolody Diagram	26/10/2024	27/10/2024
Requirement Analysis	27/10/2024	29/10/2024
Evaluation	29/10/2024	31/10/2024
Project Management	30/10/2024	02/11/2024
PLES Issues	02/11/2024	02/11/2024
Abstract & Conclusion	03/11/2024	04/11/2024
Draft Submission	04/11/2024	04/11/2024
Gantt Chart	04/11/2024	09/11/2024
Final Changes	06/11/2024	17/11/2024
Final Submission	17/11/2024	17/11/2024
Meetings	02/09/2024	13/11/2024
Meeting 1	02/09/2024	02/09/2024
Meeting 2	11/09/2024	11/09/2024
Meeting 3	17/09/2024	17/09/2024
Meeting 4	24/09/2024	24/09/2024
Meeting 5	04/10/2024	04/10/2024
Meeting 6	16/10/2024	16/10/2024
Meeting 8	22/10/2024	22/10/2024
Meeting 9	06/11/2024	06/11/2024
Meeting 10	13/11/2024	13/11/2024

Fig. 17. Timeline for Semester 1

Task Name	Start	End
Start project timeline	09/01/2025	26/02/2025
Plan data preprocessing	09/01/2025	10/01/2025
Implement audio segmentation	11/01/2025	13/01/2025
Generate multiple split lengths	14/01/2025	15/01/2025
Generate spectrograms	16/01/2025	18/01/2025
Develop ALMD feature extractor	19/01/2025	26/01/2025
Extract ResNet-101 features	27/01/2025	29/01/2025
Optimize ResNet hyperparameters	30/01/2025	31/01/2025
Train Transformer	01/02/2025	06/02/2025
Validate Transformer performance	07/02/2025	08/02/2025
Fine-tune model components	14/02/2025	18/02/2025
Merge ALMD and ResNet features	19/02/2025	20/02/2025
Test feature fusion	21/02/2025	22/02/2025
Evaluate RMSE & MAE	23/02/2025	24/02/2025
Compare with benchmarks	25/02/2025	26/02/2025
Deliverable 2	08/02/2025	25/03/2025
Receive D1 feedback	08/02/2025	08/02/2025
Implement changes	09/02/2025	13/02/2025
Introduction	27/02/2025	28/02/2025
Literature Review	01/03/2025	02/03/2025
Methodology	03/03/2025	07/03/2025
Results	07/03/2025	10/03/2025
Revise dissertation drafts	10/03/2025	23/03/2025
Submit final dissertation	25/03/2025	25/03/2025
Poster & Mini-viva	05/04/2025	04/05/2025
Design poster layout and visuals	05/04/2025	10/04/2025
Add text and finalize poster	10/04/2025	14/04/2025
Submit poster	16/04/2025	16/04/2025
Prepare for mini-viva	19/04/2025	25/04/2025
Conduct mini-viva presentation	29/04/2025	29/04/2025
Present at mini-viva	04/05/2025	04/05/2025

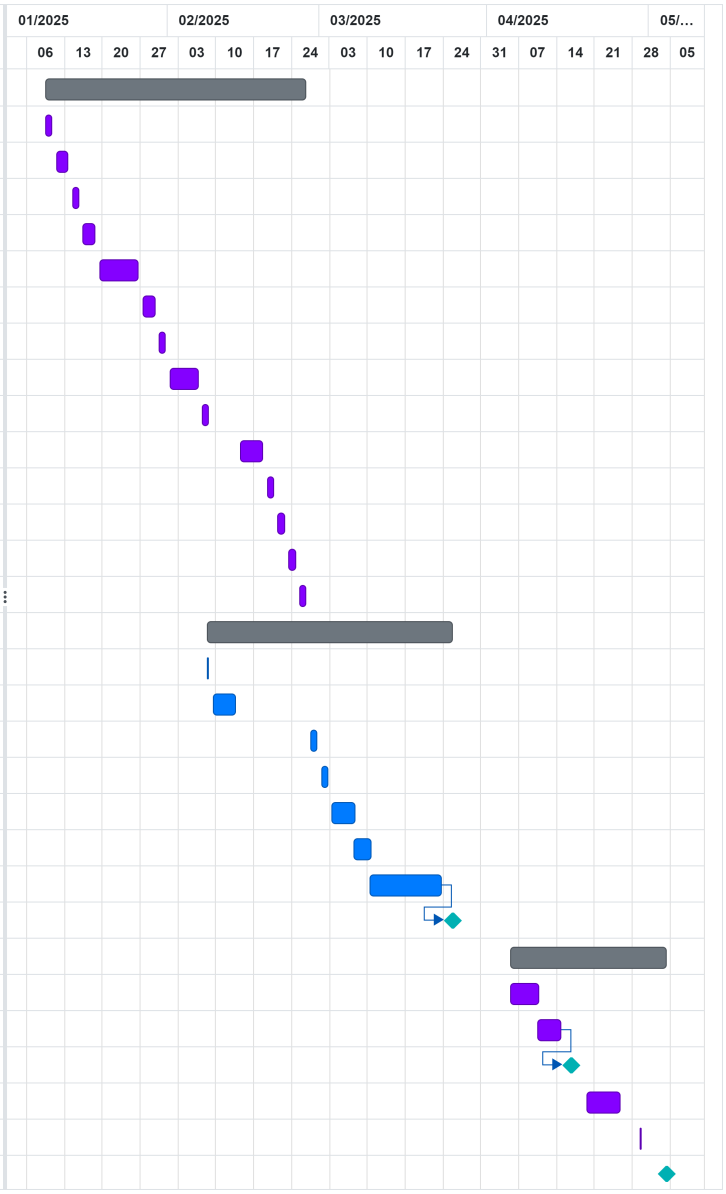


Fig. 18. Timeline for Semester 2

A.4 Risk Analysis

This sub-section talks about the possible risks that could be faced during the project. The risks, along with their level of occurrence as well as the level of negative impact it would cause are mentioned below in the following table Table 8. This is followed by mitigating strategies that could be followed to avoid or overcome the possible risks.

ID	Risk	Likelihood Level	Impact Level
1	Dataset Inadequacy	Low	Medium
2	Model Overfitting	Medium	High
3	Insufficient Computational Resources	Low	High
4	Ethics Form Rejection	Medium	High
5	Difficulty Implementing Transformer	Medium	Medium
6	Data Pre-processing Errors	High	Medium
7	High Memory Usage with Transformers	Medium	High
8	Poor Model Performance	Medium	High
9	Software Bugs	High	Medium
10	Time Management Issues	Medium	High

Table 8. Risk Assessment

A.4.1 Risk Mitigating Strategies

The strategies given below can be directly linked to the risks provided in Table 8.

(1) **Dataset Inadequacy:** If the AVEC-2014 dataset proves to be insufficient, apply audio data augmentation to increase the number of samples to be tested and trained on. Ensure proper data cleaning and pre-processing is done to maximize the dataset's utility.
(2) **Model Overfitting:** Implement techniques like dropout, early stopping, and data augmentation. Ensure that the model is using the provided test set for a balanced distribution of the data (refer to Section 5.1), ensuring no bias.
(3) **Insufficient Computational Resources:** To avoid leakage of data, using cloud computing services like Google Colab (Pro version), Amazon Web Services (AWS), etc, shall not be used. Instead other computational intensive processes shall be looked into. For

example, results of the pre-processed data shall be saved to the disk for reuse, reducing intensive repetitive computation.

(4) **Ethics Form Rejection:** Follow up with the ethics committee to understand the reasons of rejection, talk to the supervisor to confirm no ethical rules are broken.
(5) **Difficulty Implementing Transformer:** Divide the Transformer implementation into steps, such as pre-training, fine-tuning, and evaluation. .
(6) **Data Pre-processing Errors:** Validating pre-processing results by visualizing intermediate outputs such as spectrograms generated could reduce pre-processing errors.
(7) **High Memory Usage with Transformers:** Techniques like gradient check-pointing could be implemented to save memory during back-propagation. Reducing the model size or experimenting with different hyper-parameters during training could reduce memory usage accordingly.
(8) **Poor Model Performance:** Test the model's performance with smaller configurations before scaling. Use thorough hyper-parameter.
(9) **Software Bugs:** Develop in modular increments and use version controls such as GitHub to track code changes and resolve issues efficiently.
(10) **Time Management Issues:** Allocate buffer time in the gantt charts for unexpected delays and prioritize high-impact tasks early on.

B Professional, Legal, Ethical, and Social Considerations

This study adheres to principles outlined by the institution and governing research bodies ensuring integrity and value to society. The following subsections addresses the professional, legal, ethical, and social considerations in the methodology and execution of this project.

B.1 Professional Considerations

The research maintains strong professional standards for data collection, analysis, and reporting, ensuring transparency. Models and methodologies are aligned with current state-of-arts in AI and mental health research. The development of these methodologies and models rely on open-source software and libraries such as python, Keras, TensorFlow, OpenCV, Scikit-Learn and others. Any utilized software libraries are referenced and rightly licensed within their terms of use. All of these ensure that the project is well in line with standards set by the British Computing Society (BCS) code of conduct for professional integrity in software utilization, code development, and documentation.

B.2 Legal Considerations

The research complies with legal requirements, including data protection laws under the regulations of the United Arab Emirates and General Data Protection Regulation (GDPR). The dataset AVEC-2014, while restricted in access, is used solely for academic and research purposes under the guidance of the supervising professor. These procedures ensure compliance with national and international regulations governing sensitive health data in research.

B.3 Ethical Considerations

From an ethical perspective, the research ensures the handling of data AVEC-2014, is in no way traceable to any participant or their personal information. The participants are also well informed that this data collected in AVEC-2014 shall be used for research purposes. An elaborate of the data collection is mentioned in Section 5.1. The model was well-trained and tested, while also ensuring no leakage. The research is done in data compliance with the standards and data-sharing agreements of the Ethics Committee in Heriot-Watt University.

B.4 Social Considerations

This project's objective is to build an automated speech-based depression severity estimation system, with a purely technical focus that does not involve the processing of controversial data. This is ensured as the AVEC-2014 is widely used and recognized in the line of research. The system is designed to contribute positively to early detection in mental health without introducing any social or ethical concerns. Each segment of the recording is treated without any bias. Throughout the study, the project avoids any activities that could lead to any ethical dilemmas or social implications, ensuring a responsible approach to depression assessment.

The study does not involve external human participation as the performance and accuracy of the model(s) is solely based on the regression based evaluation metrics.