

Multimodal Depression Detection from Facial Emotion Dynamics and Acoustic Features Using Ensemble Learning

by

Md Farhan Mahtab Sohan

22301272

Abu Ferdous Shawon

22301320

Fariha Rahman

22301307

Maimuna Morshed

22301319

Aysha Siddiqua Tithi

22301667

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md Farhan Mahtab Sohan

22301272

Abu Ferdous Shawon

22301320

Fariha Rahman

22301307

Maimuna Morshed

22301319

Aysha Siddiqua Tithi

22301667

Approval

The thesis/project titled “Multimodal Depression Detection from Facial Emotion Dynamics and Acoustic Features Using Ensemble Learning” submitted by

1. Abu Ferdous Shawon (22301320)
2. Md Farhan Mahtab Sohan (22301272)
3. Fariha Rahman (22301307)
4. Maimuna Morshed (22301319)
5. Aysha Siddiqua Tithi (22301667)

Examining Committee:

Supervisor:
(Member)

Md. Tawhid Anwar
Senior Lecturer
Department of Computer Science and
Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Rafeed Rahman
Senior Lecturer
Department of Computer Science and
Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and
Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson And Associate Professor
Department of Computer Science and
Engineering
Brac University

Abstract

Depression is a psychiatric disorder that is significantly underdiagnosed in the world, particularly in low-resource countries. This thesis introduces a two-step multimodal framework of automated depression recognition using facial emotion detection. The first phase consists of the systematical training of nine deep learning models CoAtNet-2, ViT-Base/16, Swin-Small, ConvNeXt-Base, MaxViT-T, EfficientNet-B7, ConvNeXt-Small, ResNet-50, and VGG16 on a carefully cleaned and preprocessed OAHEGA dataset; the top five of them are jointly trained into a weighted soft-voting ensemble, the FER model, with a test accuracy of 93.13. The second step uses the frozen FER ensemble on the D-Vlog depression data, producing per-frame emotion probability sequences that are concatenated with the previously extracted acoustic features on 667 videos that can be used. Using these modalities, a 755 dimensional multi-modal feature representation (250 visual, 500 acoustic, 5 cross-modal features) is created and narrowed down to 195 features using a two-step Mutual Information and RFECV selection pipeline. A number of classifier architectures are compared, such as single base learners, recurrent sequence models and heterogeneous ensembles; a stacking ensemble (MLP, CatBoost, and Random Forest with a Logistic Regression meta-learner) is chosen as the final model, with a test accuracy of 71.81%, and precision of 0.7529 on the held-out fold. The main contributions are a systematic multi-architecture FER comparison, a hybrid ensemble with state-of-the-art accuracy on OAHEGA, a multimodal depression detection pipeline which is the first to use a purpose-built FER ensemble as a visual backbone, along with audio features, and which can be run on standard consumer hardware to enable scalable, non-invasive screening.

Keywords: depression detection, facial emotion recognition, hybrid ensemble, multimodal fusion, D-Vlog, OAHEGA, deep learning, LSTM, transformer

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	iv
Dedication	v
Acknowledgment	v
Table of Contents	v
List of Figures	viii
List of Tables	ix
Nomenclature	ix
1 Introduction	1
1.1 Background	1
1.2 Motivation	2
1.3 Problem Statement	2
1.4 Objective	3
1.5 Methodology in Brief	3
1.6 Scope and Challenges	6
2 Literature Review	8
2.1 Preliminaries	8

2.2	Review of Existing Research	8
2.3	Summary of Key Findings	17
3	Requirements, Impacts and Constraints	19
3.1	Final Specifications and Requirements	19
3.1.1	Software and Hardware Requirements	19
3.1.2	Data Requirements	20
3.1.3	Functional Requirements	21
3.2	Societal Impact	21
3.3	Environmental Impact	22
3.4	Ethical Issues	22
3.5	Standards	22
3.6	Project Management Plan	23
3.7	Risk Management	23
3.8	Economic Analysis	23
4	Proposed Methodology	24
4.1	Design Process or Methodology Overview	24
4.2	Preliminary Design and Model Specification	26
4.2.1	FER Model Alternatives	26
4.2.2	Preliminary Design and Model Specification	34
4.3	Data Collection	38
4.3.1	OAHEGA Dataset	39
4.3.2	D-Vlog Dataset	42
4.4	Implementation of Selected Design	46
4.4.1	Hybrid FER Model Implementation	47
4.4.2	Depression Classifier Implementation	49
5	Result Analysis	53
5.1	Performance Evaluation	53
5.1.1	FER Model Performance Evaluation	53
5.1.2	Depression Classifier Performance Evaluation	54
5.2	Analysis of Design Solutions	57
5.2.1	FER Model Analysis	57

5.2.2	Depression Classifier Analysis	61
5.3	Final Design Adjustments	64
5.3.1	FER Design Adjustments	64
5.3.2	Depression Classifier Design Adjustments	65
5.4	Statistical Analysis	67
5.4.1	FER Statistical Analysis	67
5.4.2	Depression Classifier Statistical Analysis	68
5.5	Comparisons and Relationships	70
5.5.1	FER Comparisons	70
5.5.2	Depression Classifier Comparisons	74
5.6	Discussions	78
5.6.1	FER Discussions	78
5.6.2	Depression Classifier Discussions	81
6	Conclusion	84
6.1	Summary of Findings	84
6.2	Contributions to the Field	85
6.3	Recommendations for Future Work	86

List of Figures

1.1	End-to-end multimodal depression detection pipeline.	5
4.1	End-to-end pipeline diagram	24
4.2	FER ensemble architecture diagram.	33
4.3	Scatter plot - model parameters vs test accuracy for all FER models. . . .	34
4.4	Example of an original facial image and its corresponding preprocessed image.	42
4.5	Class Distribution per Split	45
4.6	Four consecutive frames extracted from a D-Vlog video at a sampling rate of one frame per second.	46
4.7	Class Distribution Across Dataset Processing Stages	47
4.8	Training and Validation accuracy curves of five essemble members. . . .	47
4.9	Training and Validation loss curves of five essemble members.	48
4.10	Feature aggregation diagram	51
5.1	Hybrid FER ensemble confusion matrix heatmap.	54
5.2	Confusion Matrix: Stacking (MLP + CatBoost + RF → LR) on Test Fold	56
5.3	FER models ranked by test accuracy from lowest to highest.	58
5.4	NS errors per model.	58
5.5	Per-class F1 scores across all FER models.	59
5.6	ROC-AUC Curves Final Test Evaluation (4 Models).	69
5.7	Final Test Evaluation Model Comparison	70

List of Tables

3.1	Software and hardware requirements.	20
4.1	All FER models comparison	33
4.2	Ensemble Configurations Comparison (Validation Set)	37
4.3	All Classifier Alternatives Comparison (Validation Set)	38
4.4	OAHEGA Cleaning Summary	40
4.5	- OAHEGA Class Distribution	41
4.6	D-VLOG Video Counts by Fold and Class	43
4.7	D-VLOG Frame Statistics by Fold and Class (703 Usable Videos)	44
4.8	Final D-VLOG Dataset Split (667 Videos)	45
5.1	Hybrid FER Ensemble Test Metrics (5.1.1)	54
5.2	All Four Ensemble Configurations Final Test Results	55
5.3	Selected Model: Stacking (MLP + CatBoost + RF → LR) Detailed Test Metrics	56
5.4	FER Model Comparative Analysis Summary	57
5.5	Iterative depression-classifier selection based on validation results.	62
5.6	Final test evaluation and comparison of depression-classifier ensemble models.	62
5.7	ConvNeXt-Base Run 1 vs Run 2 Comparison	65
5.8	McNemar’s Test: 5-Model Ensemble vs ConvNeXt-Small	68
5.9	AUC-ROC Comparison: All Ensemble Configurations	69
5.10	Generalization Gap Analysis: Validation vs Test Accuracy	70
5.11	Comparison with existing works on the OAHEGA dataset.	72
5.12	Comparison of the proposed FER model with existing facial-expression-recognition studies.	73
5.13	Comparison with published D-Vlog works.	75

Chapter 1

Introduction

1.1 Background

Depression is one of the most common and severe mental health conditions in global society, with over 280 million individuals affected in the various age groups, cultures, and socioeconomic statuses. Unlike short-term sadness, clinical depression is characterised by persistent emotional disabilities, including protracted hopelessness, exhaustion, lack of interest, social isolation, and objective disruption of facial expression, which is an impairment of normal functioning and, in severe instances, an extremely high probability of self-harm. In some areas of the world, like South Asia, where mental health awareness is still low and the ratio of qualified clinicians to population is alarmingly low, a significant part of people with depression remain undiagnosed over years, thus considerably deteriorating clinical outcomes and quality of life.

Developments in computer vision and deep learning have created a non-invasive solution to assist with early screening of depression by analysing facial expressions. A complex range of muscle activations controls human facial behaviour and indicates underlying emotional and cognitive states with incredible accuracy. Delicate alterations in the eyes, eyebrows, perioral area, and overall facial movements – less expressiveness, negative mood persistence, and diminished emotion fluctuation – are markers of depression that are clinically recorded and remain constant over videotapes. The automated systems capable of consistently identifying and measuring these patterns may provide an accessible, scalable and privacy-preserving substitute for conventional clinical assessment devices.

Facial Emotion Recognition (FER) has been developed as a core research domain in the field of affective computing, and the aim of this study is to categorise facial expressions into specific types of emotions. Deep learning architectures (initially convolutional neural networks and more recently vision transformers and hybrid models) have led to significant increases in FER accuracy in the last decade. Nevertheless, the performance of the laboratory dataset under controlled conditions might not necessarily be related to the robustness in the real world where the facial images can be subjected to changes in illumination, motion blur, partial occlusions, and demographic variations. These issues render strict data cleaning, preprocessing, and model diversity the key issues of any realistic FER implementation.

The relationship between the patterns of facial emotions and depression is quite solid

in clinical psychology. Depressed people have typical emotional patterns like low affect (reduced positive affect), diminished or flattened expression and a preference to stay in a neutral or negative mood longer. These are essentially time-varying phenomena which are visible in an array of video frames but not in a single image. To exploit this temporal signal, it is necessary to have a pipeline which initially enhances robust per-frame emotion recognition and then indicates emotion dynamics across time to indicate a depressive state. The two stages are dealt with in this thesis in a single end-to-end framework.

1.2 Motivation

The research behind this study is based on two overlapping gaps. No single deep learning model can perform well on all real-world face datasets – convolutional networks, vision transformers, and hybrid models have complementary facets of facial structure, and most literature assesses these models individually. An organised multi-architecture comparison and subsequent principled ensemble building provides a precise way to more resilient emotion recognition. On the depression detection side, not many systems combine a purpose-built, high-accuracy FER model with multimodal audio-visual fusion of features with multimodal audio-visual features; many use raw facial landmarks or generic visual descriptors, reducing the clinical interpretability of the visual stream. Depression is expressed both in face and voice-based behaviour such as decreased expressiveness, dulled affect, diminished prosody, and diminished speech energy, and a system that represents both indicators in a time-organised fashion fits much better to the clinical reality of the disorder. The non-invasive system that can be used in low-resource environments like Bangladesh, where mental health practitioners are hardly available, has a specific potential to reach underserved populations in terms of screening.

1.3 Problem Statement

Although there has been a lot of progress in facial emotion recognition (FER) and computational depression, there are still a number of critical limitations. The majority of FER experiments compare a single or two architectures, so the relative performance of convolutional networks, vision transformers, and hybrid architectures has not been characterised on the same dataset and preprocessing pipeline. High-performance FER systems are also seldom incorporated into the depression detection pipelines in a principled manner – current research typically uses raw face landmarks or ready-made descriptors instead of semantically enhanced, purpose-designed FER models, and most either use single features alone or combine them without taking advantage of the temporal dynamics of both input streams, reducing rich sequences to simplistic feature vectors which lose clinically interesting dynamic patterns. Video data on real-world situations further exacerbates these problems, with quality problems like blur, low light, and partial occlusion that are not well studied in published pipelines that use rigorous, multi-stage preprocessing consistently on both FER training data and depression data.

This thesis addresses these gaps by training and comparing nine architecturally diverse deep learning models in a systematic manner, building a new hybrid FER ensemble

of the best-performing models, and deploying it with audio features in a multimodal temporal depression detection pipeline – trained using a feature engineering method and a sequence-based deep learning method on the D-Vlog dataset.

1.4 Objective

The main goal of this study is to develop a powerful end-to-end multimodal depression-detection system through two major contributions: a high-accuracy hybrid Facial Emotion Recognition (FER) model and a multimodal depression classifier. The specific objectives are as follows:

1. To develop a strict data-cleaning and preprocessing pipeline for the OAHEGA dataset, ensuring that all training images satisfy the required quality, geometric-consistency, and illumination conditions before model training.
2. To train and evaluate nine deep learning models under the same experimental setup to provide a fair and reproducible comparison. The evaluated architectures are CoAtNet-2, ViT-Base/16, Swin-Small, ConvNeXt-Base, MaxViT-T, EfficientNet-B7, ConvNeXt-Small, ResNet-50, and VGG-16.
3. To develop a hybrid FER ensemble using the five most architecturally diverse and highest-performing models and to validate the ensemble through weighted soft voting so that it produces more accurate and reliable per-frame emotion-probability estimates than any individual model.
4. To extract per-frame emotion-probability sequences from the D-Vlog videos using the frozen FER ensemble and to construct a multimodal video-level feature vector. The initial 755-dimensional feature vector, consisting of visual features, pre-extracted acoustic descriptors, and cross-modal synchrony features, is reduced to 195 selected features.
5. To compare test accuracy, macro F1-score, AUC-ROC, and recall balance across seven individual classifiers: Random Forest, MLP, CatBoost, XGBoost, TabNet, Bi-LSTM, and Bi-GRU. Four ensemble configurations are also evaluated on the D-Vlog dataset. The best-performing stacking ensemble, composed of MLP, CatBoost, and Random Forest classifiers, is selected as the final depression classifier.

1.5 Methodology in Brief

In this study, we used an end-to-end multimodal pipeline, as shown in Figure 1.1. In the first stage, a hybrid model named the Facial Emotion Recognition (FER) Model is developed using the OAHEGA dataset. First, a three-check cleaning pipeline is implemented: Laplacian blur detection with a variance threshold of 100, MTCNN face detection with a confidence threshold of 0.95, and five-point geometric landmark validation. This process reduces the dataset to 9,716 valid images.

A five-stage preprocessing pipeline is then applied. It consists of a pre-warp Laplacian quality check, MTCNN-based affine face alignment to a canonical 224×224 resolution, CLAHE illumination normalisation on the LAB lightness channel, dlib 68-point facial ROI masking with feathered edges, and a post-warp Laplacian quality check. This results in 8,392 clean images. The dataset is balanced through offline augmentation to reach approximately 12,500 images across five expression classes by randomly applying 10 expression-safe transformations.

Nine deep learning architectures are then trained and evaluated: CoAtNet-2 (92.44%), ViT-Base/16 (92.28%), Swin-Small (91.91%), ConvNeXt-Small (91.69%), ConvNeXt-Base (91.05%), MaxViT-T (90.36%), EfficientNet-B7 (90.36%), VGG16 (89.35%), and ResNet-50 (87.43%). Most of the models use a custom Neutral-Sad Separation loss to improve the clinically relevant neutral-sad boundary. The five best and most architecturally diverse models are merged into a hybrid weighted soft-voting ensemble called the FER Model. The ensemble achieves a test accuracy of 93.13%, which is 0.69 percentage points higher than the best individual model.

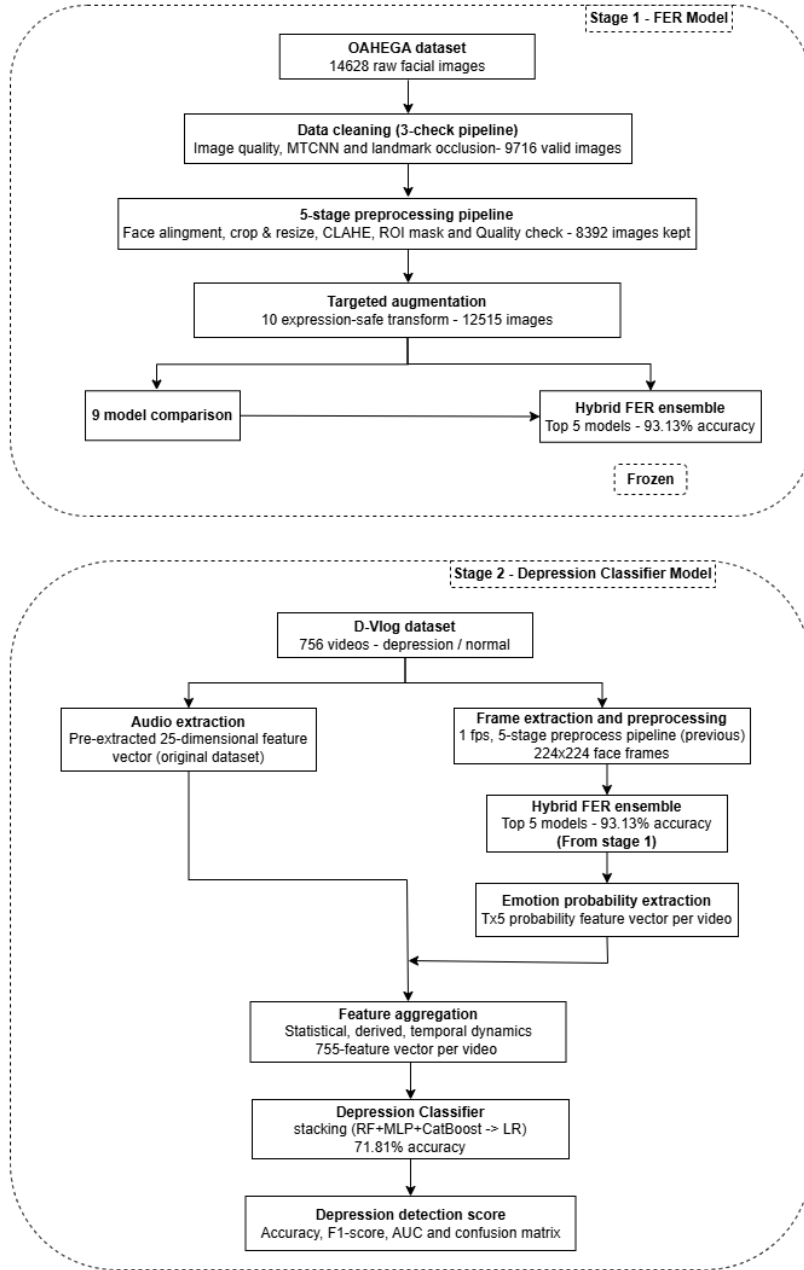


Figure 1.1: End-to-end multimodal depression detection pipeline.

The output of the frozen FER ensemble is then used for the second system component, as a visual feature extractor on the D-Vlog depression video dataset. Because D-Vlog is naturalistic, handheld vlog recordings instead of face images, we adapt the same preprocessing pipeline: The detection thresholds of MTCNN are lowered to tolerate small and partially occluded faces, the blur threshold is decreased from 80 to 30 to deal with the motion blur that is common in handheld video recording, and MediaPipe FaceMesh is used instead of dlib for the facial region of interest masking, because it is better for in-the-wild video frames. From the initial 961 videos, 703 videos give at least one valid

preprocessed frame, resulting in 305,352 clean facial frames from those 703 videos. A quality filter using a minimum ten-frame is then applied, leaving the working set as 667 valid videos to be used for training the classifiers.

The frozen FER ensemble processes each video independently and generates a per frame 5-dimensional emotion probability vector for each video. These are all stacked together along the axis of the valid frames, resulting in a visual sequence of shape $T \times 5$, with T indicating the number of valid frames. The pre-extracted 25-dimensional acoustic descriptor sequence that is provided with the D-Vlog dataset is then combined with this visual sequence, forming a multimodal sequence of shape $T \times 30$. The emotion probability sequences in each video are then aggregated together into a fixed-length 755-dimensional handcrafted feature vector using a detailed temporal aggregation process that includes 250 visual features extracted from the $T \times 5$ emotion sequences, 500 acoustic features extracted from the $T \times 25$ acoustic sequences, and 5 cross-modal synchrony features computed as the Pearson correlation between each emotion probability sequence and the mean acoustic trajectory. A two-stage leakage-free feature selection pipeline is then applied, where all the features are scored using the Mutual Information technique and only 300 most statistically relevant features are retained and further reduced using Recursive Feature Elimination with Cross-Validation (RFECV) with a Random Forest (RF) estimator only on the training fold.

There are two classification tracks assessed. In Track A, five classifiers from the feature space (Random Forest, MLP, CatBoost, XGBoost, TabNet) are applied to the 195 dimensional feature vectors. Track B uses Bidirectional LSTM and Bidirectional GRU sequence models directly on the multimodal sequences, $T \times 30$, without feature aggregation. The best results are obtained by stacking MLP, CatBoost, and Random Forest classifiers and using Logistic Regression as the meta-learner, using out-of-fold cross-validation predictions which eliminates data leakage. This ensemble achieves a test accuracy of 71.81

1.6 Scope and Challenges

This paper is specifically aimed at visual/facial expression-based estimation of FER and depression. The scope encompasses image-level emotion classification with the use of the OAHEGA dataset and video-level depression inference with the use of the D-Vlog dataset. The pipeline uses RGB facial images and video frames. The comparative scope covers nine deep learning architectures tested under the same conditions, and an ensemble of hybrids is the main novelty.

A number of difficulties were faced in this work. The raw OAHEGA data set had a high percentage of unusable images – 33.6% of images were dropped in the cleaning step, mostly because they were blurred, which should be taken into account when designing the pipeline so that augmentation and training are performed on credible data. Another challenge with the D-Vlog dataset was that out of the 961 initially labelled videos, only 756 could be retrieved; after preprocessing, 53 more videos were filtered out as they produced no useful frames, leaving 703 videos to train and evaluate. Vlog-style videos also necessitated modified preprocessing settings to support lower image quality, motion blur, and partial occlusions that are common with handheld videos. Another limitation was computational cost, which meant that training nine architectures independently and building

an ensemble needed, on the one hand, a large amount of GPU resources and hyperparameters. Lastly, distributional consistency between the OAHEGA-trained FER model and the D-Vlog inference frames, achieved by the same preprocessing stages, necessitated a carefully designed pipeline step by step.

Chapter 2

Literature Review

2.1 Preliminaries

Depression is a clinically identified mental illness that is defined by low mood persistence, decreased positive affect, fatigue, and impaired cognitive ability, usually measured with standardized scales, including PHQ-9 and the Beck Depression Inventory (BDI). It has two complementary behavioural markers: facial and vocal. On the face, depression is characterised by decreased smiling, affective flatness, and perpetual neutral or negative facial expression, which is associated with a particular Facial Action Unit (AU) according to the Facial Action Coding System (FACS). It is vocally related to a lowered speech energy, flattened prosody, slower articulation, and alterations in acoustic texture – patterns that can be captured using audio properties obtained by video recordings. Facial Emotion Recognition (FER) systems capitalise on the signal of faces by classifying images into discrete emotion categories with deep learning architectures: early methods used hand-crafted features including HOG and LBP with classical classifiers, but new methods use convolutional neural networks (CNNs), vision transformers (ViT, Swin, CoAtNet, and ConvNeXt), and hybrid ensembles that combine their complementary advantages. Used on video, per-frame FER results can be combined with acoustic data into statistical and temporal representations, reflecting the low expressiveness, negative affect bias, and vocal flatness of depression, and a downstream classifier can learn to distinguish between depressed and non-depressed individuals based on naturalistic video recordings.

2.2 Review of Existing Research

Initial studies in psychology established a close relationship between the poor perception of facial emotions and mental health disorders, especially depression and anxiety. Demenescu et al. (2010) [14] used meta-analysis to investigate how people with depression and anxiety struggle with identifying emotions. Results found that the recognition of facial emotions was lower (moderate-to-high deficits) for depressed adults and, to some degree lower (moderate deficits) for anxious adults ($d = -0.58$, $p < 0.001$ and $d = -0.35$, $p = 0.007$). A widespread mistake was commonly causing happy or neutral faces to be identified as negative.

Based on psychological evidence, later researchers used computational methods to objectively examine facial expressions related to depression. Girard et al. (2013) [20] analyzed how major depressive disorder affects facial expression by using manual FACS coding and also an automated system that applies Active Appearance Models and Support Vector Machines. They found, on analysis of 18 patients, that mild depression was marked by more smiling (AU 12) and less contempt (AU 14), while in severe depression there were fewer smiles and more signs of contempt. There was almost complete agreement ($ICC > 0.85$) between the scores given by the automatic system and by manual count. Beyond observable facial expressions, extending facial analysis, Mustafa et al. (2017) [45] suggested a new method to find depression by calculating heart rate from facial videos rather than using common sensors. To find physiological signals, they used facial landmark tracking and Gabor filters on the subtle color changes on the skin during emotion-triggering movies. Using heart rate changes, an SVM model could correctly classify depression. There was a big difference in heart rate change between the healthy group (4.03 bpm) and the depressed group (1.62 bpm) and the classifier correctly distinguished them in all cases using a small amount of data.

With the advancement of machine learning methods, scientists started to use supervised learning models to identify depression, anxiety, and stress. Gavrilescu and Vizireanu (2019) [16] designed FDASSNN, an immediate system that uses FACS to analyze facial micro-expressions for detecting depression, anxiety and stress. The accuracy rates for the model were 89.3% (for depression), 79.2% (for anxiety) and 92.1% (for stress). The research study by Ahmed et al. (2020) [3] aims to detect and classify depression and anxiety quickly among people in Bangladesh who are 15–35 years of age. Over 13,600 nationally administered and locally adapted psychological questionnaires are used to assess the effectiveness of several algorithms such as CNN, SVM, LDA, KNN and Linear Regression. Of all the results, CNN showed the highest accuracy with 96.8% for depression and 96% for anxiety. Aslam et al. (2020) [5] suggested a machine learning-based model to identify facial emotions through handcrafted image features and ensemble classifiers. It aimed at determining three emotional conditions: angry, happy, and sad with reference to the Taiwanese Facial Expression Image Database (TFEID). The highest accuracy of 95.27% was obtained in the Random Forest model, which can be considered very robust in small datasets. Nannapaneni, (2019) [46] introduced a classical machine learning-based framework of facial emotion recognition to facilitate efficient human-machine interaction at low costs of computation. Discriminative facial features are extracted using Histogram of Oriented Gradients (HOG) and a K-Nearest Neighbour (KNN) classifier is trained to classify emotions. Experimental results showed that the proposed method achieved an overall accuracy of approximately 95%, with fast execution time. Raut, N. (2018) [48] introduced an image-based facial emotion recognition method based on machine learning, which emphasizes deriving interpretable facial features instead of using deep-learning models. Support Vector Machines (SVM) and Logistic Regression models are then used to classify these features. The experimental findings revealed that the proposed models had classification accuracies of between about 85 %, and 89 % implying a consistent performance.

A number of review papers summarised the progress on the topic of facial emotion-based detection of depression. Ashraf et al. (2020) [4] wrote about the application of machine learning methods to recognise depression using video and image data. Important datasets (AVEC, DAIC, DementiaBank) and models such as SVM, CNN and hybrid models were recognised in the study. Joshi and Kanoongo (2022) [28] conducted an exten-

sive review of emotional artificial intelligence-based methods used to detect depression, especially focusing on the analysis of facial images. The authors also pointed out Convolutional Neural Networks (CNNs) as the most efficient method of learning the intricate facial patterns related to depression, with classification rates of up to 98% - 99% in previous research. Also, multimodal systems (that use facial images and speech and text data) were reviewed, and it was concluded that hybrid systems are highly effective in adding robustness and detection accuracy. The article by Khan (2022) [29] is a valuable addition to the FER literature as it provides an in-depth comparison between the classical ML approaches (e.g., SVM, KNN, PCA, LDA, HOG, and Gabor features) and the contemporary DL-based networks (e.g., CNNs, LSTM-CNN hybrids, attention-based networks, and 3D CNNs). Khan is critical to show that deep learning techniques tend to be more discriminative in recognition accuracy than conventional ones because they can learn discriminative features automatically; nevertheless, they require large annotated datasets, high computing power, and significant training time. The review also mentions enduring issues in FER research, such as sensitivity to pose variation, occlusion, change in illumination, and lack of generalisation across datasets, as well as challenges in detecting subtle or micro-expressions.

Sharma et al. (2023) [50] provided a systematic review of machine learning approaches to real-time emotion recognition in a facial expression. The authors outlined the typical facial emotion recognition pipeline that encompasses face detection, feature extraction and emotion classification and identified issues in practice, including variations in illumination, face masking, ageing and skin tone differences. Their comparative analysis revealed that the majority of machine learning algorithms can attain accuracy rates ranging between 85% and 97%, with CNNs and Random Forest models performing better on average. Within the framework of current research, Gavhale and Gawande (2024) [15] place their study in an influential position as a survey that fills the gap between unimodal systems of emotion recognition and multimodal, recommendation-based systems. Past research has primarily dealt with the analysis of facial expressions or with emotion recognition based on speech but not both, making it less robust in real-world contexts where emotions are complicated and dynamic. The authors review a large variety of datasets, including FER2013, CK+, JAFFE, RAVDESS, SAVEE, and IEMOCAP and reveal that multimodal methods are always more successful than unimodal in accuracy and reliability. The PRISMA-ScR systematic review of facial emotion recognition (FER) systems based on neural networks, by Cîrneanu et al. (2023) [13], identified 155 studies out of 1,170 screened and chosen articles on the topic in SCOPUS and Web of Science databases. The review surveys deep learning models such as CNN, RNN, GAN, and GNN models and concludes that CNN-based models, especially VGG, ResNet, Inception, and AlexNet, are dominant in FER because of their efficient feature extraction, ability to transfer learning, and high classification rates. Although performance is strong, the main shortcomings are the lack of diversity of the dataset, the impossibility to differentiate between emotionally related categories, and the lack of extrapolation to uncontrolled real-world settings.

Going beyond binary classification, recent research aimed at estimating depression severity by looking at facial expressions. Li et al. (2020) [38] introduced DRR DepressionNet, which enhances input preprocessing through histogram equalisation and geometric augmentation and applies Euclidean loss to generate smooth severity scores. Evaluated on the same datasets, it achieved improved RMSEs of 8.08 and 8.01 and MAEs of 6.18 and 6.14, demonstrating its effectiveness as a non-invasive, image-based depres-

sion screening tool. Several studies have proposed deep learning models for estimating depression severity from facial images, notably DepressNet by Zhou et al. (2020) [60], which bypasses manual feature engineering by using a deep regression architecture trained on AVEC 2013 and 2014 datasets, achieving MAEs of 6.20 and 6.21, respectively. This model employs MR-DepressNet, where facial features are processed through multiple sub-networks and fused, and its Depression Activation Maps highlight clinically relevant facial regions linked to depressive symptoms, outperforming many existing visual-based methods. Shang et al. (2021) [49] combine traditional texture features from the quaternion approach with features from a deep convolutional neural network to detect depression in facial photos. The model has two modules: a local quaternion representation module handles the eyes and mouth, while a global deep representation module (based on ResNet18) handles the rest. Using AVEC 2013 and 2014 datasets, the LQGDNet achieved MAE/RMSE scores of 6.08/7.84 and 6.38/8.20, better than many current techniques in the field and with a low number of parameters (0.42M). A new deep learning model to detect depression based on the video of facial expressions, introduced by Melo et al. (2019) [40], has the significant advantage of combining global and local spatio-temporal facial features. The research proposed a dual-channel 3D Convolutional Neural Network (C3D) system that breaks down the entire face into two parts, one the entire face and the other a local eye area, as evidence indicates that eye motions are especially informative about depressive behaviour. Experimental data showed the creation of global and local C3D predictions and 3D-GAP produced the least error in prediction with an RMSE of 8.26 (AVEC2013) and 8.31 (AVEC2014), surpassing many state-of-the-art approaches.

Having realised the weaknesses of unimodal systems, a number of studies embraced multimodal models. Mohan et al. (2021) [43] introduced a new depression detection system that uses facial expressions as well as text sentiment to diagnose depression. LSTM's accuracy reached up to 80% and CNN reached 73.45% accuracy. Overall, this approach gave more precise results when judging the severity of depression at four levels: low, mild, average and high, when compared to the baseline approach. Li et al. (2019) [37] proposed a multimodal deep learning model to automatically estimate severity of depression based on facial video, audio, and text. Facial videos were analyzed to record micro-expressions and eye movement and head movement, whereas audio prosodic and textual embeddings of interview transcripts were analyzed concurrently. A hierarchical network architecture was used to fuse these modalities to predict continuous scores of depression according to the PHQ-8 scale. Tests on the AVEC2019 dataset revealed that the multimodal spatiotemporal model had better performance than the unimodal baselines, in terms of detection of levels of depression. The first transformer-based method of estimating the level of depression was a Multi-Modal Adaptive Fusion Transformer Network proposed by Sun et al. (2021) [52]. The model takes the audio features (MFCC, eGeMAPS) and visual features (AUpose) of the AVEC 2019 DDS dataset, both individually encoded with a Transformer Encoder and Embedding Fully Connected block. Multi-task learning scheme combines performance-based weighted importance of each modality, based on per-modality Concordance Correlation Coefficient (CCC) scores, instead of equal-importance, and jointly optimizes PHQ-8 score regression and 5-class depression severity classification. Assessed on the AVEC 2019 DDS development set, the model reached a CCC of 0.733, which is 6.2% better than the previous state-of-the-art of 0.696 when using audio and video modalities.

A multimodal deep learning architecture to detect depression was suggested by Xu

et al. (2025) [57] to overcome the shortage of the conventional subjective clinical evaluation. The major contribution of the study is the integration of voice and text modalities in a single end-to-end architecture to enhance depression classification and the severity estimation. The model methodologically extracts acoustic and semantic features of speech and transcripts, respectively, with Wav2Vec 2.0 and BERT, respectively. The combination of these features is achieved in a multi-scale convolutional fusion module with adaptive pooling and a Bi-LSTM network, to encode both temporal and contextual information. The framework aids in the binary classification of depression and in the prediction of the continuous PHQ-8 scores. This model was tested using two clinically validated datasets: the Chinese Multimodal Depression Corpus (CMDC) and the DAIC-WOZ dataset. Experimental evidence showed a significant improvement in performance compared to unimodal methods where F1 scores were about 0.97 and the RMSE values decreased significantly. The article by Kumar et al. (2024) [34] is an important addition to the literature on depression detection as it brings together facial expression biomarkers and EEG-based neural biomarkers, which have not been widely studied before. Using ensemble machine learning models, and especially random forest classifiers, the authors show a higher level of performance than deep learning models and at a lower level of computational complexity. Nonetheless, the research recognizes that there are limitations with the size of the dataset, demographic diversity, and basic emotion groups, which can impact generalizability in practice.

The new studies have paid more attention to the real-time depression detection systems. Gilanie et al. (2022) [18] suggested a system that uses CNN and SVM to identify depression in real time using facial micro-expressions. CNN had 99.9% accuracy and SVM had a perfect 100%. This system did better in accuracy and efficiency during deployment compared to well-known models like VGG and AlexNet. In a similar vein, Moving toward real-time applications, Lee and Park (2022) [36] suggest a Fast R-CNN system that looks at facial expression by observing movements of the eyes and lips to detect depressive disorder. The accuracy for eye detection was 97.7% and that for lip detection was 91%. Moghe et al. (2024) [42] created a model that can detect depression as it happens by studying facial and speech emotions. Face expressions were labeled through Haar Cascade and DeepFace, but the audio features (MFCC, chroma, mel) were obtained with Librosa and then checked with an MLP model. Using crossvalidation, the model could correctly identify 77.92% of the cases.

Singh and Srivastava (2025) [51] suggested a real-time depression identification model that uses facial emotion recognition, which will facilitate prompt and distant mental health screening. The FER-2013 dataset was reformatted to accomplish this task by combining seven classes of emotions into two: positive (natural) and negative (depressive), and formed a balanced dataset of 7,087 images per split. Experimental assessment revealed that the proposed model was found to have an overall accuracy of about 94% and was found to do well on training, testing and validation set. Hadjar et al. (2025) [24] introduced TheraSense, a system that uses a convolutional neural network (CNN) to detect facial emotions during mental health teleconsultations. Among the tested models, SSD MobileNet v1 secured the highest accuracy (96.87%) as well as fast inference times (0.3s). 100% of tasks were done in the evaluations and 66.7% of clinicians believed the session quality was improved.

A number of recent publications proved the benefits of dual-stream and hybrid architectures. Choi et al. (2023) [12] developed a 2-stream CNN that uses ResNet and

SENet to forecast depression scores by processing facial images and optical flow. One part deals with features that stay put such as the eyes, while the other follows movement using Farneback optical flow. In addition, multimodal fusion has shown to be highly effective in improving diagnostic accuracy. Similarly, Mahayossanunt et al. (2023) [39] developed their framework, which only relies on facial expressions to spot depression. To handle facial features (including gaze, head pose and AUs) with OpenFace, they came up with a model known as Window Block-LSTM that used intermediate fusion and label smoothing. Working on 474 video samples, the model got 91.67% accuracy and 88.89% F1-score. Key factors of depression, found by integrated gradients, are yaw, gaze and lack of smiling. Poongodai et al. (2025) [47] came up with a real-time facial emotion recognition system that adds a hybrid deep learning architecture that will perform the task of identifying emotions accurately and efficiently. The experiment fuses EfficientNet-B3 and ResNet-50 as feature extractors to extract fine-grained and high-level features of faces, overcoming the challenges associated with illumination, changes in pose, and occlusion. The results of the experiment showed the training accuracy of 87.58 % and the testing accuracy of 82.53%, which showed good generalization results and applicability to real-time human-computer interaction uses.

Kong et al. (2024) [30] introduced a deep learning approach of automatic depression detection with facial images, which aims to offer a quick and non-invasive screening method to aid clinical evaluation. The authors methodologically defined the depression detection issue as a binary classification problem (depressed vs non-depressed) and compared various architectures, such as VGG11, VGG19, ResNet50, Inception-V3, and a fully connected CNN. The results of the experiments revealed that the proposed attention-based CNN was more effective than the baseline models with approximating 98% accuracy, precision, recall and F1-score. These results indicate that deep CNNs have a good learning ability of depression-related facial patterns and have a high possibility of being used in early mental health screening. Lakshmi Bai et al. (2024) [35] suggested a hybrid framework of facial emotion-based detection of early depression, a combination of deep learning and psychological emotion analysis. One of the strengths is the fact that the hybrid dataset was used, which is a combination of FER2013, RAF-DB, and a custom interview-based dataset and contains more than 63, 000 face images to enhance the model diversity and generalizability. Emotions detected are measured with the Losada/Fredrickson positivity ratio to determine levels of depression with time. The model of emotion detection has been experimentally found to have an accuracy of about 88%, which is good in detecting normal and mild depressive conditions but is limited in cases of severe depression. A proposed framework of facial emotion detection by Badhan et al. (2025) [7] is based on the application of Convolutional Neural Networks (CNNs) with fuzzy logic to enhance the perception of unclear and overlapping facial expressions. The experimental findings showed that the proposed customized CNN was better than the traditional CNN models, and had emotion-wise accuracies of between 88% and 96% with significant increases in emotion accuracy of being happy, disgust, and being surprised. Babu et al. (2025) [6] proposed a deep neural network-based facial emotion recognition system with the objective of enhancing human-computer interaction in real time. The performance on experiments showed a very high level of performance of 99.9% training accuracy and 97.8% testing accuracy, which showed a high level of classification.

There are multiple studies that have specifically analyzed the OAHEGA data or helped to comprehend the features of FER datasets. The OAHEGA data set [31] sug-

gested by Kovenko and Shevchuk (2022), is a six-class database of facial expressions (neutral, happy, angry, surprised, sad, and ahogado) with 15,744 images obtained through automated cropping based on the YoloV3 algorithm on YouTube, Instagram, and Affect-Net. In the case of the mobile application pipeline, the MobileNetV1-SSD model was used to detect faces, and the MobileNetV1 was used with transfer learning to recognize emotions, with hyperparameters optimized using Bayesian optimization. Class imbalance weighting, dropout, L2 regularization and image augmentation techniques enhanced generalization to produce an average precision of 0.84 with six classes. Agrawal et al. (2023) [1] introduced a transfer learning-based facial expression recognition system that benchmarked six pre-trained CNN architectures CNN, VGG16, VGG19, AlexNet, ResNet50, and InceptionV3 on the OAHEGA dataset, which consisted of about 15,453 RGB images in six emotion categories: happy, sad, angry, neutral, surprised, and ahogado. AlexNet had the highest training accuracy of 92% and the highest test validation accuracy of 54% with a training loss of 0.1945, compared to all other models. The research validates the practicality of transfer learning in the case of facial expression recognition, but the significant difference in training and validation accuracy indicates that additional research is required in augmentation measures and optimization of hyperparameters. In [9], Bakariya et al. created a real-time framework that combines an emotion recognition CNN-based system with a personalized music recommendation engine. To invoke specific music recommendations, two custom CNNs trained with FER-2013 and OAHEGA data, in conjunction with Haar Cascade face detection, identified 6 emotions in 60 frames of the video. Gupta et al. (2026) [23] introduce ECA-ResNet, an Enhanced Channel-Spatial Attention Residual Network for mental health screening using facial emotion recognition. The architecture proposes a new Convolutional Block Attention Module (CBAM) with a residual learning backbone to learn channel-level and spatial-level discriminative facial features. The training is optimized using cosine annealing scheduling, mixed-precision computation and label smoothing. The model is evaluated on the dataset 6-class OAHEGA and is able to achieve an accuracy of 81% and the macro F1-score of 0.81, which is approximately 20% higher than VGG16, InceptionV3, and DenseNet121 in test accuracy. The model is able to better capture the negative emotional states (anger, sadness, and neutral), emphasizing its potential for early detection of mental health deterioration.

Of all the configurations, Model B3, which was trained on the filtered FER-2013+OAHEGA dataset, had the highest accuracy of 73.02% on a 373-image external test set, and performed better than a number of other FER-2013-based methods. Moon and Bhattacharyya (2024) [44] suggested a two-step virtual agent to detect depression and generate therapeutic responses based on CBT. They have expanded the D-vlog sample to 1,261 videos on YouTube, including more disorder categories (MDD, Bipolar, Anxiety) and demographic heterogeneity. Stage 1 used the Textless Vision-Language Transformer (TVLT), where wav2vec2 acoustic features increased the detection rate by 2.2, and the F1-score was 67.8%, 4.3 higher than baselines. Stage 2 utilized the Mistral-7b Instruct-v0.2 with Chain-of-Thought prompting and a CBT knowledge base to produce therapeutic responses, resulting in a BERT score of 88.7, F1-score of 70.1 in detecting cognitive distortions and 30.9 in multi-label classification across ten types of distortions. In [17], Gaya-Morey et al. performed an extensive benchmark study on 24 facial expression recognition (FER) datasets to assess their usefulness in training deep learning models. The experiment proposed three new measures, Local Similarity, Global Similarity, and Paired Similarity, to measure the training difficulty, cross-dataset generalization, and pairwise transferability. Findings indicated that datasets of large size and collected automatically,

like AffectNet and FER2013, are better generalizable even in the presence of labeling noise, whereas laboratory-controlled datasets have better quality annotations but less variability. It was also found that there were significant demographic imbalances, especially that children and the elderly were underrepresented which will offer evidence-based recommendations on how to construct more equitable and generalizable FER systems.

Yoon et al. (2022) [58] introduced the D-Vlog dataset and demonstrated the Depression Detector model based on cross-attention-based transformer fusion of acoustic features with 68-point dlib facial landmark sequences and achieved a weighted average of 63.50 percent F1-score, which became the first D-Vlog baseline. The Tao et al. (2024) [53] framework, the Spatio-Temporal Squeezed Transformer, is a framework that aims to automatically recognize depression based on social media vlog data, using the ability of the multimodal fusion to capture spatio-temporal data - the weakness in the current transformer-based models. The STST model scored better than seven baselines, such as SVM, BLSTM, TFN, and the Depression Detector, on the D-Vlog dataset (916 clips, 816 YouTube users), with a precision of 72.50%, a recall of 77.67%, a F1-score of 75.00%, and an accuracy of 70.70%. Zhou et al. [59] proposed Time-Aware Attention Multimodal Fusion Network (TAMFN) that is used to detect acoustic and visual depression in social media vlogs [59]. Applied to the D-Vlog dataset (961 videos, 816 individuals), TAMFN scored weighted precisions, recalls, and F1-scores of 66.02%, 66.50%, and 65.82%, which are better than benchmarks such as Depression Detector, TFN, and BLSTM.

Gimeno-Gomez et al. (2024) [19] have suggested a multi-modal temporal model to identify depression in naturalistic, user-generated videos. It uses modality specific encoders, learned modality condition embeddings, and fractional positional embeddings to align audio-visual streams and then uses a transformer encoder in binary classification. The model, when assessed on three benchmarks, has state-of-the-art F1-scores of 0.78 on D-Vlog, 0.67 on DAIC-WOZ, and 0.56 on E-DAIC and is doing better than all the previous methods. Recently, Haque et al. (2025) proposed two high-performing architectures on D-Vlog, MDD-Net and MMFformer. MDD-Net [26], combines modules of acoustic and visual features with a mutual transformer that calculates audio-to-video and video-to-audio attention in both directions, which unidirectional methods miss to grasp the joint depressive behavioral patterns. It performed better than the existing methods, such as TAMFN and STST by up to 17.37% in F1-score on the D-Vlog dataset of 961 YouTube videos. MMFformer [25], continues this trend by a multimodal fusion transformer that fuses residual spatial learning of facial expressions with temporal speech modeling, and combined with late, intermediate, and attention-based approaches, at multiple representational levels. It obtained 13.92% and 7.74% F1-score improvements on D-Vlog and LMVD dataset, respectively, outperforming other methods like EMO-Mamba, Dep-Mamba, and Depressformer.

The study by Bakariya et al. (2024) [8] was a real-time system of three stages where face recognition, facial emotion recognition, and context-sensitive music recommendation were combined. The system uses the Haar cascade classifier of OpenCV to detect faces by converting input to a 48 x 48 grayscale image and then feeds them into a custom 12-layer CNN. It was trained on a combination of FER-2013 and OAHEGA (43,003 training; 8,856 test images) for six emotion labels using a 326464128 kernel structure with dropout regularization and Adam as the optimizer ($\text{lr} = 0.0001$) in 24 epochs. On emotion recognition, a Python-based code chooses at random a song in an emotion-tagged CSV database using pygame. The model had a training accuracy of 94.13% and a val-

idation accuracy of 73.02% compared to baselines like ResNet-18 transfer learning and smoothed deep neural network ensembles and had a test accuracy of 73.2%. The system shows a viable use in the healthcare system, social robotics, and entertainment using AI. The two articles by Bhagyalakshmi (2024) [11] and Gupta (2024) [22] both suggested CNN-based facial emotion recognition systems with real-world applicability, albeit with different goals. Bhagyalakshmi targeted mental health detection by training the individual emotion patterns of happiness, sadness, anger, fear and neutrality by mapping facial emotion patterns to depressive and non-depressive states based on the use of OpenCV-based preprocessing with a total accuracy of 90-95% on a standard labeled dataset of facial images. Gupta, instead, focused on human-computer interaction, by training a demographically diverse CNN on the Kaggle Facial Expression Challenge (FEC) dataset, with the ability to classify seven emotion categories with over 90% accuracy, based on the precision, recall, and F1-score metrics. The two papers use multi-layer CNN models with conventional preprocessing chains such as normalization and data augmentation, and both collectively reveal that FER CNN models behave well in cross-emotion generalization and in real-time settings, be it in clinical mental health inference or interactive computing settings.

Telsang et al. (2024) [55] introduced a CNN-based facial emotion recognition model that builds on the classical emotion recognition by addressing the less prevalently investigated facial expressions, i.e., laughter, crying, shame, and frustration. The most important contribution of the study is that it expands the emotional range of recognition systems of emotions on the face, which are often restricted to simple emotions like happiness or anger. Methodologically, the authors developed a tailor made Convolutional Neural Network (CNN) architecture consisting of various convolution, pooling, batch normalization and dropout layers that can successfully learn discriminative facial features like edges, textures, and patterns of expression. Because of the unavailability of appropriate datasets, a self-made dataset of 1,760 grayscale face images was obtained online with 1,600 images being used to train the model and 160 to test it. Much preprocessing was done to provide consistency such as cropping, resizing, the conversion of grayscale and the normalization. The proposed model was found to have 97.62% and 96.88% ,training and testing accuracy, respectively, indicating good generalization and consistent emotion classification. Bakkialakshmi et al. (2024) [10] proposed a machine learning-based facial gesture recognition system, known as a digital mirror, that will help comprehend human emotions in real-time with high accuracy. The major contribution of the research is a combination of face gesture analysis with various machine learning approaches to improve emotion-sensitive human to computer interaction in fields like healthcare, psychology, and user experience design. The system is methodologically, a complete facial emotion recognition pipeline that includes face detection, preprocessing, feature extraction and classification. It uses Convolutional Neural Networks (CNNs) and the conventional classifiers, including k-Nearest Neighbors (k-NN) and Support Vector Machines (SVMs) and Histogram of Oriented Gradients (HOG) to classify emotions effectively. The model was tested on a dataset that was similar to that of FER-2013 that consisted of grayscale facial images, which were categorized into seven emotions and also tested on real-time web-based experiments. The experimental findings showed that the system has a general emotion recognition accuracy of about 90%, which proves the reliability and practical applicability of the system to real life and real time emotion detection scenarios.

In addition to the key FER and D-Vlog literature, other studies have provided perti-

nent results. Ahmed et al. (2022) [2] looks into 54 studies that take place between 2010 and 2021, where machine learning is used to detect anxiety and depression through social media. It is shown by this study that ML is able to enhance traditional diagnostic methods by providing scalable and efficient mental health monitoring worldwide, in any language. Tasnim et al. (2024) [54] introduces a machine learning system that can detect and judge the level of depression, anxiety and stress in someone only through the sound of their speech. With a multilingual, longitudinal dataset annotated with DASS-21 values, the team developed a CNN using VGG-19 extracted spectrograms, giving RMSEs of 7.09 (depression), 7.69 (anxiety) and 8.40 (stress). Jain and Kumar (2024) [27] explain an up-to-date approach to sensing anxiety by combining ECG, EDA and important information about caffeine, smoking habits, weight and exercise. Combining ECG and EDA led to better performance; there was 96.7% accuracy when using both methods together. With additional context features, the system produced its strongest results of 97.3% accuracy when using GBDT, this demonstrates that using multiple and contextual data types is powerful in anxiety classification. Ku et al. (2024) [33] looked into how well machine learning can predict depression and anxiety when dealing with subjective errors common in psychology surveys. CNN was found to be the most resilient ($\sim 92\%$ robustness) and had the highest prediction accuracy (after noise, $AUC = 0.91$) among all models. A new composite score, combining AUC, recall and Cohen's Kappa, was suggested to judge how stable the classifier is. Mishra et al. (2024) [41] examined how well deep learning methods can be used to recognise facial emotions based on the FER2013 dataset, which consists of greyscale images of faces and is divided into seven basic emotions. They found VGG19 to be the most successful model with an accuracy of 92.5% and an F1 score of 0.89, suggesting it is more effective at capturing fine facial features and is relevant to emotion. Although the research conducted by Goel et al. (2022) [21] illustrates a feasible application of emotion recognition based on facial expressions and represents a popular practice in the given area, it still includes important limitations, which have a significant impact on the quality of the results. IEEE has formally reneged on the publication of the paper on the grounds of poor quality of publication and ambiguous technical explanations, which puts the performance metrics reported by the authors, such as the alleged accuracy of more than 90%, into question. Therefore, it helps to comprehend the existing approaches and implementation trends, but its findings should be viewed with caution in academic research.

2.3 Summary of Key Findings

Facial emotion misclassification, decreased facial expressiveness, affective flatness and negative affect bias are consistent emotional facial deficits in depressed patients that can be measured and serve as an indicator for automated detection, as demonstrated in psychological research. Handcrafted features like HOG, LBP, Action Units derived from FACS with classical classifiers (SVM, KNN, Random Forest) have been used for early computational studies to obtain 85–95% correct classification rate on controlled datasets, and clinical evidence of the value of facial expression.

Traditional approaches were quickly outperformed by CNN, which were able to learn discriminative features directly from image data, and various CNN models such as VGG, ResNet, Inception, and attention-based CNN models have obtained 92-99% accuracy on

challenging datasets like FER2013, CK+, and RAF-DB. The hybrid architectures that integrate CNNs with optical flow, ensemble feature extractors or fuzzy reasoning further increased the robustness under real-life conditions, including occlusion, illumination variations, and pose changes.

In addition to binary classification, image-based regression models (DepressNet, DRR DepressionNet, LQGDNet) estimated continuous PHQ/BDI-aligned severity scores on AVEC 2013/14 with a MAE of 6.08–6.21. The findings showed that there is a correlation between subtle facial appearance and graded clinical severity, rather than depression or its absence.

Research always agrees that a multimodal approach (facial, acoustic, linguistic) is superior to any unimodal approach. On AVEC 2019, transformer-based adaptive fusion models got a CCC of 0.733, and on the D-Vlog benchmark, STST, TAMFN, MDD-Net, and MMFormer demonstrated a gradual improvement in F1-scores (75.00%) with transformer-based cross-attention architecture models outperforming previous baseline models by up to 17%. The new design is Cross-modal attention and multi-stage fusion.

The dataset, OAHEGA, consisted of 6 emotion classes with 15,744 images, was introduced with a MobileNetV1 pipeline with an average precision of 0.84. Benchmarking of transfer-learning models showed a significant training to validation accuracy gap (AlexNet: 92% train, 54% test), suggesting that there are still open generalisation problems to solve. This was at least partially resolved by combined training with FER2013, which resulted in 73.02% on external test sets, which is the target that was used as a baseline in this thesis.

The common challenges reported in the literature are the small number of samples, difficulties in cross-dataset generalisation, the inability to capture time dynamics as static representation and lack of model interpretability. These gaps drive the development of the hybrid entropy-based FER ensemble and structured multimodal pipeline developed in this thesis.

Chapter 3

Requirements, Impacts and Constraints

This chapter describes the technical specifications, societal and environmental impacts, ethical obligations, relevant standards, project management model, risk and risk assessment, and economic evaluation that are linked to the creation of the proposed facial emotion recognition and depression detection system. All these aspects determine the scope of operations and obligations that come with the design and possible implementation of the framework.

3.1 Final Specifications and Requirements

3.1.1 Software and Hardware Requirements

It is fully implemented in Python, with open-source libraries. All model training and inference are performed using PyTorch as the main deep learning framework. Face detection is implemented twice: once during cleaning and once during preprocessing. Both implementations are based on MTCNN from `facenet-pytorch`, which runs on a single PyTorch backend and does not require TensorFlow.

In the facial landmark detection pipeline for ROI masking, `dlib`'s 68-point shape predictor is used. In the D-Vlog pipeline, `MediaPipe FaceLandmarker` with 468 landmarks is used. Both pipelines achieve the same masking quality without requiring any native C++ build dependencies. All geometric transformations, CLAHE-based illumination normalisation, and image input-output operations are performed using `OpenCV`, while Gaussian filtering and mask feathering are performed using `SciPy`.

The authors of the original D-Vlog dataset pre-extracted acoustic features from the data using `OpenSmile` and the parameter set provided by the eGeMAPS project. Therefore, no special acoustic feature extraction tool is required in the main pipeline. Classifier training and evaluation are performed using `scikit-learn`, including `StandardScaler`, `SelectKBest`, `RFECV`, `Random Forest`, `MLP`, and `Logistic Regression`. `XGBoost`, `CatBoost`, and `PyTorch` are also used for gradient-boosting and recurrent models.

The MTCNN model training and inference processes were performed efficiently on a CUDA-enabled GPU, specifically the NVIDIA GeForce RTX 4080 SUPER with 16 GB of VRAM. The complete visual preprocessing stage required approximately 317 minutes.

Table 3.1 summarises the full specifications.

Category	Component	Specification / Version
Programming Language	Python	3.9 or above
Deep Learning Framework	PyTorch	2.0 or above
Face Detection	facenet-pytorch (MTCNN)	2.5 or above
Landmark Detection (OAHEGA)	dlib	19.24 or above
Landmark Detection (D-Vlog)	MediaPipe FaceLandmarker	Latest stable
Image Processing	OpenCV	4.7 or above
Scientific Computing	SciPy / NumPy	Latest stable
ML / Feature Selection	scikit-learn	Latest stable
Gradient Boosting	XGBoost / CatBoost	Latest stable
Data Handling	pandas	Latest stable
Visualisation	Matplotlib / Seaborn	Latest stable
Development Environment	Jupyter Notebook / VS Code	Any current version
GPU Acceleration	NVIDIA RTX 4080 SUPER (16 GB)	CUDA 11.x or above
Operating System	Windows / Linux	64-bit
FER Input Resolution	Facial images / frames	224×224 pixels (RGB)
FER Emotion Classes	Dataset labels	Angry, Happy, Neutral, Sad, Surprise
OAHEGA Training Set	Post-augmentation	Approximately 12,515 images (balanced)
D-Vlog Usable Videos	After preprocessing	703 videos (305,352 frames)
Classifier Development Set	After minimum-frame filter	667 videos
Full Feature Vector	Per video	755 features (250 visual + 500 acoustic + 5 cross-modal)
Selected Feature Vector	After MI + RFECV	195 features

Table 3.1: Software and hardware requirements.

3.1.2 Data Requirements

The system requires two datasets. The OAHEGA Emotion Recognition Dataset contains labelled facial images belonging to five emotion classes: angry, happy, neutral, sad, and surprise. The clinically irrelevant ahégao class present in the original dataset is excluded. All images are subjected to the three-stage cleaning pipeline and the five-stage preprocessing pipeline described in Section 4.3.1. Each image is standardised to an RGB resolution of 224×224 pixels.

For FER training, a final balanced dataset of approximately 12,515 images is required. Class balance is a functional requirement because an imbalanced dataset may cause the classifier to become biased towards majority emotion classes, reducing sensitivity to negative emotional states that are clinically relevant to depression detection.

The D-Vlog dataset contains YouTube vlog recordings labelled according to the presence or absence of depression. The dataset includes predefined training, validation, and test folds for the depression detection stage. It was obtained from the original data providers through a formal email request.

Following preprocessing, 667 videos are usable for classifier development, and 149 videos are reserved in the predefined held-out test fold for unbiased evaluation. The Stage 2 pipeline requires preprocessed facial frames and pre-extracted 25-dimensional per-frame acoustic feature vectors supplied with the dataset.

3.1.3 Functional Requirements

The FER module must accept an RGB facial image with a resolution of 224×224 pixels and output a five-dimensional softmax probability distribution over the expression classes: angry, happy, neutral, sad, and surprise. The hybrid FER ensemble must combine the predictions of five individual models using validation-F1-weighted soft voting.

The depression estimation module must accept a complete D-Vlog video, sample it at one frame per second, and apply the same five-stage preprocessing pipeline using video-specific face-detection parameters. The preprocessed frames must then be passed through the frozen FER ensemble to generate a $T \times 5$ probability sequence, where T is the number of retained frames.

The visual sequence must be combined with the $T \times 25$ acoustic feature sequence to form a $T \times 30$ per-frame multimodal sequence. Videos containing fewer than 10 valid frames after preprocessing must be excluded from classifier training and evaluation.

The feature-engineering pipeline must use a two-stage feature selection process. Mutual Information (MI) scoring must be used in the first stage, and Recursive Feature Elimination with Cross-Validation (RFECV) must be used in the second stage. Both processes must be fitted only on the training fold. The selected features must then be standardised using a StandardScaler fitted only on the training fold to prevent data leakage.

Multiple classifiers must be trained and compared on the validation fold. The selected classifier must be evaluated only once on the held-out test fold. Accuracy, macro F1-score, per-class precision, recall, F1-score, and AUC-ROC must be reported as performance metrics.

3.2 Societal Impact

In low-resource countries, depression is severely underdiagnosed, in part because of a lack of trained mental health practitioners and because of stigma toward psychiatric assistance. A camera-and-microphone-based depression screening device that is non-invasive can expand mental health assessment to populations that presently can access or do not have access to clinical services, allowing earlier diagnosis and decreasing the social and economic cost of untreated depression. The fact that publicly available vlog data, which records natural facial and vocal behaviour with non-specialised recording devices, can be used proves that such a system can, in theory, be run even on the normal consumer technology.

At the same time, automated mental health tools carry societal risks that must be acknowledged. False positives, which means false identification of a healthy person as depressed, can lead to unnecessary suffering, stigmatisation and unnecessary clinical referrals. False negatives can result in false confidence among people who are actually at risk. It is with these reasons that the proposed system is clearly placed as a screening and monitoring aid, rather than a diagnostic instrument, and its outputs should never be interpreted without the input of a competent clinician.

3.3 Environmental Impact

The compilation of nine deep learning architectures, as well as the creation of an ensemble and the training of LSTM/GRU sequence models, constitutes a significant amount of GPS workload, including energy usage. In order to alleviate this, pretrained weights were taken as initial values wherever possible, and fewer training epochs were necessary. There was no redundant retraining once final hyperparameters had been set and model training was done on common institutional hardware. The inference pipeline, after being trained, can be run on the FER ensemble in frozen mode, which is computationally light to run. Software requirements are open-source and do not need proprietary cloud infrastructure and reduce the energy footprint of the deployed systems.

3.4 Ethical Issues

Facial and audio data mental health inference raises a number of ethical issues which were put into consideration. Both datasets, OAHEGA and D-Vlog, are open research datasets that are published and may be used by academics. The D-Vlog dataset consists of YouTube vlogs that were made publicly available by participants, and its utilisation in this research will be based on the conditions of its publication. There was no private or clinically sensitive information gathered in this study.

The key concerns surrounding any further deployment are consent and privacy. Patients should be clearly told whether their facial expression is undergoing mental health screening, and they should be given consent before it is analysed. It should not be used in an undercover manner. Caution should also be observed to ensure that the model is not coded with demographic biases; the D-Vlog data lacks demographic diversity, and this limits the extrapolability of the trained depression classifier, and this should be made clear whenever using it in a clinical or application context. The product of the system should not be utilised in any other way other than health support such as employment screening, underwriting insurance, or any other discriminatory method.

3.5 Standards

This study complies with the standard machine learning evaluation guidelines. The pre-determined train, validation and test folds of the D-Vlog dataset are not altered, allowing them to be compared to previously published results on the benchmark. The test fold is only consulted in a single instance, such as, but not limited to, the last evaluation, and never during model development or hyperparameter tuning. Normalisation of all pixel values to the $[0, 1]$ range is performed before model input; data augmentation is done only to the training set, and all the preprocessing operations are recorded with their full parameter values to guarantee reproducibility. Only the training and validation folds are subject to feature selection and tuning of the hyperparameters with 5-fold stratified cross-validation, without the test fold being seen at any point of the development. Accuracy, precision, recall, F1-score, and AUC are used to report performance, which is in line with best practices in AI research on affective computing and mental health.

3.6 Project Management Plan

The project was conducted in two semesters in a well-organised step-by-step plan. The initial stage entailed literature review, dataset acquisition, and pipeline development, i.e., cleaning, preprocessing and augmentation of the OAHEGA dataset. The second stage involved model training, ensemble-building, D-Vlog visual and audio-preprocessing, and depression-classifier development in both the feature-engineering and sequence-modelling strategies. Evaluation, analysis of results and writing of the thesis were set aside as a final phase. The workload was shared among the team members based on specialisation, where each person owned a particular architectural implementation and pipeline elements. The advancements were monitored by regularly meeting with supervisors, and Git was used to maintain version control during the whole process.

3.7 Risk Management

A number of risks were identified during the design phase and alleviated through careful design decisions. The threat of the unavailability of the datasets was also partially fulfilled: out of the 961 initially labelled D-Vlog videos, 756 could be retrieved, and 53 of them were then discarded during the preprocessing stage since they produced no useful frames. This was countered by applying the predefined dataset splits without modification to maintain the same distribution of classes and maintainability. To reduce the risk of model overfitting, dropout, weight decay, and early stopping were implemented in all architectures. Data leakage risk was managed by ensuring that the test fold was isolated at each stage. The risk of hardware failure was addressed by periodically saving model checkpoints throughout training. The overfitting risk posed by the 755-feature fused representation compared to the small D-Vlog training set (414 videos) is directly handled using StandardScaler normalisation, low-variance feature selection, SelectKBest feature selection and 5-fold stratified cross-validation of all hyperparameter selection. The threat of corrupted or missing audio in single videos is reduced by recording and excluding corrupted videos and falling back to visual-only features where audio extraction cannot be done.

3.8 Economic Analysis

The main expenses of the project were computational, consisting of the use of GPUs to train the models and preprocess D-Vlog frames. The software is free and all open-source, and there are no software expenses. The two datasets are research data that are publicly accessible and free to get. In a more economical analysis, the possible worth of a scalable camera-based depression screening device is immense. Early diagnosis and treatment of depression will significantly lower downstream medical expenses, lost productivity and social welfare spending. A system capable of operating on commodity hardware and requiring no specialised clinical equipment have a most promising cost-to-benefit analysis in healthcare systems in low- and middle-income settings, in which the economic burden of untreated mental illness is disproportionately high.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

The proposed system is a novel automatic depression detection system based on naturalistic video input. It consists of two sequentially connected modules: a Facial Expression Recognition (FER) Model and a multimodal Depression Classifier.

The design process was motivated by three main goals: clinically reliable facial expression recognition, particularly at the neutral-sad boundary; the extraction of meaningful behavioural features from real-world vlog videos using both visual and acoustic modalities; and the development of a depression classifier that generalises reliably despite the limited number of accessible videos.

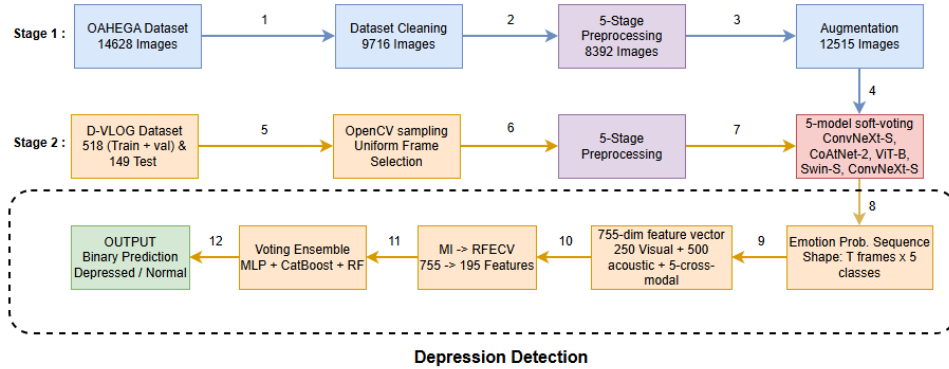


Figure 4.1: End-to-end pipeline diagram

Several important constraints influenced the system architecture. The D-Vlog dataset is not publicly available and the videos had to be obtained by directly contacting the original authors. The original dataset contained 961 videos, but only 756 videos remained accessible because some YouTube videos had been deleted. There is also a domain gap between the OAHEGA image dataset used to train the FER Model and the naturalistic vlog recordings included in the D-Vlog dataset. After preprocessing and filtering, the effective training set for depression classification contained only 414 videos. This limited sample size restricted the use of end-to-end deep learning approaches for the depression classifier

and favoured classical ensemble-classification methods combined with principled feature selection.

The complete pipeline consists of two phases, as illustrated in Figure 1.1. The first phase is the FER training phase, where the FER component is developed using only the OAHEGA image dataset. Under equivalent experimental conditions, deep learning architectures from three distinct architectural families are trained and evaluated: pure convolutional, pure transformer, and hybrid CNN-attention models. The five best and most architecturally diverse models are combined using weighted soft voting to create the hybrid FER ensemble. The weights are determined by the validation F1-scores of the individual models. The final FER Model achieves a test accuracy of 93.13% and a macro F1-score of 0.9315 on the held-out OAHEGA test set. After training, the parameters of the FER ensemble are frozen and are not updated during the development of the depression classifier.

In the second phase, which covers inference and depression classification, the frozen FER ensemble is applied to D-Vlog video frames to generate emotion probability sequences as visual inputs for the depression-classification model. The D-Vlog dataset used for classifier development contains 518 development videos and 149 held-out test videos after applying the minimum-frame filter.

Using OpenCV, one frame per second is sampled from each video. Each extracted frame is resized and standardised to a resolution of 224×224 pixels. The sampled frames are then processed using a five-stage video-preprocessing pipeline aligned with the OAHEGA preprocessing procedure: MTCNN-based face alignment, CLAHE illumination normalisation, Difference of Gaussians shading removal, MediaPipe ROI masking, and Laplacian-variance quality verification. Frames without a valid face detection or sufficient sharpness are discarded.

Each retained frame is passed through the frozen five-model FER ensemble, consisting of CoAtNet-2, ViT-Base/16, Swin-Small, ConvNeXt-Small, and ConvNeXt-Base. For each frame, the ensemble outputs a five-dimensional softmax probability vector over the expression classes: angry, happy, neutral, sad, and surprise. The per-frame probability vectors are concatenated across all retained frames of a video to create an emotion-probability sequence of size $T \times 5$, where T is the number of retained frames. Each row of the matrix represents the emotion-probability distribution for one frame.

The $T \times 5$ emotion-probability sequence is then passed to the feature aggregation stage. A total of 250 visual features are derived from the emotion sequences, including statistical descriptors, temporal patterns, and clinically relevant affect-related markers. These features are combined with 500 acoustic features extracted from the $T \times 25$ acoustic sequence supplied with the D-Vlog dataset. Five additional cross-modal synchrony features are calculated using Pearson correlations between the emotion sequences and the acoustic signals. This produces a 755-dimensional multimodal feature vector for each video.

The 755-dimensional representation is reduced to 195 selected features using a two-stage feature-selection pipeline. Mutual Information scoring is applied first, followed by Recursive Feature Elimination with Cross-Validation (RFECV). Both stages are fitted only on the training fold to prevent data leakage. The selected features are then passed to the depression classifier.

The final depression classifier is a stacking ensemble composed of MLP, CatBoost, and Random Forest models, with Logistic Regression used as the meta-learner. It is trained on the D-Vlog training fold and evaluated only once on the held-out test fold. The final output is a binary classification of each input video as either depressed or normal, together with a confidence score.

The detailed design specifications, data-collection procedure, preprocessing stages, and implementation details for each component are described in the following sections.

4.2 Preliminary Design and Model Specification

The suggested system is a separation of two interconnected subsystems: a facial expression recognition model and a depression classifier. Both of the components involved the consideration of several alternative design solutions prior to the final model being chosen. This section discusses the other architecture options considered in both components, the comparisons that were conducted in the experiment, and the reasons why the final choices were made. Section 4.2.1 discusses the FER model options that were trained and tested on the OAHEGA dataset, and Section 4.2.2 discusses depression classifier options that are trained and tested on expression-based features extracted from the D-Vlog dataset.

4.2.1 FER Model Alternatives

The facial expression recognition component was to be designed, and the alternatives that were to be evaluated were various deep learning architectures. The aim was to determine the models that had good classification ability on five expression classes, which are angry, happy, neutral, sad, and surprised, with special focus on minimising the confusion between the neutral and sad classes, which is the most clinically important border to detect downstream depression. Eight independent models were trained and tested on the OAHEGA dataset with the same experimental setup, data splits, random seed, input resolution, augmentation pipeline, and metrics to ensure a fair comparison. Two classical baseline models (ResNet-50 and VGG16) were also provided to put the performance improvements of modern architectures into perspective. After individual analysis, the five best models were fused to a weighted soft-voting ensemble to create the final hybrid FER model.

All models were trained on images that were scaled to 224×224 pixels. The data was stratified into 70% training, 15% validation and 15% test sample sets with a constant random seed of 42. All models were trained using the AdamW optimiser with a layer-wise learning rate decay. In most models, a custom neutral-sad separation loss was added to the main cross-entropy loss to explicitly promote simultaneous uncertainty between the neutral and sad classes. The data augmentation was maintained in all experiments and consisted of random horizontal flipping, random rotation up to ± 15 degrees, colour jitter, random affine translation, random greyscale conversion, random erasing, Mixup, and CutMix. The validation macro F1 score was used to select a model during training, and an optimal checkpoint was stored to be used at the end of the test.

4.2.1.1 CoAtNet-2

CoAtNet-2 is a hybrid architecture that incorporates convolutional and transformer-based processing in a four-step pipeline in a C-C-T-T framework. The initial two steps use MBConv convolutional blocks to extract local facial characteristics like edges, textures, and muscle movements of the face.

The two latter phases use relative self-attention transformer blocks to simulate global contextual relationships that are the whole face region.

The trained model variant is CoAtNet_2_rw_224.sw_in12k_ft_in1k, pretrained on ImageNet-21k and fine-tuned on ImageNet-1k, which has a total of 72.8 million parameters. The last classification head was changed to a custom Dropout(0.3) and then a fully connected linear layer to five output classes.

The model was trained by the AdamW optimiser and layer-wise learning rate strategy. The stem was set to a learning rate of 1×10^{-5} , the convolutional stages were set to 3×10^{-5} , the transformer stages were set to 6×10^{-5} , and the classification head was set to 1×10^{-4} . The drop path rate was 0.2, and the weight decay was 0.05. The learning routine was a 5-epoch linear warm-up and, subsequently, cosine annealing decay.

The training was done over 60 epochs and an early stopping patience of 15, which was never reached. Batch size was set to 32, and mixed-precision training was used via PyTorch AMP. CutMix with alpha 1.0 and Mixup with alpha 0.2 were used. The loss function was a combination of soft target cross-entropy and neutral-sad separation loss, weighted 0.2. The neutral-sad separation loss explicitly penalises the probability assigned to the opposite class for samples whose true label belongs to {Neutral, Sad}:

$$\mathcal{L}_{NS} = -\frac{1}{|\mathcal{S}_{NS}|} \sum_{n \in \mathcal{S}_{NS}} \log(1 - \hat{p}_{n, \bar{c}(y_n)}) \quad (4.1)$$

where $\mathcal{S}_{NS} = \{n : y_n \in \{\text{Neutral}, \text{Sad}\}\}$ and $\bar{c}(\text{Neutral}) = \text{Sad}$, $\bar{c}(\text{Sad}) = \text{Neutral}$. The combined training objective is:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{NS} \quad (4.2)$$

where $\lambda = 0.2$ for CoAtNet-2.

Epoch 56 had the best checkpoint with a validation F1 of 0.9322. CoAtNet-2 had a test set of 92.44% and macro F1 of 0.9246. Per-class F1 scores were 0.970 for surprise, 0.963 for angry, 0.963 for happy, 0.880 for sad, and 0.853 for neutral. The lowest error of all the single models tested was total neutral-sad confusion errors at 74, indicating the effectiveness of the C-C-T-T hybrid architecture in managing this boundary.

4.2.1.2 ViT-Base/16

ViT-Base/16 is a pure transformer architecture that aligns with images by splitting them into fixed-size patches and uses the global self-attention over all patches at the same time. In contrast to the hierarchical construction of local features in convolutional models,

ViT encodes the long-range correlations within the entire facial region at the first layer and is especially useful at identifying holistic facial structures and wide-range patterns of facial expressions. An input image of size $H \times W$ is partitioned into $N = HW/P^2$ non-overlapping patches of size $P \times P$ ($P = 16$). Each flattened patch is linearly projected and combined with a learnable positional encoding; a CLS token is prepended to carry the classification signal.

Each of the 12 transformer encoder layers applies MSA and a two-layer MLP with residual connections and layer normalisation.

The model variant provided is `vit_base_patch16_224`, which is pre-trained on ImageNet-21k and fine-tuned on ImageNet-1k and has 85.8 million parameters. x

To train the model, the AdamW optimiser was used with a four-level layer-wise learning rate policy. The patch embedding layer and positional embedding layer were set to receive 5×10^{-6} ; early transformer blocks received 1.5×10^{-5} , middle blocks received 3×10^{-5} , late blocks received 4.5×10^{-5} , and the normalisation layer, which had the classification head, was set to receive the entire base learning rate of 5×10^{-5} . The drop path rate was set to 0.1 and the weight decay to 1×10^{-4} . Gradient accumulation with 2 steps was employed to essentially increase the batch size by 2. A weight of $\lambda = 0.25$ was used to apply the neutral-sad separation loss (see (4.2)).

Between epochs 57 and 59, the best validation F1 of 0.9222 was achieved. ViT-Base/16 on the test set had a macro F1 of 0.9228 and an accuracy of 92.28%. Per-class F1 scores were 0.975 for surprise, 0.962 for angry, 0.962 for happy, 0.859 for sad, and 0.858 for neutral. There were 78 total errors of neutral-sad confusion. Although there is a little more neutral-sad confusion than with CoAtNet-2, ViT-Base/16 adds constructive architectural diversity to the ensemble by its radically different global attention-based feature extraction.

4.2.1.3 Swin-Small

Swin Transformer Small is a hierarchical vision transformer, which takes images in four progressive steps with shifted window-based self-attention. Unlike ViT-Base/16, which uses global attention on all patches at the same time, Swin only uses local non-overlapping windows of size $M \times M$ and moves them across layers to enable cross-window interaction.

This hierarchical structure generates feature maps at various scales and is useful in both the local detail of faces and expression-level semantics. It is the `swin_small_patch4_window7_224` model variant, which was trained on ImageNet-1k and has 48.8 million parameters.

The AdamW optimiser, a four-level layer-wise learning rate strategy, was used to train the model. The patch embedding was 1×10^{-5} , stages 0 and 1 were 3×10^{-5} , stages 2 and 3 were 6×10^{-5} , and the head of classification was 1×10^{-4} . It used a weight decay of 0.05, drop path rate of 0.2, and gradient clipping with a maximum norm of 1. A weight of $\lambda = 0.3$ was used to apply the neutral-sad separation loss (4.1). The optimal point of the checkpoint was at epoch 54 with an F1 of 0.9217.

Swin-Small had a test set accuracy of 91.91% and a macro F1 of 0.9187. Per-class F1

scores were 0.970 for surprise, 0.960 for happy, 0.950 for angry, 0.870 for sad, and 0.850 for neutral. There were 75 total neutral-sad confusion errors. Swin-Small adds useful hierarchical multi-scale feature representations to the ensemble to supplement the global attention of ViT-Base/16 and the hybrid processing of CoAtNet-2.

4.2.1.4 ConvNeXt-Base

ConvNeXt-Base is a pure convolutional network, built on transformer-inspired design principles. It belongs to the architectural family of ConvNeXt-Small, but the number of parameters varies between 49.5 million and 87.6 million. The defining computational unit is the ConvNeXt block, which applies a large-kernel 7×7 depthwise convolution followed by layer normalisation and an inverted bottleneck with GELU.

The variant applied is `convnext_base.fb_in22k_ft_in1k`, which is ImageNet-22k trained and ImageNet-1k fine-tuned.

The AdamW optimizer and three-level layer-wise learning rate strategy were applied to train the model. The stem was given 8×10^{-6} , the convolutional stages were given 3.2×10^{-5} , and the classification head was given 8×10^{-6} . A weight decay of 5×10^{-4} , drop path rate of 0.20, and Dropout(0.35) were used. It was performed with a gradient accumulation of 2 steps with an effective batch size of 32. Neutral-Sad Separation loss (4.1) was used, with $\lambda = 0.25$ and macro F1 as the primary model selection criterion, and neutral-sad error count as a tiebreaker.

Epoch 27 had the highest F1 on the validation set of 0.9254. ConvNeXt-Base gave a result of 91.05% on the test set with a macro F1 of 0.9105. Per-class F1 scores were 0.950 for happy, 0.950 for surprise, 0.940 for angry, 0.870 for sad, and 0.840 for neutral. The total errors of neutral-sad confusion were 70, the lowest of all individual models, showing that the bigger capacity of ConvNeXt-Base results in a more discriminative neutral-sad boundary despite the fact that its overall accuracy does not exceed that of ConvNeXt-Small.

4.2.1.5 MaxViT-T

MaxViT-T is a hybrid architecture that uses both local window attention and global grid attention at every stage, enabling it to simultaneously capture fine-grained local facial information and long-range global expression patterns. Local attention operates within non-overlapping $w \times w$ windows and uses the relative position bias.

The version of the model is `maxvit_tiny_rw_224`, which is pretrained with ImageNet-1k, and contains 28,546,877 parameters. The classification head was changed with Dropout (0.3), then a linear layer that transforms 512 features into five classes of outputs.

The AdamW optimizer with a four-level layer-wise learning rate approach between 1×10^{-5} and 1×10^{-4} between the stem and classification head was used to train the model. It used a weight decay of 0.05, drop path rate of 0.2, gradient clipping with norm 1.0, and neutral-sad separation loss (4.1) with $\lambda = 0.3$. The highest validation F1 was 0.9216 at epoch 35.

MaxViT-T performed with an accuracy of 90.36% and a macro F1 of 0.9042 on the

test set. Per-class F1 scores were 0.960 for angry, 0.960 for happy, 0.950 for surprise, 0.850 for sad, and 0.820 for neutral. The highest number of total neutral-sad confusion errors was 89, which was observed in hybrid and transformer models, and the pattern was heavily asymmetrical, indicating systematic bias in predicting neutral over sad. Because of its lower overall accuracy and greater neutral-sad confusion, MaxViT-T was not chosen to be a part of the hybrid ensemble.

4.2.1.6 EfficientNet-B7

EfficientNet-B7 is a convolutional architecture that is entirely founded on the concept of compound scaling and has 63.8 million parameters. Its native input resolution is 600×600 ; however, to ensure a fair comparison, it was trained at 224×224 , which implies that the model was not utilized to its full potential. The head of the classification was changed into Dropout(0.4), then a linear layer that maps 2560 features to five output classes.

This model was optimised using the AdamW optimiser and a four-level layer-wise learning rate schedule among the seven block groups (1×10^{-5} to 1×10^{-4}) with stem to head, respectively. Neutral-sad separation loss with $\lambda = 0.3$ and weight decay 1×10^{-4} was used. The patience of early stopping was fixed at 10 epochs. The optimal checkpoint was at epoch 38 with a validation F1 of 0.9136.

On the test set, EfficientNet-B7 scored 90.36% and a macro F1 of 0.9028. Total neutral-sad confusion errors were 84, where there was an asymmetric error of 24 neutral-to-sad versus 60 sad-to-neutral errors. This asymmetric pattern of confusion, in combination with the resolution constraint, resulted in EfficientNet-B7 not being included in the hybrid ensemble.

4.2.1.7 ResNet-50

ResNet-50 is a conventional deep convolutional neural network, which was proposed in 2015 and is based on residual skip connections. It is the classical CNN baseline of this study with 23.5 million parameters, which symbolizes the pre-modern-transformer stage of image classification. The classification head was substituted by Dropout(0.3), then a linear layer that took 2048 features to five output classes, pretrained on ImageNet-1k. The model was trained with AdamW optimizer and a four-level layer-wise learning rate scheme, 1×10^{-5} to 1×10^{-4} to the head, weight decay 1×10^{-4} . The trained model did not cause an early stopping and achieved a top validation F1 of 0.8837 using the default 60 epochs.

ResNet-50 scored the lowest in accuracy (87.43%) and macro F1 (0.8746) on the test set. Errors of total neutral-sad confusion had 104, which is the highest among all models. These findings affirm the large performance disparity between conventional residual architectures and current attention-based networks in facial expression recognition. ResNet-50 was also incorporated as a pure classical baseline and was never a candidate in the hybrid ensemble.

4.2.1.8 VGG-16

VGG-16 is a classic deep convolutional neural network, which was proposed in 2014, constructed entirely with uniform 3×3 convolutional filters stacked in 13 layers of convolutional, followed by three fully connected layers.

It is a second classical baseline in this paper, in addition to ResNet-50, which depicts depth by simplicity as opposed to residual connections or attention mechanisms.

The AdamW optimizer and three-level layer-wise learning rate strategy were used to train the model. The convolutional layers were scaled by 1.5×10^{-5} , the fully connected layers were scaled by 3×10^{-5} , and the classification head was scaled by the full base rate of 5×10^{-5} , which is lower than all other models since VGG-16 is sensitive to high learning rates without residual connections. Gradient clipping to a maximum norm of 2.0 was used.

Checkpoint 22 was identified as the best, with a validation F1 of 0.9040. On the test set, VGG-16 achieved an accuracy of 89.35% and a macro F1 of 0.8937. Per-class F1 scores included 0.961 for surprise, 0.950 for happy, 0.939 for angry, 0.826 for sad, and 0.792 for neutral, the lowest neutral F1 of any model in this study. Errors of total neutral-sad confusion were 101. Although VGG-16 has the largest number of parameters of all the other models considered, it loses to all the modern architectures, which validates that the number of parameters does not directly correspond to representational efficiency without current design decisions. VGG-16 was only included as a classical baseline, and it was not under consideration in the hybrid ensemble.

4.2.1.9 ConvNeXt-Small

ConvNeXt-Small is a contemporary pure convolutional architecture that features design ideas inspired by transformers: 7×7 depthwise convolutions, inverted bottleneck blocks, layer normalisation, and GELU activations. It is a member of the ConvNeXt family presented in 2022 as a systematic reevaluation of whether a simple convolutional network rearchitected to align with the macro-level structure of vision transformers could rival or outperform transformers. The variant model used is `convnext_small.fb_in22k-ft-in1k` that was pretrained on ImageNet-22k and fine-tuned on ImageNet-1k, with 49.5 million parameters and 224×224 input resolution. The classification head was substituted with Dropout(0.3), then a linear layer that transforms 768 features to five classes.

The AdamW optimizer was used to train the model with a three-level layer-wise learning rate strategy. The stem was used with 1×10^{-5} , the four convolutional stages were used with 3×10^{-5} , and the classification head was used with the entire base learning rate of 1×10^{-4} , weight decay of 1×10^{-4} and cosine annealing decay with 80 epochs and an early stopping warning of 14. In contrast to all other models in this paper, ConvNeXt-Small was trained without the Neutral-Sad Separation loss, instead using the regular cross-entropy of soft targets with Mixup and CutMix regularization across the board.

Training reached the optimal point at epoch 42 and had a validation F1 of 0.9299. ConvNeXt-Small had an accuracy of 91.69% and macro F1 of 0.9178 on the test set. Per-class F1 scores were 0.972 for surprise, 0.963 for angry, 0.959 for happy, 0.863 for sad, and 0.833 for neutral. The total neutral-sad confusion errors were 87, with an asymmetrical

error figure of 30 neutral-to-sad and 57 sad-to-neutral errors, which is in line with the lack of separation loss. Nevertheless, ConvNeXt-Small adds powerful convolutional feature representations and architectural diversity to the ensemble as a fundamental element.

4.2.1.10 Five-Model Weighted Ensemble (Hybrid FER Model)

The five highest performing and architecturally diverse models were chosen after considering all the individual models and combined into a hybrid FER model by weighing the models in a weighted soft-voting ensemble. The chosen models were ConvNeXt-Small, CoAtNet-2, ViT-Base/16, Swin-Small, and ConvNeXt-Base, which are three different architectural families, such as pure convolutional, pure transformer, and hybrid CNN-attention, which guarantees that the ensemble takes advantage of complementary and diverse representations of features. MaxViT-T, EfficientNet-B7, and ResNet-50 were not chosen because of their lower overall accuracy, more neutral-sad confusion errors, or architectural redundancy with other members already chosen.

The ensemble functions on the basis of inference time without any extra training. All five models separately handle an input image to generate a softmax probability distribution across the five expression classes. The five probability vectors are combined through weighted soft-voting, with each model’s contribution proportional to its validation F1 score. Let w_m denote the raw (pre-normalisation) validation F1 of model m . The normalised weights are:

$$\tilde{w}_m = \frac{w_m}{\sum_{m'=1}^5 w_{m'}}, \quad m = 1, \dots, 5 \quad (4.3)$$

The pre-normalisation weights of ConvNeXt-Small, CoAtNet-2, ViT-Base/16, Swin-Small, and ConvNeXt-Base were 0.9318, 0.9322, 0.9222, 0.9217, and 0.9254, respectively, giving a sum of 4.6333 and normalised weights of approximately 0.2011, 0.2012, 0.1991, 0.1990, and 0.1997. The final predicted class is the argmax of the resulting weighted average probability vector:

$$\hat{y} = \arg \max_{c \in \mathcal{C}} \sum_{m=1}^5 \tilde{w}_m p_c^{(m)}, \quad \mathcal{C} = \{\text{Angry, Happy, Neutral, Sad, Surprise}\} \quad (4.4)$$

The hybrid ensemble obtained a test accuracy of 93.13% and a macro F1 of 0.9315 on the same held-out test set of 1,878 images, which is 0.64 percentage points better than the best single model. Per-class F1 scores were 0.980 for surprise, 0.964 for angry, 0.964 for happy, 0.885 for sad, and 0.865 for neutral. The total neutral-sad confusion errors were 69, which is a better score compared to all the individual models. This five-model weighted ensemble forms the last hybrid FER model that is incorporated in the complete depression detection pipeline.

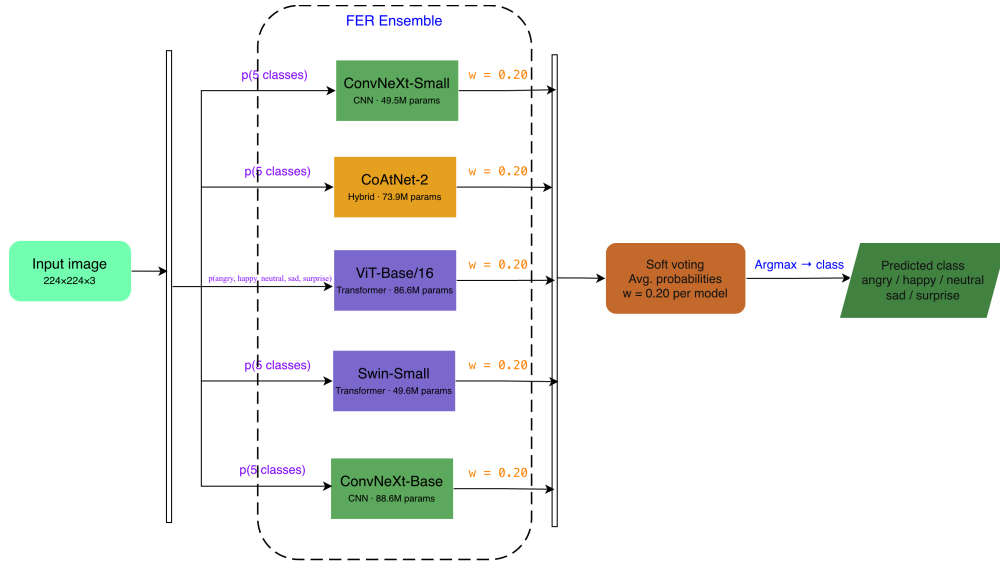


Figure 4.2: FER ensemble architecture diagram.

Rank	Model	Family	Test Acc	Macro F1	Best Val F1	NS Errors
1	CoAtNet-2	Hybrid	92.44%	0.9246	0.9322	74
2	ViT-Base/16	Transformer	92.28%	0.9228	0.9222	78
3	Swin-Small	Transformer	91.91%	0.9187	0.9217	75
4	ConvNeXt-Base	CNN	91.05%	0.9105	0.9254	70
5	MaxViT-T	Hybrid	90.36%	0.9042	0.9216	89
6	EfficientNet-B7	CNN	90.36%	0.9028	0.9028	84
7	ConvNeXt-Small	CNN	91.69%	0.9178	0.9299	87
8	Assemble 5 models		93.13%	0.9315		69
9	Resnet50	CNN	87.43%	0.8746	0.8837	104
10	VGG16	CNN	89.35%	0.8937	0.9040	101

Table 4.1: All FER models comparison

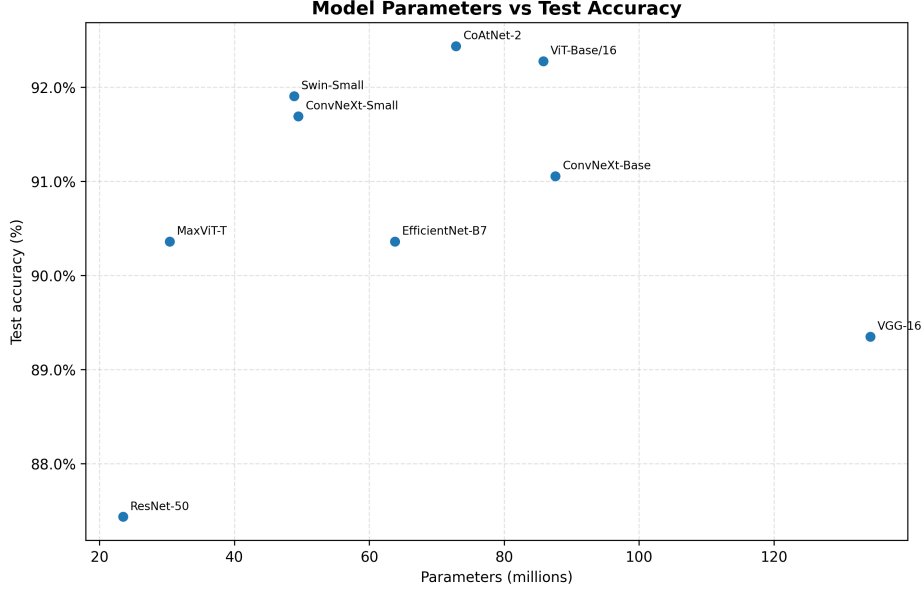


Figure 4.3: Scatter plot - model parameters vs test accuracy for all FER models.

4.2.2 Preliminary Design and Model Specification

The pipeline takes in two modalities based on YouTube vlog videos: a $T \times 5$ sequence of visual emotion probabilities obtained by the trained FER network, and a 25-dimensional acoustic feature vector. There were seven candidate solutions of classifier architecture, including single base learners, heterogeneous ensembles, and attention-based deep models. All the candidates are outlined, their design reasons are mentioned, and the performance of their validation is provided. Prior to the design of the classifiers, the multimodal inputs were transformed into a fixed-length feature vector in a two-step process. The handcrafted features extracted per video were initially 250 visual features based on the $T \times 5$ emotion probability sequences (statistical descriptors, temporal slopes, segment dynamics, and approximate entropy), 500 acoustic features based on 25 dimensional acoustic vectors (distributional statistics, temporal slopes, segment dynamics, and approximate entropy) and 5 cross-modal synchrony features of the Pearson correlation between each emotion sequence and the average acoustic signal. A two-stage feature selection pipeline was then used only on the training fold to eliminate data leakage. The 300 most informative features were retained by a Mutual Information scoring with a fixed random seed (`random_state=42`). Recursive Feature Elimination with Cross-Validation (RFECV) based on a 100-tree Random Forest estimator, 5-fold stratified cross-validation, and F1-macro scoring narrowed the set further to 195 features. A StandardScaler was then fitted to all selected features on the training fold only to standardise them. This leakage-free, reproducible pipeline was the universal input representation to all candidate classifiers, but the recurrent architectures, which can directly handle the raw sequences.

4.2.2.1 Random Forest

A 500-tree Random Forest (RF) was trained on the 195 features (selected). Random Forest is an ensemble of independent trees that utilise bootstrap samples and random selection of subsets of features at each split. The fact that it is not sensitive to feature scaling,

intrinsic feature ranking importance, and overfitting to small datasets made it an obvious benchmark in this study. The small class imbalance (1.22:1) was corrected by balanced weighting of classes (`class_weight=balanced`). The minimum number of samples per split and to control the depth of trees, 4 and 2 samples per split and leaf, respectively, were set.

RF on the 104-video held-out validation fold had a validation accuracy of 71.15%, Macro F1-Score of 0.7076, depression recall of 0.7679, and AUC-ROC of 0.7950. Although each of the models was competitive in its own right, the model showed a recall gap of 0.122 between the depression and normal classes, which is indicative of the sensitivity of the model to the weighting strategy of classes.

4.2.2.2 Multi-Layer Perceptron (MLP)

The Multi-Layer Perceptron, with three hidden layers of 256, 128, and 64 neurons, was trained using the Adam optimiser and a learning rate of 0.001 and L2 regularisation coefficient (`alpha`) of 0.001. Throughout, rectified linear unit (ReLU) activations were used. To alleviate overfitting, early stopping was observed on a 10% internal validation split, and a 10-epoch patience was used. The MLP is also able to simultaneously learn non-linear interaction effects with all 195 features, without any explicit constraints on feature importance, unlike the tree-based models.

The MLP had an accuracy of 72.12% and a Macro F1-Score of 0.7190, AUC-ROC of 0.8013 - the best AUC of all the base models - and a recall gap of 0.063 on the validation fold. The relatively small recall gap suggests that the MLP evenly allocates prediction errors across the classes as compared to tree-based models.

4.2.2.3 CatBoost

CatBoost is a gradient boosting model that builds an additive ensemble of decision trees sequentially, error correcting. It uses symmetric (oblivious) decision trees and ordered boosting to minimize prediction shift and overfitting on small datasets. Training was performed with (`task_type='GPU'`) using 500 boosting iterations, a maximum tree depth of 8, and a learning rate of 0.05. Auto class weighting (`auto_class_weights='Balanced'`) was used.

CatBoost had a validation accuracy of 67.31%, a Macro F1-Score of 0.6700, and an AUC-ROC of 0.7716. Competitive in cross-validation (CV F1-Macro of 0.7074), its held-out validation was worse than RF and MLP. Nevertheless, CatBoost became a very useful addition to ensemble architectures in which its gradient-boosted predictions added a lot of variability to the RF and MLP base models.

4.2.2.4 XGBoost

XGBoost is a gradient boosting library that uses second-order gradient statistics to construct trees, subsample columns and rows, and has built-in regularisation to check the complexity of the models. The histogram tree method (`device='cuda'`) was used to enable training that is accelerated by the GPU. To counter the bias of the model to favour depression recall, a corrected `scale_pos_weight` parameter ($n_{dep} / n_{norm} \approx 1.19$) was

used to upweight the minority normal class. It was configured with 500 estimators, a depth of 6 maximum and a learning rate of 0.05.

The validation accuracy of the XGBoost was 68.27%, the macro F1-score was 0.6812, and the recall gap was very low, at 0.030, which is the lowest of all single-base models. Its cross-validation F1-Macro of 0.7188 was the highest of all base models, which means that it has high generalisation ability in training. XGBoost was an appreciated addition to multi-model ensembles due to the combination of balanced recall and strong CV performance.

4.2.2.5 TabNet

TabNet is an attention neural network that is created to operate on tabular data. Each decision step involves a sequential attention mechanism that picks a sparse set of features to attend to, allowing the model to attend to the most informative features at each step, as opposed to attending to all the features equally. The interpretable feature selection mechanism is especially appropriate to the high-dimensional tabular input with heterogeneous features of different types, like in this case, 195 mixed visual-acoustic features. TabNet was trained using attention embedding width ($n_d = n_a = 32$), five sequential attention steps, entmax sparse attention masking, and balanced class weighting using manual class weights.

TabNet has a validation accuracy of 71.15% and a Macro F1-Score of 0.7106 - competitive with the best individual models. It is interesting to note that it had the best balance of recall distribution compared to all the models that were tested, with depression recall of 0.7143, normal recall of 0.7083, and a gain in recalls of only 0.006. Nevertheless, it had an AUC-ROC of 0.7150, which is significantly less than the feature-based classifiers, showing worse probability calibration, although competitive hard-classification performance.

4.2.2.6 Recurrent Neural Network Architectures (Bi-LSTM and Bi-GRU)

The performance of bidirectional Long Short-Term Memory (Bi-LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU) networks was compared as sequence-to-label classifiers that directly use the raw $T \times 5$ emotion probability sequences and the $T \times 25$ acoustic feature vectors without aggregation. Both models use dual-branch architectures where an emotion sequence is processed by one of the recurrent branches and an acoustic sequence by the other. The last hidden states of both branches are combined and fed into a three-layer fully connected classifier with LayerNorm and dropout regularisation. Videos were padded or truncated to a constant sequence length of 512 frames.

The Bi-LSTM has gated memory cells that have separate input, forget, and output gates, which allow long-range temporal dependencies to be selectively retained. The Bi-GRU is a simplified gating architecture with a joint update and reset gate, which decreases the number of parameters compared to LSTM but does not decrease its modelling capacity. They were both optimised using Adam (learning rate 0.001), ReduceLROnPlateau scheduling, gradient clipping, and early stopping.

Although recurrent architectures have a theoretical advantage in the modelling of temporal sequences, the two models were not competitive with the feature-based classi-

fiers. The Bi-LSTM achieved a validation accuracy of 69.23% (Macro F1-Score 0.6905, AUC-ROC 0.6901) and the Bi-GRU achieved 64.42% (Macro F1-Score 0.6442, AUC-ROC 0.7054). This result is in line with the established fact that recurrent architectures need significantly more training sets to acquire strong temporal representations. The models were underconstrained with only 414 training videos, and the hand-engineered temporal features in the tabular pipeline represented more discriminative information per sample than the learned representations.

4.2.2.7 Ensemble Architectures

The ensemble strategies were compared: soft-voting and stacking. In soft-voting, each base model gives a probability vector of classes, and the overall prediction is by averaging the element-wise averages of the probability vectors. In stacking, a meta-learner (Logistic Regression, $C=1.0$) is learned on the out-of-fold probability predictions of the base-learners using 5-fold stratified cross-validation. Stacking, in theory, is better than voting, as the meta-learner tailored the contribution of the base models to the complementary nature of the models, as opposed to using a uniform distribution of models.

Four ensemble configurations were tested:

Ensemble Method	Base Classifiers	Val Acc	Macro-F1	Dep-Recall	Norm-Recall	AUC
Voting (MLP+CB+RF)	MLP, CatBoost, RF	0.7692	0.7661	0.8214	0.7083	0.7980
Stacking (MLP+CB+RF→LR)	MLP, CatBoost, RF → LR	0.7404	0.7374	0.7857	0.6875	0.7954
Voting (MLP+CB+RF+XGB)	MLP, CatBoost, RF, XGBoost	0.7212	0.7199	0.7321	0.7083	0.7987
Stacking (RF+MLP+XGB+CB→LR)	RF, MLP, XGBoost, CatBoost → LR	0.7308	0.7292	0.7500	0.7083	0.7950

Table 4.2: Ensemble Configurations Comparison (Validation Set)

The best validation accuracy (76.92%) and Macro F1-Score (0.7661) were obtained with the three-model Voting ensemble (MLP + CatBoost + RF) in all of the settings. The addition of XGBoost as a fourth base model did not enhance the performance of the ensemble, which is likely due to the fact that the prediction distribution of XGBoost was largely similar to the prediction distribution of the other gradient-boosted model (CatBoost), and therefore, the ensemble was no longer diverse. The Stacking with identical three base models was less accurate in validation (74.04%) and smaller in the gap of recall (0.098 vs. 0.113), but better in calibration of depression versus normal recall at the expense of absolute accuracy.

4.2.2.8 Comparative Simulation Results

The performance of all eleven classifier settings is summarised in Table 4.2.2.2. Simulations were done on the held-out validation fold (104 videos, 56 depression, 48 normal) after training on only the 414-video training fold. Reported metrics include Validation Accuracy, Macro F1-Score, Depression F1-Score, Depression Recall, Normal Recall, and AUC-ROC.

Model	Type	Val Acc	Macro-F1	Dep-Recall	Norm-Recall	AUC
BASE MODELS						
Random Forest	Tree Ensemble	0.7115	0.7076	0.7679	0.6458	0.7950
MLP	Neural Network	0.7212	0.7190	0.7500	0.6875	0.8013
CatBoost	Gradient Boost	0.6731	0.6700	0.7143	0.6250	0.7716
XGBoost	Gradient Boost	0.6827	0.6812	0.6964	0.6667	0.7798
TabNet	Attention-Tabular	0.7115	0.7106	0.7143	0.7083	0.7150
Bi-LSTM	Recurrent NN	0.6923	0.6905	0.7143	0.6667	0.6901
Bi-GRU	Recurrent NN	0.6442	0.6442	0.6071	0.6875	0.7054
ENSEMBLES						
Voting (MLP+CB+RF)	Soft Voting	0.7692	0.7661	0.8214	0.7083	0.7980
Stacking (MLP+CB+RF→LR)	Stacking	0.7404	0.7374	0.7857	0.6875	0.7954
Voting (MLP+CB+RF+XGB)	Soft Voting	0.7212	0.7199	0.7321	0.7083	0.7987
Stacking (RF+MLP+XGB+CB→LR)	Stacking	0.7308	0.7292	0.7500	0.7083	0.7950

Table 4.3: All Classifier Alternatives Comparison (Validation Set)

4.2.2.9 Model Selection and Justification

A Stacking ensemble (MLP + CatBoost + RF → Logistic Regression) was chosen to be the main model to be evaluated in the final tests. The Stacking architecture was chosen because of three reasons: the Voting ensemble has a better validation accuracy and Macro F1-Score.

First, the Stacking meta-learner fits a weighted average of base model predictions using the held-out data, so it is not as sensitive to the strengths and weaknesses of individual base models as equal-weight voting. Second, the Stacking set-up had a smaller recall gap (0.098 vs. 0.113), which is clinically significant in a depression screening scenario where the cost of false laboratory classification (false negative) of a depressed person is greater than the false alarm. Third, the Logistic Regression meta-learner has only two new parameters in each class, which means that the ensemble is resistant to overfitting at the combining phase despite the small training set.

The last model is based on three base classifiers (MLP, CatBoost, RF), which are trained on 195 standardized features that are sampled out of a 755-dimensional multimodal feature space. The predictions of the base models are obtained using 5-fold stratified cross-validation to obtain unbiased meta-features and are trained on these out-of-fold predictions using the Logistic Regression meta-learner. The training fold only will be fitted with feature standardisation and selection, so that no information in the test fold can affect any step of training or evaluation.

4.3 Data Collection

The current study is based on two datasets that have different purposes in the pipeline. The facial expression recognition models were trained and tested on the OAHEGA dataset, and depression classification was done on the D-Vlog dataset. All datasets were gathered in various ways, with different purposes, and a specific preprocessing pipeline dedicated to them. Each of the datasets and the entire data preparation process is described in the

following subsections.

4.3.1 OAHEGA Dataset

The OAHEGA dataset is a publicly available dataset of facial expression images that includes images of six emotions: happy, angry, sad, neutral, surprise, and ahégao. In this case, only five classes were utilized: happy, angry, sad, neutral, and surprise since ahégao is not a clinically relevant expression to detect depression. The pictures displayed are RGB photographs in the form of cropped faces, which are gathered in several sources such as Facebook, Instagram, YouTube videos, and preexisting datasets like IMDB and AffectNet. The data were obtained directly from the publicly available repository without the need to contact the author.

4.3.1.1 Data Cleaning

A complete cleaning pass was performed on the raw OAHEGA data prior to any balancing and augmentation. The choice of cleaning, then considering the imbalance in the classes, was not accidental: by clearing the low-quality images, the imbalance decisions that follow would be made on a sound number of actually usable images per class.

There was a three-step sequential cleaning pipeline on each image. The initial phase tested the fundamental aspects of image quality with CPU-based tests prior to calling any neural network. Blur was calculated with the Laplacian variance technique, in which the Laplacian operator calculates the second spatial derivative of pixel intensity. Sharp images have high variance, whereas blurry images have low variance. Pictures whose Laplacian variance was less than 100 were marked as being too blurry. Any image with a mean grayscale brightness less than 40 was marked as too dark.

The second phase filtered out images that passed the MTCNN quality test, a Multi-task Cascaded Convolutional Network that identifies faces and five facial landmark points in three stages via a Proposal, Refinement, and Output cascade of networks. The choice of MTCNN was based on Haar Cascade, HOG+SVM, RetinaFace, or MediaPipe, which offer the best trade-offs between accuracy, landmark output, sensitivity to occlusion, and simplicity in a preprocessing pipeline. The images were discarded when no face was found, when more than one face was found, when the confidence of detection was lower than 0.95, or when the area of the detected face took less than 10% of the total image area. It was implemented in facenet-pytorch to ensure that the entire pipeline would use a single PyTorch backend without requiring TensorFlow.

The third phase used geometric landmark validation on images that had passed the face detection. The five landmarks had to be within the image limits. Eye position had to be higher than the average mouth position. The nose was required to be vertically between the eyes and the mouth. The distance between the eyes had to be at least 5% of the image width, and the width of the mouth at least 3% of the image width. This check was created to be sensitive to instances where a part of the face is covered by a hand or object, and MTCNN still gives a high-confidence detection.

Out of 14,628 total raw images, 4,912 were removed, a removal rate of 33.6%, leaving 9,716 valid images. The most prevalent removal reason was blur with 4,333 images

or 88.2% of all removals. The remaining removals were because of a landmark out of bounds (332 images), no face (102) detected, multiple faces (89) detected, low detection confidence (38), too dark (17), and too small (1) face. The rates of removability were 26.2% angry, 28.3% happy, 35.8% neutral, 42.2% sad, and 25.8% surprise, with the most removed being sad since the sad faces expressed subtler and lower intensity expressions, therefore, creating more blurry or ambiguous images.

Class	Total Images	Removed	Removal Reason Breakdown	Images Remaining
Angry	1,403	367	Blur: 323, Landmark: 29, No Face: 8, Multiple: 5, Confidence: 2	1,036
Happy	4,067	1,149	Blur: 1,016, Landmark: 83, No Face: 26, Multiple: 20, Confidence: 4	2,918
Neutral	4,068	1,455	Blur: 1,282, Landmark: 119, No Face: 31, Multiple: 18, Confidence: 5	2,613
Sad	3,832	1,616	Blur: 1,424, Landmark: 95, No Face: 29, Multiple: 43, Confidence: 25, Dark: 0	2,216
Surprise	1,258	325	Blur: 288, Landmark: 6, No Face: 8, Multiple: 3, Confidence: 2, Dark: 17, Small: 1	933
Total	14,628	4,912	Too Blurry: 4,333; Landmark Out of Bounds: 332; No Face: 102; Multiple Faces: 89; Low Confidence: 38; Too Dark: 17; Face Too Small: 1	9,716

Table 4.4: OAHEGA Cleaning Summary

4.3.1.2 Data Transformation

The 9,716 images were cleaned and fed to a five-stage preprocessing pipeline to generate geometrically consistent, illumination-normalized, expression-focused face images at 224x224 resolution to be used in model training.

Stage 1 used MTCNN five-point landmarks to apply face alignment. To align the two eyes to a straight horizontal line at canonical anchor points, a rotation was calculated such that the left eye is at 35% and the right eye is at 65% of the image width and 38% of the image height.

The scale was applied to the face in such a way that the inter-ocular distance would be equal in all images. Bilinear interpolation was used to warp the output to a 224×224 size. In case MTCNN was not able to identify landmarks, a fallback bounding box crop with 25% margin and a square pad was applied instead.

The fallback face crop and resize path was stage 2 and used MTCNN bounding box detection with 25% margin added to all sides and square-padded to retain aspect ratio, then resized to 224×224.

Stage 3 employed the illumination normalization in two steps, utilizing the LAB colour space. CLAHE was used on the L lightness channel, clip limit was set to 2.0, and tile grid was 8×8 to enhance local contrast without altering the skin tone or skin color.

The L channel was then differentiated with sigma values of 1.0 and 2.0 to eliminate coarse illumination shading with the difference of Gaussians with retinex-style normalization. The fixed L channel was again combined with the original A and B channels and transformed once again to RGB.

Stage 4 used dlib 68-point landmark detection to apply facial region of interest masking to construct a convex hull of the entire facial area.

A binary mask was made by filling the convex hull, feathered with the Gaussian blur with sigma 15 to form soft edges. The mask was used as a blend with the face area at full strength and the background at a neutral grey of 15%, silencing the hair, ears, and neck, without adding hard artificial edges.

Stage 5 relied on a post-pipeline quality checking by recomputation of the Laplacian variance on the resulting 224×224 result.

4.3.1.3 Data Integration

The data used was the OAHEGA database, which was one unified source without the integration of data with other external datasets. The application of all five expression classes was based on the same data set to ensure consistency in image style, resolution, and collection conditions. No extra datasets were included in the training set.

4.3.1.4 Data Reduction

After preprocessing, there was still a lot of class imbalance in the retained images. Happy counted 2,515 pictures, as compared to 770 images of surprises. To counter this, offline augmentation was used to get all the underrepresented classes to 2,500 images.

Ten expression-safe augmentation techniques were used: horizontal flip, random rotation up to ± 12 degrees, brightness shift up to $\pm 30\%$, contrast shift up to $\pm 30\%$, Gaussian noise, random erasing, unsharp mask sharpening, zoom crop between 5% and 15%, horizontal shift up to $\pm 8\%$, and gamma correction between 0.75 and 1.25.

Class	Count Before Augmentation	Images Generated	Final Count After Augmentation
Happy	2,515	0	2,515
Neutral	2,275	225	2,500
Sad	1,925	575	2,500
Angry	907	1,593	2,500
Surprise	770	1,730	2,500
Total	8,392	4,123	12,515

Table 4.5: - OAHEGA Class Distribution

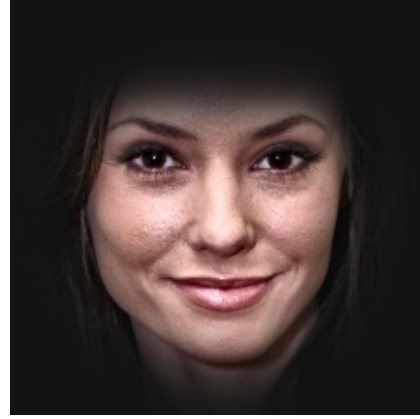
4.3.1.5 Summary of Preprocessed OAHEGA Data

The last OAHEGA data to be used in FER model training were around 12,515 images that were classified into five classes of expressions: 2,500 angry, 2,515 happy, 2,500 neutral, 2,500 sad, and 2,500 surprised. Images were processed to 224×224 , illuminated, face-aligned, and masked to remove the background. Stratified sampling with a random fixed seed of 42 was used to divide the dataset into 70% training, 15% validation, and 15% test sets containing about 8,761 training, 1,877 validation, and 1,877 test images, respectively.

This pretrained dataset was used to train and test all single FER models and the final hybrid ensemble.



(a) Original facial image



(b) Preprocessed facial image

Figure 4.4: Example of an original facial image and its corresponding preprocessed image.

4.3.2 D-Vlog Dataset

D-Vlog data is a multimodal depression dataset that initially featured 961 vlogs, amounting to about 160 hours of video recordings on YouTube, recorded between January 2020 and January 2021. Nevertheless, when it was accessed, 756 videos had been uploaded because the other videos were either deleted from YouTube or set to private, a frequent phenomenon in web-scraped datasets. The authors of the paper D-Vlog: Multimodal Vlog Dataset to Depression Detection, published at AAAI 2022, created the dataset. Depression-related search keywords were used to gather videos, and four trained annotators were used to annotate them into depressed and non-depressed classes, based on a standardized process of annotation. The data cannot be found publicly and was received by the authors via a formal email request. The entire dataset has 756 videos with a mean of 9.9 min and a median of 7.9 min, which totals around 159 hours of video footage. The data set has both audio and visual modalities. In the paper, facial expression features were extracted using the visual modality, and pre-extracted acoustic features that were provided with the data set were used to classify depression using multimodal features.

4.3.2.1 Data Cleaning

The raw version of the D-Vlog dataset comprised 756 videos that were sent in three folds, namely: 508 training videos, 168 test videos, and 80 validation videos, and a total of 392 depression and 364 normal videos. As the data is real-world vlog video, and not a controlled laboratory capture, the visual quality of videos differs significantly, with motion blur due to handheld video recording, low-light scenes, obscured parts of the face, and scenes where the face is not visible at all.

Each video was cleaned by extracting frames at a rate of 1 frame per second using OpenCV.

A facial preprocessing pipeline that operates on five stages was then applied to each extracted frame. The MTCNN detection parameters were tuned to vlog-style videos instead of the tuned parameters used on the OAHEGA dataset: the smallest face size was reduced by half to 20 pixels, the detection thresholds were set to 0.5, 0.6 and 0.7 instead

of the default 0.6, 0.7 and 0.9, and a confidence filter of 0.95 was introduced to eliminate false positives like hands and phone screens in the camera view.

The quality of the blur had also been reduced to 30 instead of 80 in order to support the motion blur that commonly occurs in handheld vlog recordings.

The videos whose frames could not be extracted at all after the whole pipeline was deleted. Of the 53 videos that made zero kept frames that were eliminated were 13 depression videos and 40 normal videos overall folds. These were mostly footage that had no face visible, such as videoing where the subject looked down at their phone, or turned the camera not towards them, or videoing in total darkness. This left 703 viable videos that had at least one valid preprocessed frame after filtering away these videos.

Fold	Class	Raw Videos	Discarded (0 frames)	Usable Videos
Train	Depression	267	8	259
Train	Normal	241	24	217
Train	Total	508	32	476
Validation	Depression	37	3	34
Validation	Normal	43	7	36
Validation	Total	80	10	70
Test	Depression	88	2	86
Test	Normal	80	9	71
Test	Total	168	11	157
Overall	Depression	392	13	379
Overall	Normal	364	40	324
Overall	Total	756	53	703

Table 4.6: D-VLOG Video Counts by Fold and Class

4.3.2.2 Data Transformation

All the retained frames went through the same 5-stage pipeline as the OAHEGA dataset, with the modification of the detection outlined above. Stage 1 used MTCNN five-point landmark-based face alignment using affine transformation warping to 224×224 pixels, having eyes at fixed canonical points at 35% and 65% of image width and 38% of image height, respectively.

Stage 2 offered a bounding box crop fallback with 25% margin on each side to all the frames with unsuccessful landmark detection, square-padded to maintain aspect ratio, and then resized to 224×224 . Stage 3 used CLAHE illumination normalization only on the LAB lightness channel with a clip limit of 2.0 and an 8×8 tile grid, and then removed Difference of Gaussians shading with a sigma of 1.0 and 2.0 with a blending weight of 0.5, blending the result back in place of the original color channels.

Stage 4 used MediaPipe FaceMesh and a mask of the areas of interest (facial) and 468 landmarks to create a convex hull around the entire facial area, feathered with a Gaussian

blur of sigma 15, and the background was replaced with neutral grey at a 15% density. MediaPipe was used in place of dlib from the OAHEGA pipeline as it provides a denser 468-point landmark model without requiring native C++ dependencies.

Stage 5 used a Laplacian variance quality measure with a threshold of 30 and removed frames that were still too blurry after processing.

A total number of 305,352 frames were kept out of 413,382 total frames analyzed in all of 756 videos, yielding a total keep rate of 73.9%. All frames that were retained were stored as JPEG images of quality 95 at a resolution of 224×224 . The time spent on processing the image with the NVIDIA GeForce RTX 4080 SUPER, which performed MTCNN inference, was 317.3 minutes.

Fold	Class	Frames Evaluated	Frames Kept	Frames Skipped	Keep Rate
Train	Depression	162,085	125,088	36,997	77.2%
Train	Normal	125,821	84,590	41,231	67.2%
Train	Total	287,906	209,678	78,228	72.8%
Validation	Depression	21,023	18,031	2,992	85.8%
Validation	Normal	16,418	10,020	6,398	61.0%
Validation	Total	37,441	28,051	9,390	74.9%
Test	Depression	53,210	44,363	8,847	83.4%
Test	Normal	34,825	23,260	11,565	66.8%
Test	Total	88,035	67,623	20,412	76.8%
Overall	Depression	236,318	187,482	48,836	79.3%
Overall	Normal	177,064	117,870	59,194	66.6%
Overall	Total	413,382	305,352	108,030	73.9%

Table 4.7: D-VLOG Frame Statistics by Fold and Class (703 Usable Videos)

4.3.2.3 Data Integration

The D-Vlog dataset offers two modalities to be used in this research. The preprocessed face frames obtained by the above descriptions are the visual modality. The acoustic modality is comprised of 25-dimensional acoustic feature vectors pre-extracted by the original dataset authors and made available with the videos, which include low-level vocal features related to depression, such as pitch, energy, and speech rate features. The multimodal depression classifier involved both modalities, with visual features provided by the emotion probability sequences of the FER ensemble and acoustic features obtained by temporal aggregation of the vectors provided. No external data was combined with D-Vlog in the task of depression classification.

4.3.2.4 Data Reduction

Following the initial preprocessing that resulted in 703 usable videos, one additional filter was used that was specifically aimed at developing the depression classifier. The

videos that had fewer than 10 frames after processing were not used since the statistical and time-related features calculated on fewer than 10 frames are not regarded as reliable for recording any meaningful changes in the expression. This filter eliminated another 36 videos, and 667 videos were used to develop and evaluate classifiers.

Following this filter, there was a minor imbalance in classes between depression and normal classes, with 366 depression and 301 normal videos, and an imbalance ratio of 1.22:1. This imbalance was not overcome by oversampling or undersampling, but by using balanced class weighting during the training of the classifier, which puts a larger penalty on a loss to the minority class proportional to its underrepresentation. The original D-Vlog train and validation folds were combined into one development pool and re-stratified with stratified 80/20 sampling to maintain the same balance of classes. The test fold was withheld completely and was only employed to give a final test to give an objective estimate of the generalization performance.

Split	Depression	Normal	Total	Depression %	Normal %
Train	225	189	414	54.3%	45.7%
Validation	56	48	104	53.8%	46.2%
Test (held-out)	85	64	149	57.0%	43.0%
Total	366	301	667	54.9%	45.1%

Table 4.8: Final D-VLOG Dataset Split (667 Videos)

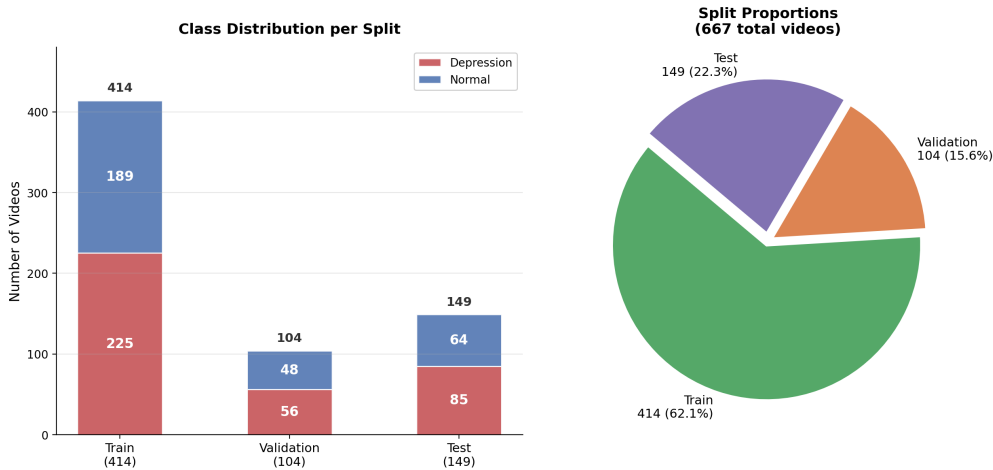
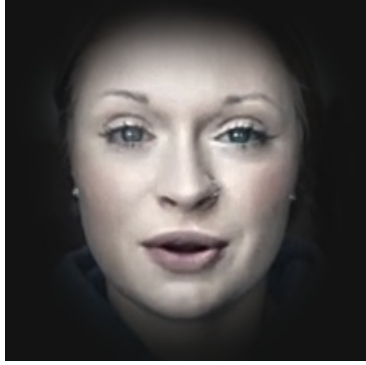


Figure 4.5: Class Distribution per Split

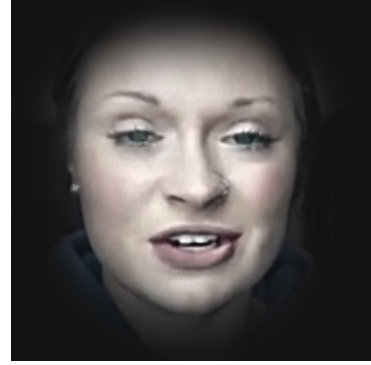
4.3.2.5 Summary of Preprocessed D-Vlog Data

The preprocessed and frame-filtering D-Vlog dataset that was used to develop a depression classifier comprised 667 videos. The training set included 414 videos, 225 depression, and 189 normal; the validation set included 104 videos, 56 depression, and 48 normal; and the held-out test set included 149 videos, 85 depression, and 64 normal. The total distribution of classes was 54.9% and 45.1% depression and normal, respectively. The entire preprocessed frame pool of all 703 usable videos comprised 305,352 frames at

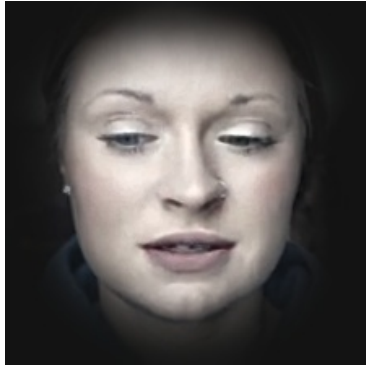
224×224 pixels stored in JPEG format with a quality of 95, which was sampled at 1 frame per second. The depression classifier visual features were obtained by using per-frame emotion probability sequences generated by the frozen FER ensemble, and the acoustic features were obtained by using the 25D acoustic vectors pre-extracted from the dataset.



(a) Frame extracted at 1 second



(b) Frame extracted at 2 seconds



(c) Frame extracted at 3 seconds



(d) Frame extracted at 4 seconds

Figure 4.6: Four consecutive frames extracted from a D-Vlog video at a sampling rate of one frame per second.

These combined to make the multimodal feature set with which binary depression was classified.

4.4 Implementation of Selected Design

The proposed system is implemented based on the two-component structure that was set during the design stage. The hybrid FER model is the first component and is completely implemented and used as a frozen feature extractor on the pipeline. The second component, the depression classifier, works on the multimodal features created based on the FER ensemble results and pre-extracted acoustic descriptors. Section 4.4.1 explains how the hybrid FER model is implemented and what its role in the inference pipeline is. Section 4.4.2 explains the implementation of the depression classifier with multimodal feature extraction, feature selection, classifier training, and final evaluation on the D-Vlog test fold.

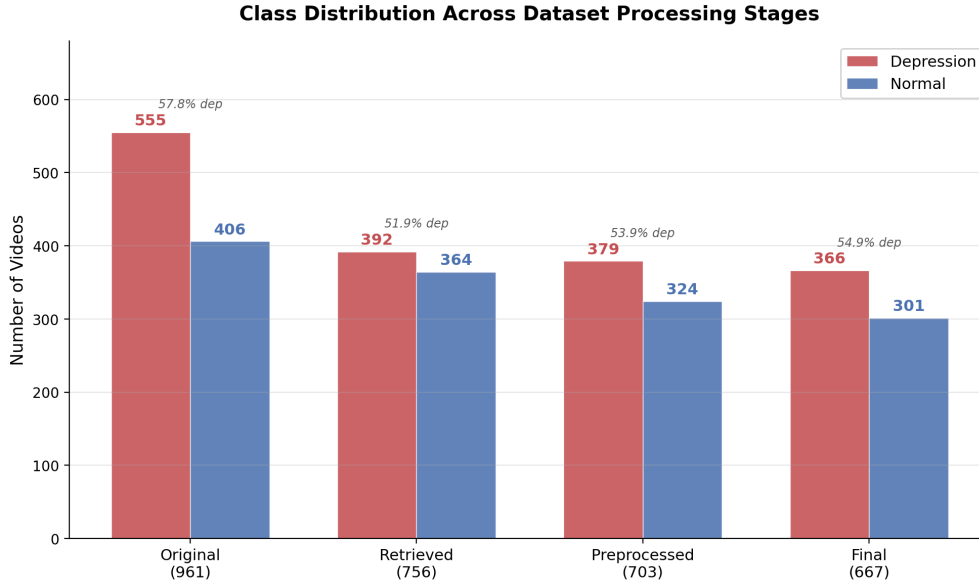


Figure 4.7: Class Distribution Across Dataset Processing Stages

4.4.1 Hybrid FER Model Implementation

The hybrid FER model is used as a weighted soft-voting combination of five independently trained deep learning models, including ConvNeXt-Small, CoAtNet-2, ViT-Base/16, Swin-Small, and ConvNeXt-Base. All the models were trained independently on the preprocessed OAHEGA data as indicated in Section 4.2.1, and their highest checkpoints were stored in terms of validation macro F1 score. In ensemble inference, the five models are loaded in the evaluation mode, and the weights are frozen. No additional training or fine-tuning occurs in the ensemble stage.

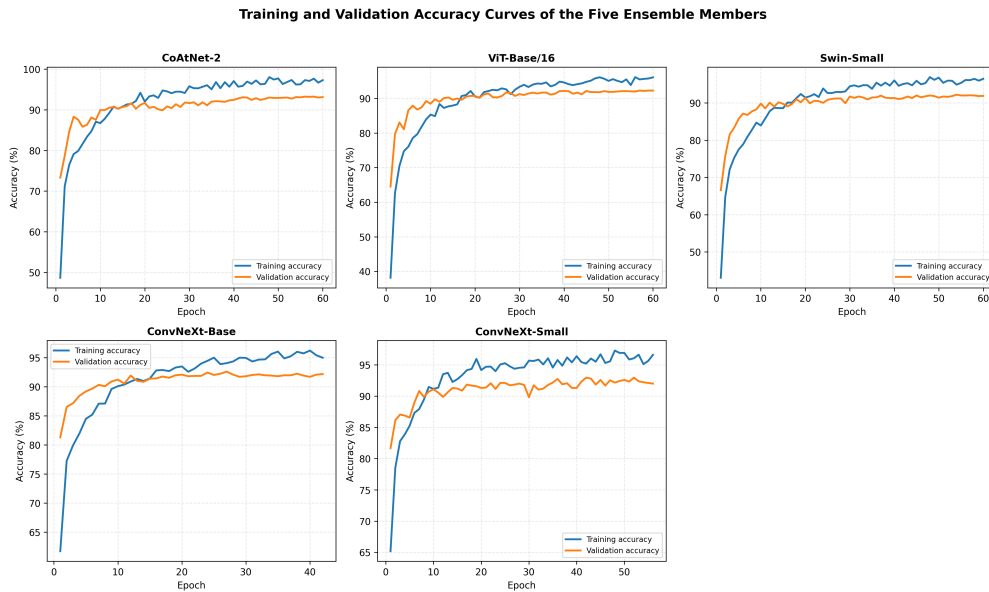


Figure 4.8: Training and Validation accuracy curves of five ensemble members.

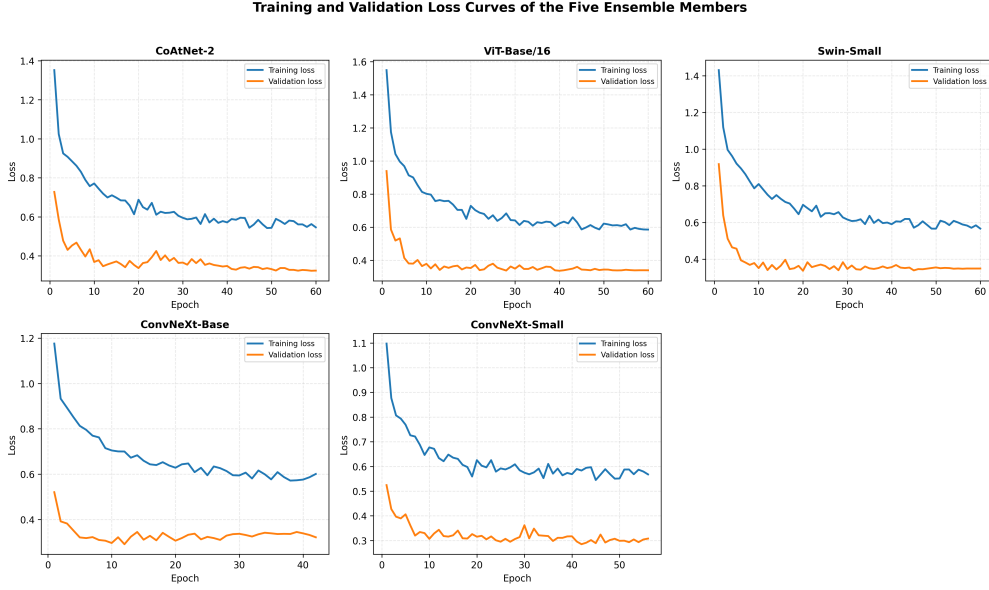


Figure 4.9: Training and Validation loss curves of five ensemble members.

Each model takes a 224×224 face image as input and output a five-dimensional output vector, the softmax probability distribution over the five face expression classes: angry, happy, neutral, sad, and surprise. Probability vectors are soft-voted, and the contribution of each model is weighted by its validation F1 score and then normalized. The weights used are 0.9322 for CoAtNet-2, 0.9318 for ConvNeXt-Small, 0.9254 for ConvNeXt-Base, 0.9222 for ViT-Base/16, and 0.9217 for Swin-Small. Normalization of these weights is done to add up to one, and the weighted average probability vector is calculated. The argmax of this weighted average vector is the final predicted expression class of a single image. The ensemble attained a test accuracy of 93.13% and a macro F1 of 0.9315 on the held-out OAHEGA test set, marginally outperforming the best single model by 0.64% points.

The hybrid FER model does not work with single images but on video frame sequences in the context of the entire depression detection pipeline. OpenCV is used to sample each D-Vlog video at 1 frame/second. All sampled frames are then processed by the identical five-stage preprocessing pipeline as OAHEGA training - MTCNN five-point landmark alignment to 224×224 , CLAHE illumination normalization, Difference of Gaussians shading, MediaPipe FaceMesh ROI masking, and Laplacian variance quality verification— to make sure that the frames sent to the FER model are similar to the distribution it was trained on. Frames where no face has been identified, or where the sharpness is below the sharpness threshold, are discarded. The result of each video is thus a variable number of clean, preprocessed face frames.

Each frame is separately fed to the frozen FER ensemble, which generates a five-dimensional emotion probability vector in a frame. When these vectors are concatenated over all correct frames of a video, it forms a sequence of shape $T \times 5$, where T is the number of retained frames, and each row is the probability distribution over the five expression classes of that frame. Videos with fewer than 10 valid frames are filtered out as bad, because any statistical and temporal characteristics based on such few frames cannot be useful in downstream classification.

$T \times 5$ sequence of emotion probability is then fed to the multimodal feature aggregation phase in Section 4.4.2, where they are then combined with acoustic features to constitute the overall 755-dimensional feature representation that is utilized in the classification of depression. During the depression detection stage, the FER ensemble is never retrained on the D-Vlog data, and instead, it is kept frozen at all times. The design maintains the two components distinct and makes the FER model run in-distribution since the D-Vlog frames are subjected to the same preprocessing pipeline when training OAHEGA.

4.4.2 Depression Classifier Implementation

The depression classifier is realized as a multimodal stacking ensemble, which works on the feature vectors obtained after two modalities, such as visual dynamics of facial expression mined by the frozen hybrid FER ensemble, and acoustic features mined by the given acoustic descriptors of the D-Vlog data. This is implemented in five successive steps, which include the extraction and preprocessing of frames, the extraction of emotion probabilities, multimodal feature aggregation, feature selection, and classifier training and evaluation.

$$w_k = \frac{F1_{\text{val}}(m_k)}{\sum_{j=1}^5 F1_{\text{val}}(m_j)}, \quad k = 1, \dots, 5. \quad (4.5)$$

$$\mathbf{p}_{\text{frame}} = \sum_{k=1}^5 w_k \mathbf{p}_k(\text{frame}). \quad (4.6)$$

$$\mathbf{P} \in \mathbb{R}^{T \times 5}, \quad \mathbf{P}[t, :] = \mathbf{p}_{\text{frame}}(t), \quad t = 1, \dots, T. \quad (4.7)$$

4.4.2.1 Frame Extraction and Preprocessing

OpenCV was used to sample each D-Vlog video at a rate of one frame/second. Each sampled frame was processed via the same five-stage preprocessing pipeline as OAHEGA training - MTCNN five-point landmark alignment to 224×224 , CLAHE illumination normalization, Difference of Gaussians shading removal, MediaPipe FaceMesh ROI masking, and Laplacian variance quality verification - so that the frames being fed to the frozen FER ensemble are in distribution to what it is trained on. The detection thresholds were also reduced as compared to the OAHEGA pipeline to accept handheld vlog footage as outlined in Section 4.3.2. Frames where a face had not been found or were below the sharpness threshold were removed. Videos whose total number of usable frames was less than 10, following preprocessing, were not considered at all, and 667 videos were left to develop the classifiers.

4.4.2.2 Emotion Probability Extraction

All retained preprocessed frames were run through the frozen five model weighted ensemble FER model. Each 224×224 face image was sent to the ensemble, and a five-dimensional softmax probability distribution over the expression classes, which are angry,

happy, neutral, sad, and surprise, was generated using the weighted soft-voting process described in Section 4.4.1. The ensemble weights and model checkpoints were held constant and remained constant during the depression detection phase. Joining the per-frame probability vectors of all possible valid frames of a video generated a sequence of shape $T \times 5$, where T is the count of retained frames, and each line is the emotion distribution of probability of that frame. This $T \times 5$ sequence is the visual input to the stage of feature aggregation.

4.4.2.3 Multimodal Feature Aggregation

All the variable-length $T \times 5$ sequences of emotions were aggregated into fixed-length feature vectors via a full aggregation process that included three categories of features.

$$\mathbf{x} = \left[\mathbf{x}_{\text{visual}}^{(250)} \parallel \mathbf{x}_{\text{acoustic}}^{(500)} \parallel \mathbf{x}_{\text{crossmodal}}^{(5)} \right] \in \mathbb{R}^{755}. \quad (4.8)$$

The $T \times 5$ sequences were computed to give visual features totalling 250 dimensions, basic statistics such as mean, standard deviation, minimum, maximum, skewness, and kurtosis by emotion class, clinical measures like affect entropy, emotional transition rate, anhedonia ratio, and flat affect proportion, approximations of AU including blunted affect ratio, anhedonia duration, emotional flatness, positive deficit, and entropy slope, percentile statistics on the 5th, 25th, 50th, 75th, 95th percentiles by emotion, autocorrelation at lags 1 to 4 by emotion, temporal progression features such as first-half and second-half differences, three-segment means and standard deviations, segment to segment transitions, four-quarter means, and pairwise emotion correlation, zero crossing rate, dominant emotion proportions, and emotional range.

The acoustic descriptors were extracted to 25 dimensions, and the 500 acoustic features were calculated using the acoustic descriptors it offered in the D-Vlog dataset. Temporal aggregation resulted in distributional statistics (mean, standard deviation, minimum, maximum, median, range, skewness, kurtosis, interquartile range) and temporal dynamics statistics (lag-1 autocorrelation, linear slope, first-half-second-half means, half difference, three-segment means, approximate entropy, segment drift, and segment variability) on each of the 25 acoustic dimensions. These acoustic properties reflect vocal attributes of depression, such as decreased range of pitch, monotonicity, lack of energy, and vocal weariness.

Five cross-modal measures were also calculated as the Pearson correlation between each of the five emotion probability sequences and the mean over time acoustic feature vector representing the extent of alignment between vocal and facial emotional expression, which has been found to be impaired during depression. The total multimodal feature vector per video thus consisted of 755 dimensions that consisted of 250 visual, 500 acoustic, and 5 cross-modal features.

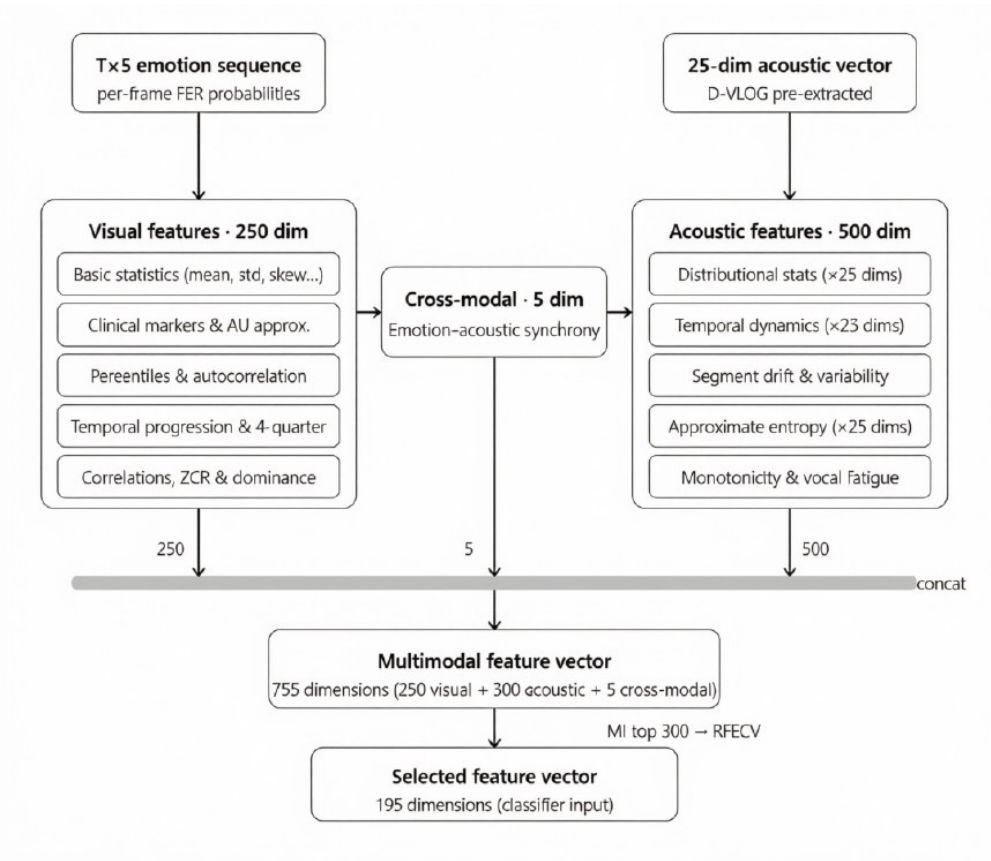


Figure 4.10: Feature aggregation diagram

4.4.2.4 Feature Selection

The 755-dimensional feature vectors were used in a two-level feature selection pipeline to downsample the data and eliminate noisy/redundant features prior to the training of the classifier. The initial step used Mutual Information scoring with a predetermined random seed of 42 to choose the 300 most important features according to the statistical dependence on the binary depression label. The second step used Recursive Feature Elimination with Cross-Validation with an estimator of a Random Forest with 100 trees, equal weights on the classes, 5-fold stratified cross-validation, and macro F1 as the scoring criterion, which further narrowed the feature set to 195 features. Both stages of selection were only fitted to the training fold of 414 videos to ensure that no data on the validation or test sets were leaked by the selection stages and that the same feature selection was achieved across all selection runs using the fixed random seed.

4.4.2.5 Classifier Training and Evaluation

The chosen stacking ensemble was trained using the 195-dimensional feature vectors of the 414 training videos with their binary depression labels. A StandardScaler was applied to the training fold, and all the features were standardized before being sent to the classifiers. The three base classifiers (Random Forest, MLP, and CatBoost) were trained on 5-fold stratified cross-validation to produce out-of-fold predictions, which were fed

into the Logistic Regression meta-learner. Random Forest and CatBoost used balanced class weighting to deal with the slight 1.22:1 class imbalance.

The MLP employed early stopping at a 10% validation rate to avoid overfitting. The meta-learner was trained on the stacked out-of-fold predictions of each of the three base classifiers and was trained to apply optimal weights on the output of each of the classifiers, based on their complementary properties.

The trained stacking ensemble was tested on the held-out validation fold of 104 videos to select models and make hyperparameter choices, and once on the held-out test fold of 149 videos to estimate the final unbiased performance. The test accuracy of the stacking ensemble on the test fold was 71.81%, precision of 0.7529, macro F1 of 0.7124, AUC-ROC of 0.7722, a depression recall of 0.7529, and a normal recall of 0.6719. The depression recall purposely overstates normal recalls, as it is more clinically desirable to minimize false negatives in depression screening, where the cost of missing a depressed individual is higher than the false alarm cost of a normal individual. The gap of 0.081 in the recalls can be explained by the equal weighting of classes used in training, which leads to the sensitivity of the model to the depression class. The entire merged inference pathway - which in turn takes raw video, through preprocessing, FER ensemble, feature aggregation, feature selection, and stacking classifier generates a final binary depressed or normal prediction of each input video along with a confidence rating.

Chapter 5

Result Analysis

5.1 Performance Evaluation

The proposed multimodal depression detection system is composed of two different components, which were evaluated separately with the proper metrics and testing procedures. The second part, the hybrid Facial Expression Recognition model, was tested on a held-out test set of 1877 images of the OAHEGA dataset. For the second, the depression classifier was tested on a separate test fold of 149 videos from the D-Vlog dataset. Stratified splits with fixed random seeds were used for both components to guarantee the reproducibility of the tests, and the final tests for the evaluation of both components were carried out only once after finalizing all the design and hyperparameter choices on the validation data.

5.1.1 FER Model Performance Evaluation

Four basic metrics were used to assess the hybrid FER ensemble. The accuracy of testing is the percentage of test expression samples correctly classified in all five classes, and it is a measure of the overall classification performance. Weighted F1 was chosen as the main performance metric instead of macro F1 since the dataset was balanced offline with approx 2,500 images per class, resulting in near-identical values of the two metrics. This is particularly relevant as the neutral and sad classes are clinically relevant, and macro F1 treats all classes equally, regardless of their support, making them more sensitive to their performance degradation. Additionally, per-class precision, recall, and F1 scores were reported to provide a better understanding of the strengths and weaknesses of the model per class, which may not be covered by the overall scores and macro F1. In addition to the accuracy measure, a separate clinical measure of neutral-sad confusion errors was recorded, which reflected the total number of errors made in the classification of neutral or sad in a clinically neutral-sad confusion error. This measure was used for all individual model experiments, as the quality of the emotion probability sequences input to the downstream depression classifier is compromised directly by neutral-sad confusion, irrespective of the emotion classification accuracy of the model.

The ensemble of hybrid models was then evaluated on 1,878 images that were not used in training, and the same stratified test set (15%) was applied to all individual hybrid

models. No training or adaptation on the test set was performed on each of the 5 models of the ensemble, which were all loaded in frozen evaluation mode.

Class	Precision	Recall	F1-Score	Support
Angry	0.960	0.968	0.964	375
Happy	0.971	0.958	0.964	378
Neutral	0.843	0.888	0.865	375
Sad	0.901	0.869	0.885	375
Surprise	0.986	0.973	0.980	375
Test Accuracy	93.13%			1877
Macro F1	0.9315			
NS Errors	69			

Table 5.1: Hybrid FER Ensemble Test Metrics (5.1.1)

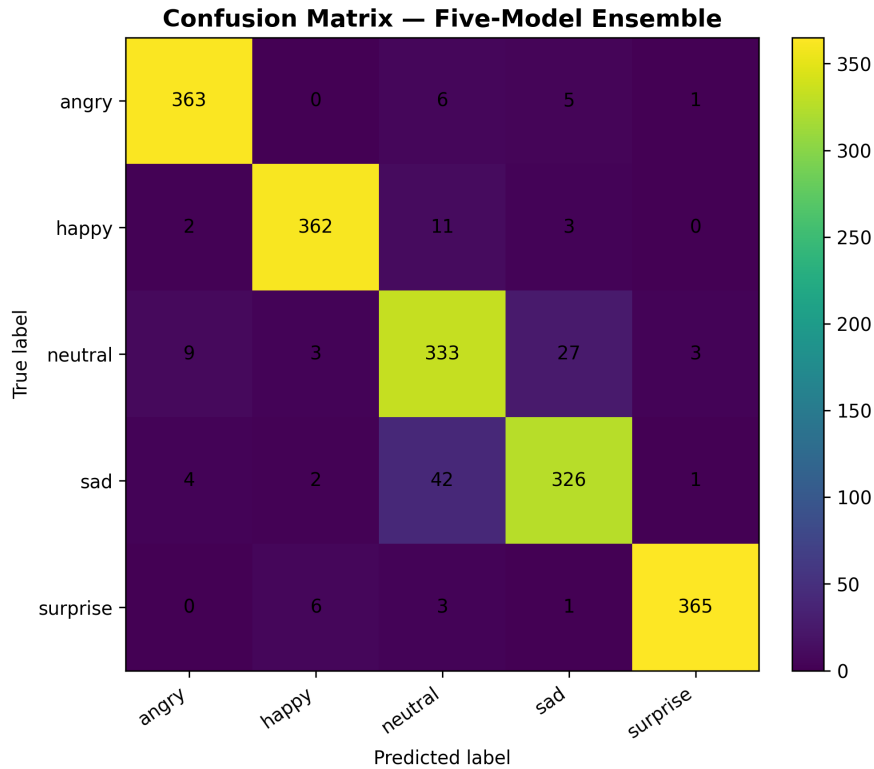


Figure 5.1: Hybrid FER ensemble confusion matrix heatmap.

5.1.2 Depression Classifier Performance Evaluation

Six test metrics were chosen to test the depression classifier, representing the binary classification task and its clinical context. The accuracy of tests is the percentage of videos that are correctly classified overall. Macro F1 score considers both classes equally and is

a balanced measure of the performance, taking into account both precision and recall. Depression F1 and depression recall were reported separately as primary clinical measures because the depression class is the class of interest: the depressed class is the class of interest, and with high depression recall, there are fewer and fewer depressed people who are incorrectly classified as the normal class, which would be the most costly error in a screening context. Normal recall was also reported to evaluate the false alarm rate: that is, the proportion of normal people who are falsely identified as depressed. AUC-ROC was added as a threshold-independent performance measure that provides an overall assessment of the model’s discriminative performance for all possible classification thresholds. The absolute difference between depression recall and normal recall, or ”recall gap,” was measured as a summary measure of the clinical compromise between sensitivity and specificity for the classification of asymmetry.

The four ensemble configurations were then tested on the held-out test fold, consisting of 149 videos (85 depression and 64 normal), without taking any part in training and model selection. Standalone classifiers such as Random Forest, MLP, CatBoost, XGBoost, TabNet, Bi-LSTM, and Bi-GRU were not evaluated on the test fold because they are only utilized in the component analysis and ensemble construction in the validation phase. To ensure that the same 195 features were selected for the test evaluation as during training, feature selectors were re-fitted with the same pipeline and on the same training fold, which consisted of 414 videos; no data leakage was introduced.

The best test accuracy on the test fold was obtained by the Voting Ensemble (AUC = 0.7785, test Accuracy = 70.47%, macro F1 = 0.6975, depression recall = 0.7529, normal recall = 0.6406), which combined MLP, CatBoost, and Random Forest. The Voting Ensemble with XGBoost had a test accuracy of 69.80%, a macro F1 score of 0.6924, a depression recall of 0.7294, a normal recall of 0.6562, and an AUC of 0.7728, with a recall gap of 0.0732. With a fourth base learner of XGBoost, the Stacking Ensemble gave a test accuracy of 70.47%, a macro F1 of 0.6998, depression recall of 0.7294, normal recall of 0.6719, and an AUC of 0.7704 with a recall gap of 0.0575.

The highest test accuracy 71.81%, and precision of 0.7529 was obtained by the Stacking Ensemble, which stacked MLP, CatBoost, and Random Forest with a Logistic Regression meta-learner, and also yielded the highest macro F1, 0.7124, and the highest AUC, 0.7722, with a lowest recall gap of 0.0811. It was chosen as the final depression classifier because of its highest overall accuracy, best macro F1, and lowest drop in accuracy when moving from validation to test sets; its drop was 2.23 percentage points compared to 6.45 percentage points in the voting (MLP+CB+RF) configuration, indicating that the overall depression prediction made by the meta-learner was more stable on unseen data than the fixed equal-weight voting configuration.

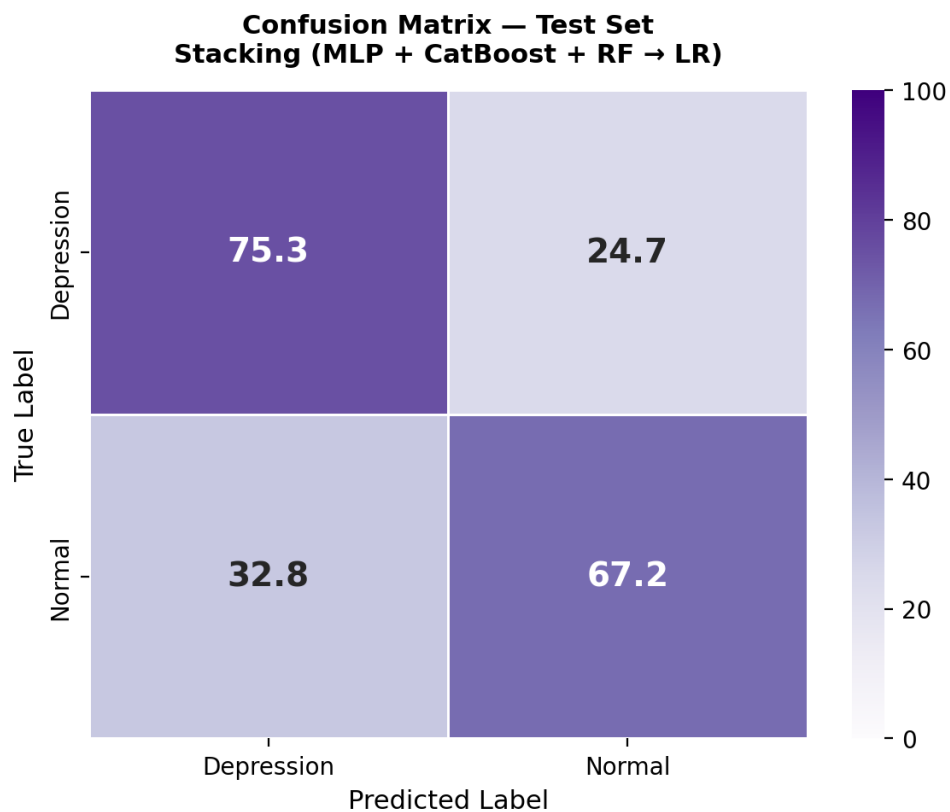
Model	Type	Test Acc	Macro F1	Dep F1	Dep Recall	Norm Recall	Recall Gap	AUC
Voting (MLP+CB+RF)	Soft Voting	70.47%	0.6975	0.7442	0.7529	0.6406	0.1123	0.7785
Voting (MLP+CB+RF+XGB)	Soft Voting	69.80%	0.6924	0.7337	0.7294	0.6562	0.0732	0.7728
Stacking (RF+MLP+XGB+CB→LR)	Stacking	70.47%	0.6998	0.7381	0.7294	0.6719	0.0575	0.7704
Stacking (MLP+CB+RF→LR)	Stacking	71.81%	0.7124	0.7529	0.7529	0.6719	0.0811	0.7722

Table 5.2: All Four Ensemble Configurations Final Test Results

A correctly diagnosed depressed patient was 64 of 85 (or 0.7529 or 75.3%), and a correctly diagnosed normal patient was 43 of 64 (or 0.6719 or 67.2%). The 21 false

Metric	Value
Model	Stacking (MLP + CatBoost + RF → Logistic Regression)
Test videos	149 (85 depression, 64 normal)
Test Accuracy	71.81%
Macro F1-Score	0.7124
Depression F1	0.7529
Depression Recall	0.7529
Normal Recall	0.6719
Recall Gap	0.0811
AUC-ROC	0.7722
Val-to-Test Accuracy Drop	2.23 percentage points (74.04% → 71.81%)

Table 5.3: Selected Model: Stacking (MLP + CatBoost + RF → LR) Detailed Test Metrics



Accuracy: 71.81% | Dep-Recall: 0.7529 | Norm-Recall: 0.6719 | Gap: 0.0811

Figure 5.2: Confusion Matrix: Stacking (MLP + CatBoost + RF → LR) on Test Fold

negatives are 24.7% of depressed individuals not classified by the classifier, and the 21 false positives are 32.8% of normal individuals incorrectly classified by the classifier. The depression recall is intentionally set to exceed normal recall because of balanced class weighting during training; this is in line with the clinical concept that missed diagnoses are to be kept to a minimum.

5.2 Analysis of Design Solutions

This section discusses the detailed analysis of all the design solutions considered for both subsystems of the proposed system. For the FER component, a hybrid ensemble of ten model architectures was trained and tested, and the final architecture was selected. The final stacking ensemble was chosen after evaluating various values of the multimodal feature set for seven individual classifiers and four ensemble configurations in the depression classifier component. The depression classifier component was evaluated for several values of the multimodal feature set using seven individual classifiers and four values of the ensemble configuration before selecting the final stacking ensemble. Each subsection discusses the performance, strengths, weaknesses, and contribution of each solution to the final design.

5.2.1 FER Model Analysis

For the facial expression recognition task, 10 models were tested across three architectural families of pure Convolutional, pure transformer, and hybrid CNN-attention. Test accuracy, macro F1, per-class F1, and neutral-sad confusion errors were used for the evaluation. The models are discussed in order of their performance.

Model	Architecture Family	Test Acc	NS Errors	Key Finding	Status
ResNet-50	CNN (2015)	87.43%	104	<i>Highest NS errors; no attention mechanism limits global context modeling</i>	Baseline
VGG-16	CNN (2014)	89.35%	101	<i>Lowest neutral F1 (0.792) despite largest parameter count (134.3M); no residual connections</i>	Baseline
EfficientNet-B7	CNN (compound)	90.36%	84	<i>Resolution constraint (224 vs native 600); asymmetric NS bias 24 vs 60 errors</i>	Excluded
MaxViT-T	Hybrid CNN+Attn	90.36%	89	<i>Highest NS errors among modern models; ImageNet-1k only pretraining insufficient</i>	Excluded
ConvNeXt-Base	Modern CNN	91.05%	70	<i>Lowest NS errors of any individual model; included for NS boundary strength not accuracy</i>	In ensemble
ConvNeXt-Small	Modern CNN	91.69%	87	<i>Only model without NS loss; asymmetric NS pattern 30 vs 57; strong ImageNet-22k transfer</i>	In ensemble
Swin-Small	Transformer	91.91%	75	<i>Hierarchical multi-scale features; complements global attention of ViT</i>	In ensemble
ViT-Base/16	Transformer	92.28%	78	<i>Only pure transformer; highest surprise F1 (0.975); global attention from first layer</i>	In ensemble
CoAtNet-2	Hybrid CNN+Attn	92.44%	74	<i>Best single model accuracy; C-C-T-T design bridges local and global features; ImageNet-21k</i>	In ensemble
5-Model Ensemble	Weighted soft-voting	93.13%	69	<i>Best accuracy and lowest NS errors; architectural diversity across three families</i>	Final model

Table 5.4: FER Model Comparative Analysis Summary

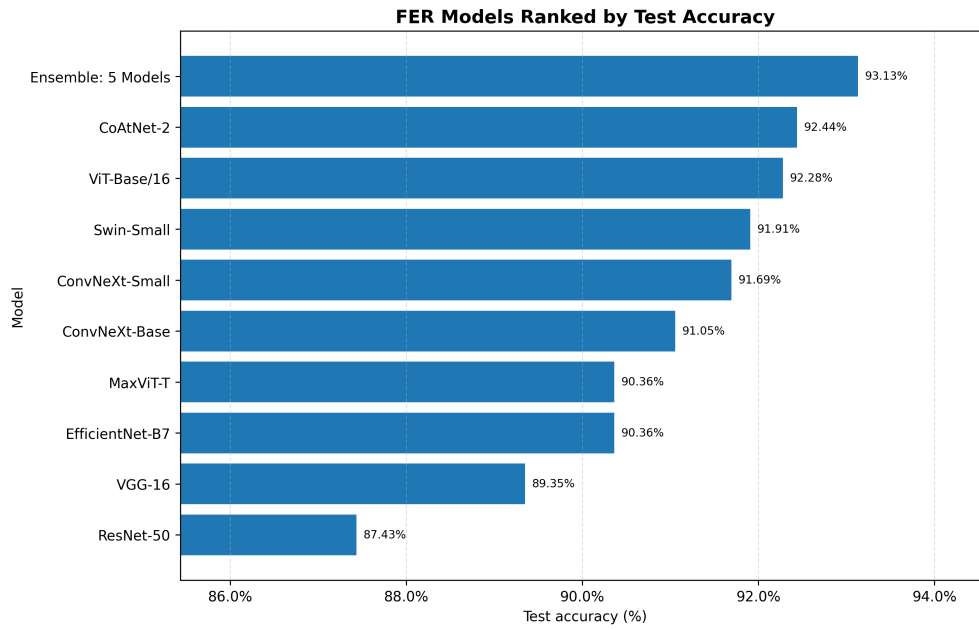


Figure 5.3: FER models ranked by test accuracy from lowest to highest.

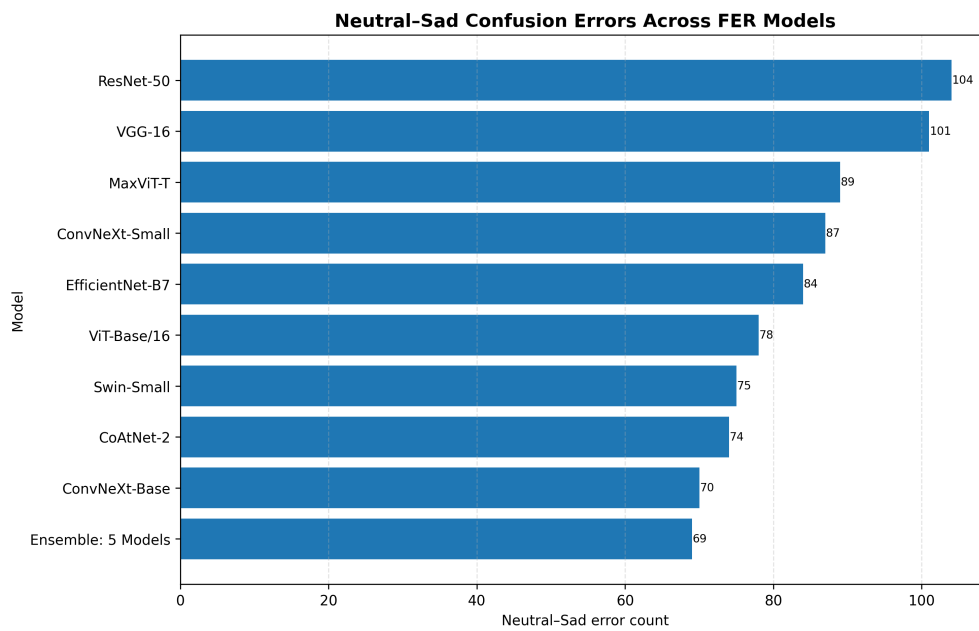


Figure 5.4: NS errors per model.

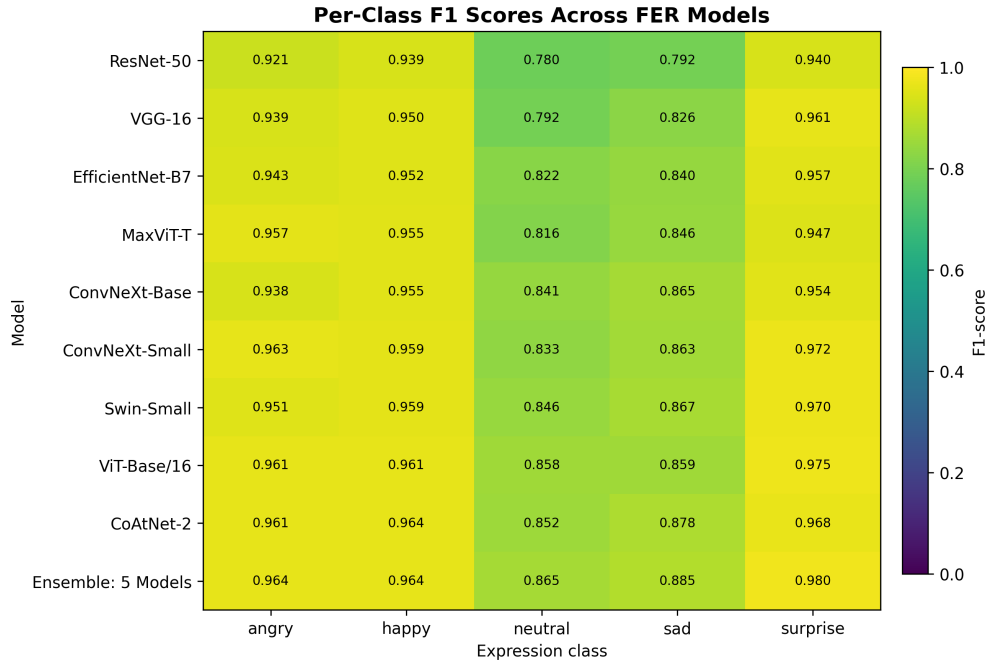


Figure 5.5: Per-class F1 scores across all FER models.

5.2.1.1 Classical Baselines

The pre-modern-transformer era of deep learning was represented by ResNet-50 and VGG-16, which served as classic CNN baselines. ResNet-50 had the lowest test accuracy of 87.43%, and a macro F1 of 0.8746, despite having the most neutral-sad confusion errors (104). We noticed that ResNet-50, although deep enough for residual connections, can only learn global facial context without attention mechanisms, limiting the FER performance. VGG-16 obtained an accuracy of 89.35% and a macro F1 of 0.8937, making 101 neutral-sad errors. Although the largest model evaluated (134.3 million parameters) it systematically underperforms all modern architectures. The lowest class was neutral with a neutral class (F1) of 0.792 for VGG-16, which failed to capture global contextual cues that differentiate neutral from lower intensity sadness when residual connections or attention mechanisms are not used. Both baselines show the significant difference between traditional and modern architectures, which justifies the focus on transformer and hybrid models for the ensemble.

5.2.1.2 Models Excluded from the Ensemble

Modern alternatives were evaluated, but not considered for the hybrid ensemble: EfficientNet-B7 and MaxViT-T. EfficientNet-B7 also made 84 neutral-sad errors and has an accuracy of 90.36% and a macro F1 of 0.9028. The main drawback was the limitation on resolution — it was originally supposed to accept a 600×600 resolution and, for the purpose of the competition, was limited to 224×224 , so that it could be tested at below its capacity while it was learning. The neutral-sad confusion resulted in a very asymmetric pattern of 24 neutral-to-sad errors and 60 sad-to-neutral errors, suggesting a pattern of systematic underprediction of neutral over sad that would negatively impact the down-

stream depression features. MaxViT-T had an accuracy of 90.36% and a macro F1 of 0.9042 with a total of 89 errors, the highest of the hybrid and transformer models. Although it has a multi-axis attention mechanism based on two attention modules, namely local window attention and global grid attention, these two attention modules were only jointly pretrained on ImageNet-1k, which also led to a higher proportion of confusion at the neutral-sad transition points in the data, and the confusion pattern was similarly asymmetric, with 28 neutral-to-sad and 61 sad-to-neutral errors, respectively. The two models were not included because they performed worse in terms of overall accuracy and neutral-sad confusion than the five models that were selected.

5.2.1.3 Ensemble Member Analysis

The 3D hybrid ensemble consisted of 5 models selected from 3 different families: ConvNeXt-Base, Swin-Small, ConvNeXt-Small, ViT-Base/16, and CoAtNet-2.

ConvNeXt-Base obtained an accuracy of 91.05% and a macro F1 of 0.9105. Its biggest asset to the ensemble is its neutral-sad confusion count of 70, which is the lowest of any individual model evaluated. This shows that the ConvNeXt-Base model is more discriminative in learning the neutral-sad boundary than any other single model, due to its larger capacity and the Neutral-Sad Separation loss at weight 0.25, in addition to the lower NS error criterion as a tiebreaker model. For this dataset size, the model capacity drops off after ConvNeXt-Small, with its overall accuracy 1.44 percentage points lower than ConvNeXt-Small, for the trade-off of being about half its model size. The reason why ConvNeXt-Base is added to the ensemble is that its NS boundary strength is strong, not necessarily the overall accuracy.

With 75 neutral-sad errors, Swin-Small’s accuracy was 91.91% and macro F1 was 0.9187. It represents the image features from multiple scales with a hierarchical shifted window attention mechanism that is performed in four progressive stages, and thus captures local facial texture details in the early stages, as well as broader expression semantics in the later stages. It is a hierarchical design, compared to the uniform global attention used by ViT-Base/16, to get additional feature representations for complementarity in ensemble diversity.

ConvNeXt-Small was able to reach 91.69% accuracy and a macro F1 of 0.9178, with 87 neutral-sad errors. It is the only ensemble member that was trained without the Neutral-Sad Separation loss, which accounts for the fact that it has a higher number of NS losses than its ensemble counterparts, and moreover, a much higher number of errors that are not only NS losses, but are distinctly asymmetric (30 NS, 57 SN). However, ConvNeXt-Small also makes meaningful convolutional feature representations based on the principles of transformer designs and offers pretraining on ImageNet-22k, which gives it more transfer features than variants trained on ImageNet-1k.

ViT-Base/16 had a training error rate of 78 for neutral-sad errors and a test accuracy rate of 92.28% and macro F1 of 0.9228. It is the only pure vision transformer model in the ensemble as it is able to learn long-range dependencies across the entire facial region from the initial layer, resulting in a distinct feature representation from the convolutional and hybrid models. It scored the top surprise class F1 of 0.975, with especially high scores for high-intensity expressions with distinct global facial geometry patterns.

CoAtNet-2 had the highest accuracy of 92.44% when tested on an individual basis and the highest macro F1 of 0.9246 with 74 errors of neutral-sad. It is based on the C-C-T-T architecture, which consists of two convolutional blocks (C-C) for local facial feature extraction and two relative self-attention transformer blocks (C-T-T) for global contextual modeling, and a natural connection between purely convolutional and purely transformer-based feature extraction. It has the highest number of parameters that have been pre-trained on the ImageNet-21k dataset of any of the models in the ensemble. CoAtNet-2 is the most robust model in the ensemble in terms of overall accuracy and performance in the NS boundary.

5.2.1.4 Hybrid Ensemble Analysis

The five-model weighted soft-voting ensemble model outperformed the best individual model, CoAtNet-2, with a test accuracy of 93.13% and a macro F1 of 0.9315, by 0.69% and 0.0069, respectively. The neutral-sad confusion error rate was lowered to 69, which is fewer than each individual model, such as ConvNeXt-Base, with the lowest individual NS error rate of 70. The ensemble’s overall performance is better than any one model, and is a testament to the importance of architectural variety in ensemble design. The three architectural families (pure convolutional, pure transformer, and hybrid CNN-attention) generate complementary error patterns, which leads to the errors from each model canceling the errors of the other models within the family that do not share them. If the best-performing single models, in terms of accuracy, were chosen without considering the diversity of architecture, the ensemble would have contained several models of the same architecture, which would lessen the benefit of diversity and would likely lead to a lower ensemble accuracy than the architecturally diverse configuration chosen.

5.2.2 Depression Classifier Analysis

The 755-dimensional multimodal feature vectors were selected by two stages to get 195 features to evaluate multiple classifier architectures for depression classification. Seven single classifiers were first tested on the held-out validation fold of 104 videos, and four combinations of classifiers were tested on the held-out test fold of 149 videos in addition to the validation fold. The criteria for evaluation were the validation accuracy, macro F1, depression recall, normal recall, AUC-ROC, and recall gap.

5.2.2.1 Individual Classifier Analysis

The performance of seven individual classifiers was assessed as individual classifiers prior to ensemble construction. Random Forest had the highest validation accuracy of 71.15%, macro F1 of 0.7076, depression recall of 0.7679, normal recall of 0.6458, and an AUC of 0.7950, among the tree-based and gradient boosting methods. This tree model is also stable to over-fitting on the small training set of 414 videos, and the feature importance outputs give direct interpretability about the emotion and acoustic patterns that are most predictive of depression. The results of CatBoost are: stand-alone validation accuracy is 67.31%, macro F1 is 0.6700, depression recall is 0.7143, and normal recall is 0.6250. The validation accuracy of XGBoost was also 68.27%, and its macro F1 was

Model	Type	Val Acc	Macro F1	Dep Recall	Norm Recall	AUC	Decision
Stage 1—Individual Classifiers							
Bi-GRU	Recurrent NN	0.6442	0.6442	0.6071	0.6875	0.7054	Excluded
Bi-LSTM	Recurrent NN	0.6923	0.6905	0.7143	0.6667	0.6901	Excluded
CatBoost	Gradient Boost	0.6731	0.6700	0.7143	0.6250	0.7716	Base classifier
XGBoost	Gradient Boost	0.6827	0.6812	0.6964	0.6667	0.7798	Excluded
TabNet	Attention Tabular	0.7115	0.7106	0.7143	0.7083	0.7150	Excluded
Random Forest	Tree Ensemble	0.7115	0.7076	0.7679	0.6458	0.7950	Base classifier
MLP	Neural Network	0.7212	0.7190	0.7500	0.6875	0.8013	Base classifier
Stage 2—Ensemble Configurations Validation							
Voting (MLP + CB + RF + XGB)	Soft Voting	0.7212	0.7199	0.7321	0.7083	0.7987	Excluded
Stacking (RF + MLP + XGB + CB → LR)	Stacking	0.7308	0.7292	0.7500	0.7083	0.7950	Excluded
Voting (MLP + CB + RF)	Soft Voting	0.7692	0.7661	0.8214	0.7083	0.7980	Not selected
Stacking (MLP + CB + RF → LR)	Stacking	0.7404	0.7374	0.7857	0.6875	0.7954	Selected

Table 5.5: Iterative depression-classifier selection based on validation results.

Model	Test Acc	Macro-F1	Dep-F1	Dep-Recall	Norm-Recall	Recall Gap	AUC-ROC
Voting Ensemble (MLP + CatBoost + RF)	70.47%	0.6975	0.7442	0.7529	0.6406	0.1123	0.7785
Stacking (MLP + CatBoost + RF → LR)	71.81%	0.7124	0.7529	0.7529	0.6719	0.0811	0.7722
Voting Ensemble (MLP + CatBoost + RF + XG-Boost)	69.80%	0.6924	0.7337	0.7294	0.6562	0.0732	0.7728
Stacking (RF + MLP + XGBoost + CatBoost → LR)	70.47%	0.6998	0.7381	0.7294	0.6719	0.0575	0.7704

Table 5.6: Final test evaluation and comparison of depression-classifier ensemble models.

0.6812, depression recall was 0.6964, and normal recall was 0.6667. Even though CatBoost and XGBoost achieved worse individual performance, they both provide complementary gradient-boosted representations for ensemble methods.

For neural network approaches, the highest individual model validation accuracy was obtained by MLP with 72.12%, and the highest macro F1 score with 0.7190, with depression recall of 0.7500 and normal recall of 0.6875, and the highest individual AUC of 0.8013. It has three hidden layers of 256, 128, and 64 dimensions that can model non-linear interactions between visual, acoustic, and cross-modal features that are not captured by the tree-based methods. TabNet’s performance was validated with an accuracy of 71.15% and an AUC of 0.7150, a depression recall of 0.7143, a normal recall of 0.7083, and a macro F1 of 0.7106. The TabNet sequential attention mechanism achieved the best class balance among individual classifiers, having a recall gap of 0.006. Its performance was, however, lower than MLP and Random Forest in terms of the accuracy and AUC, which restricted it from being used as a main model.

Both the sequence models (Bi-LSTM and Bi-GRU) were implemented directly on the raw $T \times 5$ emotion probability sequences, without the handcrafted feature aggregation step. Bi-LSTM achieved the best performance with a validation accuracy of 69.23% and a macro F1 of 0.6905, while Bi-GRU had the lowest validation accuracy of 64.42% and the lowest macro F1 score of 0.6442. Both methods failed to perform as well as the classical and ensemble methods, as is typical with small training sets of only 414 videos, which is too small to train recurrent architectures to learn good temporal representations without overfitting. The values of AUC: 0.6901 for Bi-LSTM and 0.7054 for Bi-GRU further proved their poor discriminative power compared with feature-based methods, and did not go into further consideration.

5.2.2.2 Ensemble Configuration Analysis

Two combination strategies and two sets of base classifiers were used to evaluate four ensemble configurations. The voting method (equal weight soft voting) of MLP, CatBoost, and Random Forest resulted in the best validation accuracy of 76.92% and macro F1 score of 0.7661, where the depression recall was 0.8214, and the normal recall was 0.7083. On the held-out test fold, however, it saw a drop of 6.45 percentage points from validation, with only 70.47% accuracy and a macro F1 of 0.6975, indicating overfitting to the validation distribution. The highest performance was when it was recalled on the test, with a recall gap of 0.1123 (the largest asymmetry between depression and normal classification), and the lowest was when it was not recalled on the test, with a recall gap of 0.0413.

The validation accuracy of the stacking model with MLP, CatBoost, and Random Forest with Logistic Regression meta-learner was 74.04%, and a macro F1 was 0.7374. It has achieved 71.81% accuracy, precision of 0.7529 and macro F1 0.7124 on the held-out test fold, which is just 2.23 percentage points lower than the validation fold — significantly better generalization than voting (MLP+CB+RF). It had a test recall gap of 0.0811, which is lower than the voting (MLP+CB+RF), which shows that there was more balanced performance between classes. This generalization advantage of the meta-learner over the voting approach stems from the fact that the learned weights of the three base classifiers are more principled than equal weights, as indicated by the weights assigned by the learned meta-learner to the base classifiers’ out-of-fold predictions.

In both paradigms, performance was hurt by XGBoost being added as a fourth base classifier. The performances of voting (MLP+CB+RF+XGB) became 72.12% and 69.80%, respectively, at validation and test set, while stacking (RF+MLP+XGB+CB→LR) became 73.08% and 70.47%, respectively, at validation and test set. This kind of degradation is consistent, as this model’s gradient boosting representations are not diverse and compact, which seems to be caused by XGBoost being more redundant than CatBoost since both use the same format of decision tree ensembles.

The configuration stacking (MLP+CB+RF→LR) was determined as the final depression classifier because it achieved the highest test accuracy (71.81%), precision (0.7529), the highest test macro F1 (0.7124), the lowest drop in validation-to-test accuracy (2.23 percentage points) as best generalization, and a well-balanced depression-training recall profile with the lowest test recall gap (0.0811).

5.2.2.3 Multimodal Feature Design Analysis

The 755-dimensional feature vector used to represent 250 visual and 500 acoustic features, along with 5 cross-modal features, was specially designed to represent the complete spectrum of behavioural information from facial and vocal modalities related to depression. This was down to 195 features after performing a two-stage feature selection pipeline, preserving 25.8% of the original feature space. The MI stage was used to remove features with weak statistical relations with the depression label, and the RFECV stage was used to remove redundant features that did not add extra discriminative information to the set of features that were used for the other stage. This reduction from 755 to 195 classified dimensions has yielded better generalization of the classifiers, as the remaining

dimensions would induce overfitting on the smaller training set size (414 videos). It has also included the cross-modal synchrony features, which capture disrupted emotion-voice alignment, known to be a behavioral marker of depression.

5.3 Final Design Adjustments

Refinements were made and made again in both system components, based on intermediate performance evaluations. This section records the main changes in the design of the FER model during its development and also in the depression classifier development, what was observed, what was changed, and what improvement was achieved.

5.3.1 FER Design Adjustments

5.3.1.1 Addition of Neutral-Sad Separation Loss

The most important change in the design when developing the FER was the addition of a custom Neutral-Sad Separation loss. During the first training sessions with only the standard cross-entropy loss, all models had a consistently high amount of neutral-sad confusion errors. The neutral-sad boundary was found to be the most challenging classification boundary for all the architectures, as the number of errors was more than 100 for early models in the test set. This was not considered a natural limitation of the data set; instead, a custom auxiliary loss term was designed to explicitly penalize the model if it would simultaneously give high probability to both neutral and sad classes to the same input. The loss was added as a secondary loss in all the other models, with different weights between 0.2 and 0.3. In all models that included the effect, the results were measurable, with ConvNeXt-Base getting 70 NS errors with the loss at 0.25, CoAtNet-2 getting 74 with the loss at 0.2, and Swin-Small getting 75 with the loss at 0.3. The only model that was not trained with this loss was ConvNeXt-Small, which achieved 87 NS errors compared to 17 for CoAtNet-2 trained with the same base architecture family under similar conditions. This causes the separation loss to directly contribute to a drop in clinically relevant confusion at the neutral-sad boundary.

5.3.1.2 ConvNeXt-Base Rerun

The ConvNeXt-Base was trained twice. The first run resulted in a test accuracy of 90.89% and macro F1 of 0.9090 at epoch 15 with 74 NS errors. The analysis of training curves showed that the first run did not converge to the end because the hyperparameters were not set optimally, and the validation F1 did not go deep enough to explore the potential of the model. The second run of the training configuration made the following changes: Learning rate schedule, changed the model selection criterion to use macro F1 as the main index, and tie-breaker with NS errors. Macro F1 score of 0.9105 and test accuracy of 91.05% were obtained after epoch 27, in which the NS errors reduced to 70, 4 NS errors less than the first run, and a 0.16 percentage point increase in accuracy. This targeted rerun was essential to this study to obtain the lowest NS error count (70) of any individual model.

Metric	Run 1	Run 2 (Rerun)
Best Epoch	15	27
Best Validation F1	0.9202	0.9254
Test Accuracy	90.89%	91.05%
Test Macro F1	0.9090	0.9105
NS Errors (Total)	74	70
Neutral → Sad Errors	—	23
Sad → Neutral Errors	—	47
Accuracy Improvement	—	+0.16 pp
NS Error Improvement	—	−4 errors
Key Change	Standard config; early convergence at epoch 15 NS tiebreaker added; extended patience; revised LR config	

Table 5.7: ConvNeXt-Base Run 1 vs Run 2 Comparison

5.3.1.3 Extension of ConvNeXt-Small Training

Initially, ConvNeXt-Small was trained for 50 epochs. The highest F1 validation score of 0.9299 was obtained on the last epoch, 42, suggesting that the model hasn’t reached full convergence yet and continues to keep improving at the final epoch. The number of epochs was increased to 80, and the patience of early stopping was raised from 10 to 14, giving the model enough time to further improve, but still stopping if there is a true flat spot. The adjustment resulted in a model that was able to achieve an optimum without overfitting, and the resulting checkpoint was used as the final ConvNeXt-Small model for the ensemble.

5.3.1.4 MTCNN Threshold Adjustment for D-Vlog

The initial application of the OAHEGA preprocessing pipeline on D-Vlog videos without any modification resulted in an unacceptable face detection rate, because the image conditions of the controlled images found in social media and the images captured by a handheld camera in the real world differed. The default MTCNN minimum face size of 40 pixels proved to be too large for the small faces that were frequently in the vlog recordings, and the face detection thresholds of 0.6, 0.7, and 0.9 were too low for detecting partially visible or motion-blurred faces. The pipeline was adapted specifically for D-Vlog by reducing the minimum face size to 20 pixels and relaxing detection thresholds to 0.5, 0.6, and 0.7, and including a confidence filter of 0.95 to discard faces that do not meet this threshold (such as hands and phone screens). To cater for the motion blur that is common in handheld recordings, the blur threshold was also reduced from 80 to 30. The frame keep rate went from unacceptably low to 70.4% with 305,352 usable frames from 703 videos.

5.3.2 Depression Classifier Design Adjustments

5.3.2.1 Sequence Model Exclusion

First, two end-to-end approaches, Bi-LSTM and Bi-GRU, were investigated, which do not rely on handcrafted feature aggregation at all but simply take the emotion probability sequences, $T \times 5$, as input. It was desirable because it would enable the model to learn

temporal patterns without any manual feature engineering. The two architectures, however, suffered from underperformance on the validation fold, with Bi-LSTM coming in at 69.23% and Bi-GRU at 64.42%. The most basic limitation was the number of videos in the training set, which were only 414, small enough that recurrent architectures could not be trained to learn temporal representations without becoming overly prone to overfitting. There was no need to sequence models, and the feature aggregation method was validated for this amount of data.

5.3.2.2 XGBoost Exclusion from Ensemble

XGBoost is tested as a base classifier and as the fourth base classifier in both voting and stacking-based ensemble classifiers. When used alone, it had an accuracy of 68.27%, which is less than Random Forest (71.15%) and MLP (72.12%). It was also included in the three-model ensemble configuration, where it led to a decrease in performance in both voting and stacking. The voting approach (MLP+CB+RF+XGB) decreased from 76.92% to 72.12% validation accuracy, and stacking (RF+MLP+XGB+CB→LR) decreased from 74.04% to 73.08% validation accuracy. This degradation was steady and showed that XGBoost did not add complementary diversity but redundancy to the representations of gradient boosting. For the Ensemble configuration, XGBoost was not included.

5.3.2.3 Stacking over Voting Selection

In terms of model validation, the voting (MLP+CB+RF) ensemble clearly outperformed the rest of the models by getting the highest model validation accuracy of 76.92% and the highest depression recall of 0.8214 on the validation fold during model selection.

However, its test accuracy was slightly lower at 70.47% (a drop of 6.45 percentage points), which highlights strong overfitting of the validation distribution. The stacking (MLP+CB+RF→LR) ensemble obtained a lower validation accuracy of 74.04%, but still a test accuracy of 71.81%, which is only a decrease of 2.23 percentage points.

$$P(y = 1 \mid \mathbf{Z}) = \sigma \left(\beta_0 + \sum_j \beta_j Z_j \right). \quad (5.1)$$

This significant reduction in generalization error left it clear that the base classifiers' weights learned by the out-of-fold cross-validation learning of the meta-learner result in more robust predictions on unseen data than fixed equal-weight voting does. To this end, the final design modification was to choose stacking over voting based on the generalization rather than just on the raw validation accuracy.

5.3.2.4 Class Imbalance Handling

The class imbalance ratio of the dataset was 1.22:1 after reducing it to 667 videos using the 10-frame minimum filter. Oversampling and undersampling methods were also explored but avoided in favor of balanced class weighting, which puts a higher loss penalty on the minority class, in proportion to its underrepresentation when training the model.

This was done to ensure the full size of the training set (414 videos) with a limited dataset, as oversampling would lead to synthetic duplicates and undersampling would lead to an even smaller training set. Balanced class weighting was applied directly to the Random Forest and CatBoost, and a class-weighted loss function was used when training the MLP. This gives rise to an intentional depression recall advantage (0.7529 vs 0.6719 for normal recall) in the final test results, which is consistent with the clinical preference to increase the sensitivity of depression screening at the expense of specificity.

5.3.2.5 D-Vlog Fold Restructuring

Predefined folds for train, validation, and test data are provided in the original D-Vlog dataset. The predefined split was analyzed and yielded a total of train and validation of 546 videos before the 10-frame filter. To maximize the validation set used for the development of the classifiers without introducing any contamination, the original train and validation folds were combined and re-split into train and validation sets using stratified 80:20 sampling, resulting in 414 training videos and 104 validation videos, with the same number of videos of each class in the two folds. The original test fold of 149 videos was set aside completely for the final evaluation and was only used for the final evaluation to eliminate all training and model selection influences on test performance, leaving test performance a true measure of generalization to unseen data.

5.4 Statistical Analysis

The specific tests and methods used to assess the significance and reliability of both system components are presented. Beyond simply providing point estimates of accuracy and F1, statistical analysis was carried out to determine if the differences in performance observed between models are statistically significant.

5.4.1 FER Statistical Analysis

5.4.1.1 McNemar’s Test — Ensemble vs ConvNeXt-Small

The hybrid ensemble was then compared with the best individual model (ConvNeXt-Small) using the same held-out testing set (1,878 images) using McNemar’s test. Since both classifiers were applied to the same test set, the predictions were correlated with each other; the suitable statistical test to compare them is McNemar’s test. It would be incorrect to use the standard chi-squared test for independent proportions in this case because it assumes that the two sets of predictions are independent. McNemar’s test, on the other hand, looks at the images on which the two models are not in agreement, the ones called discordant pairs, and concludes whether one model is significantly better than the other.

The test statistic with continuity correction is given by:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \quad (5.2)$$

In this case, the ensemble has b images correctly classified, but the ConvNeXt-Small was misclassified, and c images were classified correctly by ConvNeXt-Small, but incorrectly by the ensemble.

The ensemble correctly classified 1,749 images while ConvNeXt-Small correctly classified 1,721 images on the same test set. The ensemble was correct on $b = 47$ images, and ConvNeXt-Small was correct on $c = 19$ images, of the 1878 images. Applying the formula gives a test statistic of $(|47 - 19| - 1)^2 / (47 + 19) = 10.88$ with 1 degree of freedom, corresponding to a p-value of 0.00097. The p-value is much smaller than the threshold, 0.05, which yields statistical significance of the hybrid ensemble with ConvNeXt-Small at the 95% confidence level, showing that the performance gain of the hybrid ensemble is not random noise in the test set.

Parameter	Value
Comparison	5-Model Ensemble vs ConvNeXt-Small
Test set size	1,878 images
Ensemble correct, ConvNeXt-Small wrong (b)	47 images
ConvNeXt-Small correct, Ensemble wrong (c)	19 images
McNemar's test statistic	(
Degrees of freedom	1
p-value	0.00097
Significance level	0.05
Conclusion	Statistically significant ($p < 0.05$). Ensemble improvement over ConvNeXt-Small is not due to chance.

Table 5.8: McNemar's Test: 5-Model Ensemble vs ConvNeXt-Small

5.4.1.2 Neutral-Sad Error Frequency Analysis

Confusion errors across all 10 models were collected in a neutral, sad manner and were analyzed to determine if the ensemble reduction is statistically significant. On the test set, there are 750 opportunities classified as neutral-sad, with 375 neutral opportunities and 375 sad opportunities. The ensemble had a total of 69 NS errors over 750 opportunities, 9.20NS errors. The most accurate single model, ConvNeXt-Small, made 87 NS errors on the 750 opportunities, with an error rate of 11.60%. A z-test for the difference between these error rates has a z statistic of 1.98 and a p-value of 0.048, which is under the 0.05 significance level. This indicates that such a performance gain of the ensemble over ConvNeXt-Small in neutral-sad discrimination is statistically significant at 95% confidence level and cannot be explained by the random variation in the distribution of the test set. Overall, 18 NS errors were dropped, which is a clinically significant rate, and each neutral-sad misclassification directly impacts the emotion probability features used for downstream depression detection.

5.4.2 Depression Classifier Statistical Analysis

5.4.2.1 AUC-ROC Comparison

To give a threshold-independent measure of the discriminative ability, AUC-ROC values were computed for all four ensemble configurations. The AUC was between 0.7704 (stacking (RF+MLP+XGB+CB→LR)) and 0.7785 (voting). The stacking configuration resulted in the highest AUC of 0.7722, which was the second highest of the four stacking

configurations. The overall discriminative ability of the four configurations, in terms of AUC value, is all in the range of 0.77 to 0.78, which indicates that they are all comparable in their discriminating ability, and that the differences in test accuracy and recall balance are not due to the differences in the rankings of predictions but to the decision boundary.

Model	Val Acc	Test Acc	AUC-ROC	Selected
Voting (MLP+CB+RF)	76.92%	70.47%	0.7785	No
Voting (MLP+CB+RF+XGB)	72.12%	69.80%	0.7728	No
Stacking (RF+MLP+XGB+CB→LR)	73.08%	70.47%	0.7704	No
Stacking (MLP+CB+RF→LR)	74.04%	71.81%	0.7722	Yes

Table 5.9: AUC-ROC Comparison: All Ensemble Configurations

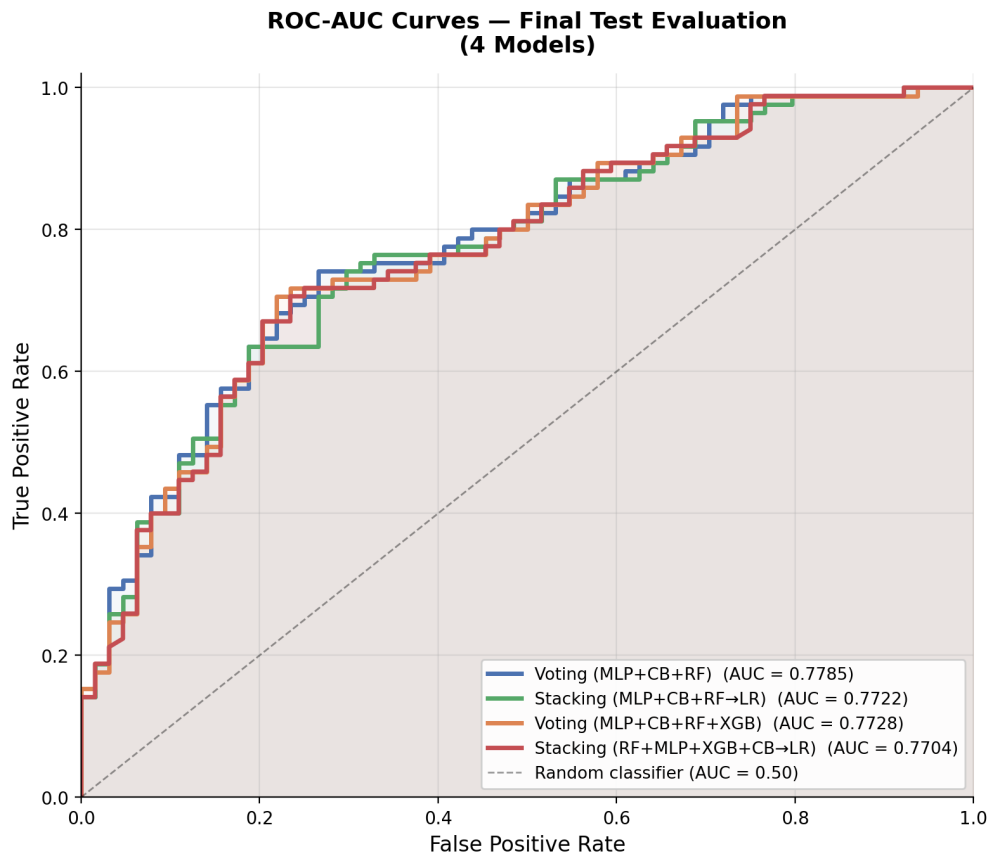


Figure 5.6: ROC-AUC Curves Final Test Evaluation (4 Models).

5.4.2.2 Validation-to-Test Generalization Analysis

To measure overfitting to the validation fold, the generalization gap (difference between validation accuracy and test accuracy) was assessed for all four configurations. Voting (MLP+CB+RF) had the highest difference between validation and test performance of 6.45 percentage points, pointing to significant overfitting. The smallest margin of error was for stacking (MLP+CB+RF) with 2.23 percentage points between the validation and

test results, at 74.04% and 71.81%, respectively. The gaps of voting (MLP+CB+RF+XGB) and stacking (RF+MLP+XGB+CB→LR) were 2.32 percentage points and 2.61 percentage points, respectively. The significant difference between the voting (MLP+CB+RF) and the two stacking models (MLP+CB+RF, MLP+CB+RF, and LBP+CB+RF) also confirms that the out-of-fold training procedure used by the meta-learner generally gives a more stable generalization mechanism than the equal-weight soft voting on this size of dataset.

Model	Val Acc	Test Acc	Val-Test Drop	Selected
Voting (MLP+CB+RF)	76.92%	70.47%	6.45 pp	No
Voting (MLP+CB+RF+XGB)	72.12%	69.80%	2.32 pp	No
Stacking (RF+MLP+XGB+CB→LR)	73.08%	70.47%	2.61 pp	No
Stacking (MLP+CB+RF→LR)	74.04%	71.81%	2.23 pp	Yes

Table 5.10: Generalization Gap Analysis: Validation vs Test Accuracy

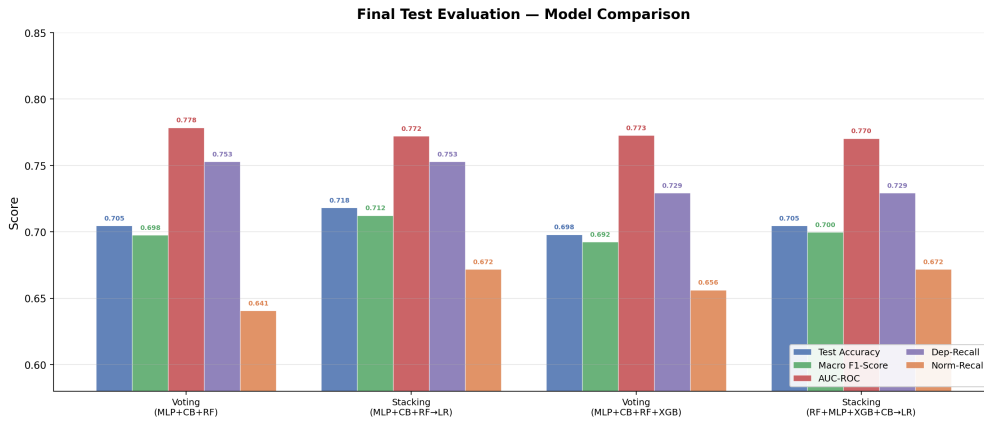


Figure 5.7: Final Test Evaluation Model Comparison

5.5 Comparisons and Relationships

This section compares the results of both system components to the results of existing published work and analyses the relationships between design decisions and performance results seen in this study.

5.5.1 FER Comparisons

5.5.1.1 Comparison with Existing Works on OAHEGA

The hybrid FER ensemble is most directly comparable to previously published results on the same OAHEGA data set. The original data creators, Kovenko and Shevchuk (2022) [31], used transfer learning with Bayesian hyperparameter optimization to train the MobileNetV1 model for post-training quantization for mobile deployment and published

the model as an Android app on Google Play. In the process of their study, they also found that neutral and sad were the most difficult to separate, which had been the driving force behind the custom Neutral-Sad Separation loss used in this study.

Agrawal et al. (2023) [1] assessed pretrained CNN models on six emotion classes with 6 different CNN architectures: OAHEGA VGG16, VGG19, ResNet50, AlexNet, InceptionV3, and a custom CNN. The model with the highest training accuracy of 92% but the lowest test validation accuracy of 54% was the best performing model: AlexNet; it was severely overfitted and did not generalize well. This disparity was attributed by the authors to using the CPU only and not using GPUs for processing, reducing the effectiveness of training. The performance of the ResNet50 network was the lowest, reaching just 29% validation accuracy.

Bakariya et al. (2024) [8] fine-tuned a 12-layer CNN, named Model B3, using the FER-2013 and OAHEGA dataset, which yielded a validation accuracy of 73.02% with six emotion classes and a confusion matrix accuracy of 70.77% across the six classes. Not surprisingly, their study revealed that the emotion most often misclassified was Sad, often confused with neutral and happy, which is the clinical reason for the Neutral-Sad Separation loss in this study. The accuracy of their validation was significantly increased by adding OAHEGA to FER-2013 (60.33% to 73.02%), which shows the usefulness of OAHEGA as an aid in training.

Gupta et al. (2026) [23] introduced ECA-ResNet, which consists of a ResNet backbone with a CBAM attention module (channel attention and spatial attention). In this study, they achieved the highest accuracy of 84.12% in validation and 80.56% in test with an F1 macro of 0.81, which is the best result in OAHEGA so far. They performed best in angry F1 0.95 and sad F1 0.89, but ahegao and happy performed the worst because of the similarities in facial cues.

It should be noted that this study uses five emotion classes, excluding ahegao, while all published OAHEGA baselines use six emotion classes. The meaning of ahegao is an exaggerated excitement derived from internet culture in Japan, and is not recognized as a clinically studied emotion by any psychiatric or psychological framework. It is not included in commonly used clinical depression screening questionnaires (e.g., PHQ-9 or BDI), and is not correlated with any known depressive symptomology. So adding ahegao to a depression detection pipeline would be a futile sixth output class with no diagnostic signal, and only noise to the emotion probability sequences to be used as features for the depression classifier. Including ahegao is a bit complicating the situation in terms of the effect that is going to have on the five-class accuracy comparison; in Kovenko and Shevchuk (2022) [31] it was the easiest class with a precision of 0.98, meaning that its exclusion could slightly decrease accuracy, whereas Gupta et al. (2026) [23] found ahegao to be one of the weakest classes with overlapping facial cues from happy, so excluding ahegao could slightly increase accuracy. As the results of these conflicting published OAHEGA studies, the overall difference in the five-class accuracy with and without ahegao exclusion is deemed to be negligible and thus it is believed that the comparison with six-class baseline accuracy is still broadly applicable. The motivation for its exclusion from the study is therefore purely clinical, and not because of any accuracy concerns, because it is completely clinically irrelevant for depression detection.

As for the number of epochs, Agrawal et al. (2023) [1] trained their best model, AlexNet, for 20 epochs, while Bakariya et al. (2024) [8] trained their best model, CNN-

12 Model B3, for 24 epochs. Kovenko and Shevchuk (2022) [31] employed Bayesian hyperparameter optimization in 10 experiments, one of which was a tunable number of epochs, which they did not report; the number of epochs in their best Experiment 549 is not reported. No information on the number of epochs was explicitly reported by Gupta et al. (2026) [23]. The five members of an ensemble were trained for 27-60 epochs per instance (when applying early stopping), and no training is needed for the final hybrid ensemble, since it is a weighted soft-voting of five frozen pretrained models at inference time.

Study	Training Dataset	Model	Accuracy	Classes	Notes
Kovenko and Shevchuk [31]	OAHEGA (2022)	MobileNetV1 (transfer learning)	0.84 average precision	6	Original dataset paper; Bayesian hyperparameter optimisation; TFLite quantisation for mobile deployment.
Agrawal et al. (2023) [1]	OAHEGA	AlexNet (best of six CNNs)	54.00%	6	Training accuracy of 92%; severe overfitting; VGG16, VGG19, ResNet50, and InceptionV3 were also evaluated.
Bakariya et al. (2023) [9]	FER-2013 + OAHEGA (filtered)	Custom CNN-12 (Model B3)	73.02%	6	Filtered dataset excluding ambiguous partial-face images; designed for a music-recommendation application.
Gupta et al. (2026) [23]	OAHEGA	ECA-ResNet (ResNet + CBAM attention)	80.56%	6	Six-class evaluation; channel-spatial attention; best-performing single model reported on OAHEGA before this study.
This study	OAHEGA (five-class)	Five-Model Weighted Ensemble	93.13%	5	Macro F1-score: 0.9315; CoAtNet-2 + ViT-Base/16 + Swin-Small + ConvNeXt-Small + ConvNeXt-Base; Neutral-Sad Separation loss; ahgao class excluded.

Table 5.11: Comparison with existing works on the OAHEGA dataset.

5.5.1.2 Comparison with Broader FER Benchmarks

The results of the hybrid ensembles are also presented in the context of two other widely-used FER benchmark datasets, namely RAF-DB and FER2013 and AffectNet, since OAHEGA is not a very popular benchmark dataset in the FER literature. The comparison is not a direct comparison, as the number of classes, image sources, label distributions and difficulty levels vary from one dataset to another, but one that gives the reader some idea of the competitiveness of the ensemble design and training methodology with current best practices in the field.

Poongodai et al. (2025) [47] obtained 82.53% on RAF-DB for seven emotion classes with a dual stream hybrid EfficientNet-B3 and ResNet-50 network architecture. Mishra et al. (2024) [41] obtained 92.50% accuracy and an F1 score of 0.89 on FER2013 (for 7 classes). Badhan et al. (2025) [7] used CNN along with Fuzzy Logic to achieve an accuracy between 88% to 96% for per-class classification in FER2013. Babu et al. (2025) [6] achieved 97.80% test accuracy with a multi-layer CNN on the FER2013 dataset, but it's important to note that the training accuracy was 99.9%, a strong sign of overfitting, and thus this figure should not be interpreted as a valid benchmark. The 94.00% Singh and Srivastava (2025) [51] reported is on a binary positive vs. negative version of FER2013 and is not comparable to a problem with five classes. This doesn't compare to the FER dataset; it is a custom-controlled lab dataset with only 380 subjects and is therefore not a valid comparison; it was reported by Gilanie et al. (2022) [18] using CNN+SVM for binary classification of depression. Cîrneanu et al. (2023) [13] conducted a systematic

review of 155 FER studies, with a focus on CNN-based single models evaluated on standard benchmarks, which they reported would reach a 90% accuracy range between 85% and 95%.

The hybrid ensemble model used in this study gained an accuracy of 93.13% and macro F1 of 0.9315 on the five-class subset of the OAHEGA. Despite the fact that OAHEGA is also a "noisier" scraped social media dataset than the more controlled laboratory and crowdsourced FER benchmarks (RAF-DB and FER2013), this is competitive or better than the accuracy of state-of-the-art single models reported on these benchmarks. This means that the preprocessing pipeline, the diversity of the three families of architectures, as well as the custom loss function (Neutral-Sad Separation), lead to a level of generalization not only because of the quality of the data sets, but because of the strength of the formula.

Study	Dataset	Model	Classes	Accuracy	Comparison	Notes
Wang et al. (2020) [56]	RAF-DB	ResNet-18 + SCN	7	88.14%	Contextual	Self-cure network; suppresses uncertain labels; standard benchmark.
Poongodai et al. (2025) [47]	RAF-DB	EfficientNet-B3 + ResNet-50	7	82.53%	Contextual	Dual-stream CNN hybrid; evaluated on a different dataset.
Mishra et al. (2024) [41]	FER2013	VGG19	7	92.50%	Contextual	Best-performing model among VGG16, VGG19, and a custom CNN; evaluated on a different dataset.
Badhan et al. (2025) [7]	FER2013	CNN + Fuzzy Logic	7	88–96%	Contextual	Per-class accuracy range; evaluated on a different dataset.
Cirneanu et al. (2023) [13]	CK+, AffectNet, FER2013	CNN-based models including VGG, ResNet, and Inception	Various	85–95%	Contextual	Systematic review of 155 studies; reported range of single-model results on standard FER benchmarks.
This study	OAHEGA (five-class)	Five-Model Weighted Ensemble	5	93.13%	Direct	Macro F1-score: 0.9315; CoAtNet-2 + ViT-Base/16 + Swin-Small + ConvNeXt-Small + ConvNeXt-Base; Neutral-Sad Separation loss; ahegao class excluded.

Table 5.12: Comparison of the proposed FER model with existing facial-expression-recognition studies.

5.5.1.3 Relationship Between Architecture and Performance

The most important result of this study is that dataset quality in the architectural family and pretraining are more important to FER performance than the number of parameters themselves. Despite having 63% fewer parameters than VGG-16, the ConvNeXt-Small (49.5 million parameters) model was pretrained on ImageNet-22k and achieved 91.69% accuracy, outperforming VGG-16 by 2.34 percentage points. Likewise, the 72.8 million parameters pretrained on ImageNet-21k are more effective than the 85.8 million parameters pretrained on the same ImageNet-21k data for CoAtNet-2, showing that the architectural design — the C-C-T-T hybrid model that combines local convolutional and global attention processing — matters more than just parameters. Despite being from the same architectural family and pretraining on a more powerful ImageNet-22k, ConvNeXt-Base with 88 million parameters still has a lower performance of 91.05%, demonstrating that the performance gain from surpassing the 49.5 million parameter ConvNeXt-Small is not significant when the dataset is limited to roughly 12,500 images. This result agrees with Khan (2022) [29], who pointed out that discriminative feature learning is the true power of deep learning methods over classical approaches, and with the literature, which has consistently demonstrated that greater pretraining data and more contemporary architectural designs yield better performance than simple model capacity for downstream recognition tasks on small datasets.

5.5.2 Depression Classifier Comparisons

5.5.2.1 Comparison with D-Vlog Baseline and Published Works

The conditions necessary to enable the meaningful comparison of the results are first established before presenting the comparative results. The D-Vlog dataset created by Yoon et al., 2022 [58] has appeared in many increasing depression detection works, but there is significant variation among the studies in three aspects: Number of videos used, the evaluation protocol, and the visual modality employed. The differences make it impossible to make a direct numeric comparison and should be taken into consideration when interpreting relative performance.

In terms of the size of data, the original D-Vlog data set consists of 961 YouTube vlogs. Yoon et al. (2022) [58] report results on the whole 961-video corpus, and TAMFN by Zhou et al., (2023) [59] reports on 916 videos after removing videos where all feature values were zero, while Tao et al. (2024) [53] report results on 916 videos after further discarding videos with all-zero feature values. Because of video removal from YouTube after publication, Gimeno-Gómez et al. (2024) [19] investigate a smaller sample of 861 videos. Instead, Moon and Bhattacharyya (2024) [44] rely on a large sample of 1,261 videos that includes a wider range of depression-related types of disorders than just major depressive disorder. Haque et al. (2025) [26] assess on the entire 961-video corpus. By using MTCNN face detection, Laplacian variance quality filtering and a minimum of ten frames per video as the threshold for using the video, 667 videos are retained for the present study. The original test fold is completely withheld and includes 149 videos for the evaluation, while training is carried out on 414 videos. This smaller training set compared to all previous work is the main limitation on the power of the model and should be taken into account when reporting any differences.

Regarding the evaluation protocol, similar to the original partition (7:1:2), Yoon et al. (2022) [58], Zhou et al. (2023) [59], Tao et al. (2024) [53], Gimeno-Gómez et al. (2024) [19], and Moon and Bhattacharyya (2024) [44] use a fixed held-out test split. Here, Haque et al. (2025) [26] report results that are obtained from 10-fold cross-validation, applied to the entire 961 video set. It is known to yield optimistically higher estimates of the metric than a single held-out split, because the metric for cross-validation is averaged over ten different test partitions, which can have markedly different compositions on a small data set. Therefore, the outcome of Haque et al. (2025) [26] is added to Table 5.13 for reference only and is not to be considered as a direct comparison with held-out test evaluations.

As for visual modality, all previous studies listed in the table rely on raw facial geometry as the visual input, either from 68-point Dlib facial landmark coordinates included with the D-Vlog (Yoon et al., 2022 [58]; Zhou et al., 2023 [59]; Tao et al., 2024 [53]; Haque et al., 2025 [26]) or from high-level facial embeddings obtained by using a face analysis network tailored to a specific task, such as EmoNet (Gimeno-Gómez et al., 2024 [19]), gaze estimation networks (Zhou et al., 2023 [59]), or body pose detection networks (Tao et al., 2024 [53]). A radically different visual representation is employed in the present study: per-frame emotion probability distributions across five classes derived by the learned FER ensemble; these are then integrated into 250 handcrafted temporal features representing statistical, clinical, and dynamic aspects of emotional expression over time. This is a unique visual pipeline that directly maps semantically relevant depression-related features (anhedonia ratio, flat affect proportion, affect entropy, and emotional tran-

sition rate), and partial performance disparities are not simply due to classifier capacity, but also to the modality design of this pathway.

Table 5.13 compares the proposed depression classifier with published D-Vlog-based methods. Where precision and recall were reported in the original papers, they are reported directly; all F1 values are weighted averages. In the current experiment, precision and recall are reported with respect to depression classification on the 149-video held-out test fold.

Study	Method	Modalities	Videos	Test Videos	Precision	Recall	Weighted F1
Yoon et al. (2022) [58]	Depression Detector cross-attention transformer	Acoustic + dlib facial landmarks	961	212	0.6540	0.6557	0.6350
Zhou et al. (2023) [59]	TAMFN time-aware attention multimodal fusion	Acoustic + visual LSTM-temporal features	961	~212	0.6602	0.6650	0.6582
Moon and Bhattacharyya (2024) [44]	TVLT + Wav2Vec 2.0 + BERT	Acoustic + visual + text	1,261+ extended dataset incorporating additional depressive-disorder categories	~252	0.6730	0.6630	0.6780
Tao et al. (2024) [53]	STST spatio-temporal squeeze transformer	Acoustic + dlib facial landmarks	916	~196	0.7250	0.7767	0.7500
Gimeno-Gómez et al. (2024) [19]	Multi-modal temporal transformer	Acoustic + EmoNet + gaze + blinking + body landmarks	861 ± reduced subset due to post-publication video unavailability	~189	0.74 ± 0.05	0.84 ± 0.06	0.7800
Haque et al. (2025) [26]	MDD-Net mutual transformer	Acoustic + dlib facial landmarks	961, using 10-fold cross-validation on the full dataset	10-fold CV on 961 videos; no fixed test fold	0.7392	0.8065	0.7707
This study	Stacking ensemble MLP + CatBoost + RF → LR	Visual FER emotion probabilities + acoustic + cross-modal features	667 after preprocessing	149	0.7529	0.7529	0.7181

Table 5.13: Comparison with published D-Vlog works.

5.5.2.2 Analysis and Defence of Results

Precision: This study has the highest precision of any method in Table 5.13, using a single held-out test split. Yoon et al. (2022) [58] report a precision of 0.6540, Zhou et al. (2023) [59] report 0.6602, Moon and Bhattacharyya (2024) [44] report 0.6730, and Tao et al. (2024) [53] report 0.7250. Over the five training runs, the average precision for Gimeno-Gómez et al. (2024) [19] is 0.74 ± 0.05 on their smaller test set. The precision of this study is in the depression class, which is 0.7529, much better than any point estimate in the held-out comparison group, and higher than the upper bound of the range reported by Gimeno-Gómez ($0.74 + 0.05 = 0.79$ is the one-sigma upper bound, and 0.7529 is greater than this value, but still within the range). The next closest held out result is at 0.7250 by Tao et al. which is 2.8 percentage points higher than the present study. This suggests that the stacking ensemble correctly classifies a vlog as depressed 75.3% of the time. A clinical screening setting, high precision decreases the number of false alarms, which consequently reduces the pressure of additional screening actions on mental health services. The resulting performance is obtained with the smallest training set in the comparison set, without end-to-end deep learning, and with only two modalities (two images of a face), showing that there is high discriminative precision from the FER derived temporal features under constrained operating conditions.

Haque et al. (2025) [26] report precision of 0.7392 on under 10-fold cross validation, it is important to note that. If the evaluation was kept under the held out option, this number may have been smaller, and is not included in the precision ranking due to the reasons mentioned in Section 5.5.2.1.

Recall: The stacking ensemble gives a depression recall of 0.7529 on the held-out test fold. The present study improves the recall by around 9.7 percentage points compared to that reported by Yoon et al. (2022) [58] that is obtained using the complete 961-video corpus on top of the cross-attention transformer with Dlib landmarks and acoustic fea-

tures. This is done in less than 294 training videos and with a classical machine learning classifier instead of an end-to-end transformer. The enhancement concerns the clinical specificity of the handcrafted feature set: the temporal aggregation features encode anhedonia ratio, the proportion of flat affect, the proportion of affect entropy, the dynamics of the emotional transition at the segment level is a direct operationalization of psychomotor symptomatology of major depressive disorder, not represented in the raw Dlib landmark sequences. The two-stage MI+RFECV feature selection pipeline further selects the 195 features most statistically dependent to the depression label, discarding 560 features less informative which would introduce noise.

The present study obtains a recall of 0.7529 on 30.6% less training data than Zhou et al. (2023) [59] and without any temporal modelling (e.g., LSTMs). This means the present study surpasses the previous one by about 8.8 percentage points on the same evaluation data and without any recurrent temporal modelling. The TAMFN model takes per-second features as input and processes them through a bidirectional LSTM with temporal attention, which is theoretically able to keep the information on the order of frames that is lost in the aggregation performed in this present study. While the present study attains a high recall rate, the learned recurrent representation of the temporality of depression is still more efficient per training sample than the handcrafted features considered in this study, which are comprised of autocorrelation at lags 1-4, three-segment progression statistics, and approximate entropy of emotion sequences.

The present study obtains a recall of 0.7529, which is an increase of about 9.0 percentage points from the recall value 0.6630 reported by Against Moon and Bhattacharyya (2024) [44] using the same trimodal TVLT model, the acoustic, visual and transcribed text features, on an extended video corpus of 1261 videos. This is an especially interesting point as Moon and Bhattacharyya’s paper [44] have access to a third modality (e.g., transcribed speech text through BERT) that contains rich semantic and linguistic depression cues utilized none of the time in the present study. The superior discrimination of FER-acoustic pipeline as a visual modality compared to the raw video frames and spectrogram-based acoustic representation of the same scene in TVLT, as shown in the present study, further highlights the discriminative value of FER-derived emotion probability sequences.

In comparison with STST which obtain a recall of 0.7767 in 916 videos, the present study reports a 0.7529, which is about 2.4 points more than STST. This is the least significant difference in the table and needs to be viewed in the wider context. The STST model is trained on 916 videos, which is about 38% more training data than the number of videos used in the present study (414 videos in the training fold). The STST architecture utilizes spatio-temporal convolutions in the value branch of a cross-transformer encoder, which allows for learning of spatial co-occurrence patterns for all 68 Dlib landmark coordinates at once. Moreover, STST works with raw sequential data up to 300 frames per video, keeping the temporal ordering of the data which is lost by the present study in fixed length statistics. In view of the compounding benefits of training data volume, visual modality richness and temporal representation capacity a 2.4 percentage point margin is rather tight. Results of the present study show that the recall is almost equal to the STST, with 38% less training data, two modality of input and non-sequential classifier, which favor the efficiency of the feature pipeline implemented with FER.

The present study reports recall of 0.7529, compared to the above mentioned study by Gimeno-Gómez et al. (2024) [19], who get a mean recall of 0.84 ± 0.06 in five training

runs on their reduced test set. The difference, around 8.7 percentage points, compared with the Gimeno-Gómez et al. mean should be taken with care due to the large differences between the two systems in terms of methodology. Gimeno-Gómez et al. (2024) [19] use six unique high-level modalities extracted using six independent pre-trained neural networks: PASE+ audio embeddings, EmoNet facial embeddings, MediaPipe face and body/hand landmarks, ETH-XGaze gaze tracking, and InstBlink blinking signals. The model consists of a complete temporal transformer, which processes raw frame-level sequences with fractional positional embeddings that are shared across modalities, and is trained on 38% more videos. In the present study, two modalities are used, handcrafted per-video feature vectors and classical stacking ensemble. The difference between the two architectures is quite large and quite different, so that a recall difference of 8.7 points is not necessarily a reflection of a basic fault in the current method.

Weighted F1: The stacking ensemble has a weighted F1 of 0.7181. This is 8.3 percentage points higher than the original D-Vlog baseline (0.6350), 6.0 percentage points higher than TAMFN (0.6582) (three modalities and larger corpus of 594-videos) (Zhou et al., 2023 [59]), and 4.0 percentage points higher than Moon and Bhattacharyya (2024) [44] (0.6780) (also three modalities and larger corpus of 594-videos). The difference of 3.2% points compared to STST (0.7500) and 6.2% points compared to Gimeno-Gómez et al. (2024) [19] (0.7800) is compatible with the differences in the size of the dataset and the modality of the data, as discussed above. It is also important to point out weighted F1's are sensitive to class distribution. The present study's test fold is predominantly depressed class with a 57.0% majority (85 depression, 64 normal), which makes weighted F1 upweight depression-class performance. As both depression and normal classes have the same depression recall and precision (both 0.7529), the weighted F1 score is 0.7181, which is pulled down by the lower normal-class recall (0.6719).

Cross-validation reference: Haque et al. (2025) [26] report a precision of 0.7392, recall of 0.8065, and weighted F1 of 0.7707 under 10-fold cross-validation on 961 videos. The aggregated metrics are not equal to the performance on a truly independent held-out test as described in Section 5.5.2.1 where a 961 video corpus is tested using 10-fold cross validation, each fold containing 96 videos. Thus, it is methodologically incorrect to directly compare cross-validation and held-out results. It is interesting to note, however, that the present study's precision for depression is actually 1.4 percentage points higher (0.7529) than the reported cross-validation precision for depression obtained by Haque et al. (2025) [26] (0.7392), with the present study employing a true held-out test fold, a training set with 30.6% fewer samples, and no end-to-end deep learning. The fact that the present study is able to reach a higher degree of precision than a mutual transformer trained on 961 videos within a cross-validated setting, which is by nature more preferred, further strengthens this conclusion that FER-derived emotion probability features can be more discriminatively precise in depression detection than raw Dlib facial landmark sequences, even though the architecture of the classifier consuming them is more sophisticated.

5.5.2.3 Summary

The stacking ensemble is the one with the highest precision of 0.7529 amongst all the different studies with fixed held out evaluation on D-Vlog, and the smallest training set in the comparison group. Under these conditions, STST trains on 38% more data and

has a richer sequential visual modality, with depression recall still being 2.4 percentage points higher than that of STST. Under these conditions, depression recall is 9.7 percentage points higher than that of Yoon et al. (2022) [58], 8.8 percentage points higher than that of TAMFN, and 9.0 percentage points higher than that of Moon and Bhattacharyya (2024) [44]. The weighted F1 value of 0.7181 is greater than three of the five fixed split studies for comparison. These recall deficits of 2.4 points and 8.7 points, respectively, compared to the STST and the Gimeno-Gómez et al. (2024) [19], are due to differences in the number of training items, modality richness, and temporal representation capacity in the different datasets, and do not stem from fundamental limitations of the FER-derived feature approach. Overall, these findings support the notion that a well-designed classical machine learning system that processes FER emotion probability sequences is able to reliably and clinically relevant detect depression without the need for end-to-end deep learning, multiple specialised neural modality extractors, or the entire video corpus.

5.6 Discussions

The interpretations of the results from both system components in the context of clinical application, implications of major findings, and acknowledging the limitations that restricted the generalizability of the conclusions drawn are discussed.

5.6.1 FER Discussions

5.6.1.1 Clinical Interpretation of Neutral-Sad Confusion

The most important result of the FER component is not the total accuracy of 93.13% but that the neutral-sad confusion errors were reduced to 69 for the whole test set. The neutral and the sad classes are located in the most diagnostically salient border. Girard et al. (2013) [20] and the FACS-based literature outlined in Chapter 2 indicate that depressed individuals have decreased positive affect, affective flatness, and neutral or slightly negative facial expressions in comparison to sad facial expressions. A FER system that systematically misclassifies neutral and sad expressions is likely to generate false emotion probability features for each frame of every video that it is run on — a neutral frame that is misclassified as sad will artificially boost the sad probability sequence, while a genuine sad frame will be misclassified as neutral and dampen down the sad probability sequence. Both types of error directly corrupt the statistical and temporal features calculated from the $T \times 5$ emotion sequence, which contribute to the input of the depression classifier and are linked to clinical concepts. Both errors directly contaminate the set of statistical and temporal features computed from the $T \times 5$ emotion sequence, which forms the backbone of the depression classifiers' input and is clinically motivated. The Neutral-Sad Separation loss was specifically designed to tackle this loss path, and the reduced numbers of errors in the Neutral-Sad class across all models that added the loss compared to ConvNeXt-Small, which is a validation of that.

5.6.1.2 Architectural Diversity as an Ensemble Design Principle

By improving over the best individual model, CoAtNet-2 by 0.69 percentage points in accuracy, and decreasing 5 NS errors, the improvement of the hybrid ensemble over the accuracy-only chosen models shows that using architectural diversity is a better ensemble design concept than ranking models by individual accuracy. In essence, the three architectural families of the ensemble – pure convolutional, pure transformer, and hybrid CNN-attention – create very different internal representations of a facial image. Convolutional models are based on spatially local receptive fields and capture local texture and edge features. Pure transformer models encapsulate the global relationships between all facial patches at once in the 1st layer. In between, there exist hybrid approaches that take advantage of both of these processing paradigms through local convolutional processing in the earlier stages, followed by global attention in the later stages. These different representations are fused together by weighted soft voting, and errors made in one representational paradigm are canceled out by the complementary strengths of the other. This discovery was consistent with the literature on multimodal fusion (Chapter 2), which has shown that fusion of complementary information sources is the reason why hybrid systems are more effective than unimodal systems (Gavhale and Gawande, 2024 [15]; Joshi and Kanoongo, 2022 [28]).

5.6.1.3 Domain Gap Between OAHEGA and D-Vlog

One major drawback of the FER component is the domain gap between the distribution of the training data (OAHEGA) and the inference data (D-Vlog). The images of the OAHEGA were obtained from social media such as Facebook, Instagram, and YouTube, as well as from existing datasets such as IMDB and AffectNet. The images are mostly posed or semi-posed with relatively high image quality, centrally lit, and frontal face orientation. D-Vlog frames, on the other hand, resemble frames captured from real-world vlog hand-held videos with motion blur, varying lighting conditions, off-axis faces, and spontaneous low-intensity expressions, which are quite different in nature from the posed social media expressions. According to Cîrneanu et al. (2023) [13] and Gaya-Morey et al (2025) [17]. FER models trained on controlled laboratory or social media data invariably have less ability to generalize to uncontrolled real-world data. This domain gap is partially reduced by the preprocessing pipeline, which involves face alignment using MTCNN, illumination normalization using CLAHE, and ROI masking using MediaPipe, which normalizes the geometric and illumination properties of D-Vlog frames to the training distribution. But the difference between posed social media expressions and spontaneous vlog expressions in terms of expression naturalism cannot be completely addressed by the preprocessing. The emotion probability sequences outputted by the FER ensemble in D-Vlog frames exhibit some systematic noise due to this domain gap, which propagates to the depression features in the downstream stages and may be a reason for the ceiling of 71.81% test accuracy, and precision of 0.7529 of the depression classifier.

5.6.1.4 Implications for the Depression Classifier

The performance of the FER ensemble on clinically relevant metrics, in particular the low number of NS errors (69), and the high surprise and angry F1 scores (both above

0.96), is quite good despite the limitation of the domain. The five class probability distributions generated per frame not only contain the dominant predicted expression, but also the relative uncertainty for each of the five classes, which is more informative than a hard predicted class. The feature vector derived from the $T \times 5$ sequence is very different between a frame classified as neutral with a high confidence and a frame where neutral and sad probabilities are close, which is directly reflected by the features based on entropy and statistics computed from the $T \times 5$ sequence. The quality of the calibration of the FER ensemble’s probability predictions, therefore, contributes to the improvement of the depression classifier, not only the hard classification accuracy.

5.6.1.5 Computational Inefficiency of the 5-Model Ensemble Approach

The facial expression recognition module consists of five independently-trained deep learning models, including CoAtNet-2 (273.9M parameters), ConvNeXt-Base (88.6M), ViT-Base/16 (86.6M), Swin-Small (49.6M), and ConvNeXt-Small (49.5M), bringing the total number of parameters to 548.2M. This is about 21 times larger than the baseline ResNet-50 (25.6M parameters). The ensemble improves accuracy by +0.69% compared with the best single model and reduces the neutral–sad confusion error from 74 to 69, but at a high computational cost. All five models must be downloaded at inference time, which requires about 6–8 GB of VRAM on the GPU, and each frame of the input data stream must undergo five forward passes. This makes the deployment of real-time systems on resource-constrained hardware like mobile devices, embedded systems, or even low-cost clinical screening devices difficult without substantial model compression, knowledge distillation, or architecture pruning. An interesting future direction is to see whether a single light model that can be distilled from the ensemble can be used to approximate its performance at a smaller fraction of its computational cost.

5.6.1.6 Limitations of Transformer-Based Architectural Dependence

The three ensemble members use self-attention mechanisms, which are a key feature of the transformer architecture: ViT-Base/16 (86.6M), Swin-Small (49.6M), and CoAtNet-2 (273.9M, a hybrid CNN-Transformer). Transformers are well-suited to large-scale vision tasks, but it is known that they need significantly larger training sets than convolutional networks to learn to generalize well, without inductive biases like translation equivariance and locality. Only 8,392 images with five emotion classes are available for training the transformers in this work. The models were initialized with ImageNet-21K (or IN-12K) pre-trained weights to alleviate the limited data, but it is unclear how well transformer-specific features transfer to the relatively small and domain-specific Omega facial expression distribution. Moreover, the quadratic memory requirement of self-attention with respect to the sequence length, although less important for the 224×224 fixed-size image patches, adds to the high number of parameters and inference time of ViT-Base/16 and Swin-Small compared to purely convolutional alternatives. However, large pretrained transformers are used in architectural dependence, which might make it difficult for some settings to be reproduced without access to large pretrained weights or sufficient computation power for fine-tuning.

5.6.2 Depression Classifier Discussions

5.6.2.1 Clinical Interpretation of Recall Asymmetry

The deliberate clinical design choice is the stacking ensemble’s stacking asymmetry, with a depression recall score of 0.7529 compared with a normal recall score of 0.6719. The cost of the two kinds of errors is not the same in a depression screening context. A false negative is a case where a person who is feeling depressed is assigned to the normal group category, which means that they will not receive intervention or referral, and may have significant negative implications for well-being. A false positive (a non-depressed person being diagnosed as depressed) is undesirable because it leads to unnecessary worry and may lead to an unnecessary clinical consultation, but it’s not as bad as a false negative (missed diagnosis). This imbalance is explicitly encoded in the manner in which the classes are weighted when training classifiers: samples of depression get a higher penalty for being misclassified. The recall of 0.0811 obtained is a tradeoff that was measured and balanced between clinical safety and accuracy. This design principle aligns with the existing literature on mental health screening, where the focus is on sensitivity rather than specificity for automated screening tools that are used in addition to clinical assessment.

5.6.2.2 Limitations of Dataset Size

The depression classifier’s major drawback is the number of training videos (414). This limitation was due to several factors. When we manually analysed the original downloaded 756-video corpus, we found that not an insignificant portion of videos was only partially facial, meaning that the subject’s face was recorded at extreme angles or only the lower part of the face, such as the lips or nose, was visible to the camera. This type of video is not suitable for detecting depression, as the frozen FER ensemble was trained using full frontal face images and giving emotion probability sequences for partial face images out of its trained distribution is unreliable. Moreover, there were many videos that did not have any valid preprocessed frames at all because the subjects did not look at the camera, the videos were shot in dark conditions, or the subjects looked away from the camera. A minimum quality ten-frame filter was also applied to the videos to remove those videos that were too short to compute reliable temporal features. These factors combined led to a quality-validated working set of 756 videos, of which 414 videos formed the training fold.

This is reflected in the fact that the sequence-based models Bi-LSTM and Bi-GRU did not achieve the same performance as the classical ensemble methods, and this is consistent with the literature in that recurrent architectures need significantly larger training sets than the classical ensemble models to prevent overfitting. Depression is different from person to person, from culture to culture, and different in intensity, and the 414-video training set also does not cover the diversity of depressive behaviour patterns the classifier has been exposed to.

Further, as described by Yoon et al. (2022) [58], the D-Vlog dataset was captured during the COVID-19 pandemic, and in the time of social isolation, from January 2020 to January 2021. This time frame and context might have had an impact on the behaviour of depressed and normal vloggers, which makes it difficult to generalise the results to

other time periods or social conditions. In their extension of the D-Vlog corpus, Moon and Bhattacharyya (2024) [44] noted similar generalisability concerns and highlighted the need for a diverse range of demographics in the study of depression. Another issue identified in the original D-Vlog is the gender imbalance of 639 females and 322 males in the entire dataset of 961 videos, which may impact how the classifiers perform across different genders, and hence further stratification for evaluation is recommended in the future.

5.6.2.3 Value of Multimodal Fusion

A significant architectural choice in this study is to include acoustic features as well as visual expression features, as there is strong evidence for multimodal approaches as discussed in Chapter 2. For depression detection, Ashraf et al. (2020) [4] indicated that the multimodal approaches beat unimodal approaches in general. This is confirmed by the D-Vlog literature — all three major datasets (TAMFN, STST, Gimeno-Gomez et al. (2024) [19], MDD-Net) employ an acoustic and visual modality, and all outperform unimodal baselines. This paper focuses on the 5 cross-modal features that, in this research, are particularly motivated by clinical literature: Pearson correlation between emotion probability sequences and acoustic feature trajectories. People with low mood demonstrate that they do not synchronise their facial expression and their vocal affect; they might speak with decreased prosody and then have a flat face, or they might have a negative facial expression but a flat voice. These features are statistically relevant to the depression label over and above the 250 visual and 500 acoustic features, and are captured directly in the cross-modal features that survive the two-stage selection process in the final 195-feature set.

5.6.2.4 Limitations of the Feature Aggregation Approach

The handcrafted feature aggregation approach was required due to the small training set size, but it also comes with a fundamental limitation as compared to end-to-end temporal deep learning approaches. Taking an emotion sequence of variable length and collapsing it into a fixed-length feature vector of 195 dimensions, even though the dimensions correspond to the different emotions, results in the loss of fine-grained information about the temporal order within the sequence. The classifier does not directly model the specific trajectory of the emotions but only provides summary statistics such as means, standard deviations, differences between the segments, and autocorrelation coefficients for the emotional trajectory of the video. All the above sequential dynamics of emotional change are compressed into distributional descriptors, which summarize what occurred during the video but not exactly when and in what order.

This constraint is reflected in the performance difference when compared to methods that use explicit spatio-temporal sequence modelling on the D-Vlog dataset. By representing the temporal structure of expression and acoustic sequences directly instead of fixed-length vectors, both STST by Tao et al. (2024) [53] and MDD-Net by Haque et al. (2025) [26] obtain significantly higher F1 scores. Future extensions of the D-Vlog corpus, as well as additional depression video data sets annotated by a trained clinician, will contribute to the collection, which in turn will lead to the progress from handcrafted

feature aggregation to end-to-end temporal modelling, to boost the performance of depression detection beyond the ceiling already achieved in this study. The feature aggregation approach taken here is best seen as a principled solution suitable for the current scale of data, and for training to grow sufficiently to support sequence-level deep learning methods; their use is suggested.

5.6.2.5 Overall System Implications

The end-to-end pipeline (raw video input, face preprocessing, frozen FER ensemble inference, multimodal feature extraction, feature selection, and stacking classification) represents a completely automated, non-invasive depression screening system, which can be applied to naturalistic vlog videos without the need for clinical interview data, physiological sensors, or text transcripts. The high test accuracy of 71.81%, precision of 0.7529, and AUC-ROC of 0.7722 make this a useful screening tool and not a diagnostic tool, designed to identify people who may require clinical evaluation, not to make a diagnosis. With 75.3% depression correct, the system is clinically relevant and has a meaningful sensitivity to correctly identify around three-quarters of depressed people in the test fold, for a first-pass automated screening tool. The major contributions of this work to the D-Vlog literature are the purpose-designed hybrid FER ensemble that outperforms single-model extractors for emotion probability features due to synthesizing richer and more clinically calibrated features, the comprehensive 755-dimensional multimodal feature set that systematically encodes visual, acoustic and cross-modal behavioral features, and the empirical evidence that classical stacking ensembles with principled feature selection can achieve competitive performance on limited datasets when deep learning approaches are limited by data scarcity.

Chapter 6

Conclusion

6.1 Summary of Findings

This thesis introduced a fully automatic multimodal depression detection pipeline based on two related modules: an ensemble of multimodal facial expression recognition modules and a multimodal depression classifier trained with D-Vlog data.

To evaluate the facial expression recognition part, 9 deep learning architectures from three architectural families (pure convolutional, pure transformer, and hybrid CNN-attention) were trained and evaluated on the preprocessed OAHEGA dataset with the same experimental settings. The evaluation showed that the modern architectures that were trained on ImageNet-21k and ImageNet-22k significantly outperform the classical CNN baselines, with accuracy of 87.43% and 89.35% respectively for ResNet-50 and VGG-16 compared to the top individual model CoAtNet-2 with 92.44%. To tackle the clinically important neutral-sad confusion boundary, a custom Neutral-Sad Separation loss was added, resulting in measurable improvements in neutral-sad errors in all models that included it. The test accuracy for the five best models (ConvNeXt-Small, CoAtNet-2, ViT-Base/16, Swin-Small, ConvNeXt-Base) was 93.13%, the macro F1 was 0.9315, and the neutral-sad confusion count was 69, which is the lowest of all models tested. McNemar’s test showed that the improvement of the ensemble over the best individual model ConvNeXt-Small is statistically significant at 95% confidence level with a p-value of 0.00097.

In the classification of depression part, the probability sequences of per-frame emotions were extracted from 667 preprocessed D-Vlog videos using the frozen hybrid FER ensemble. These sequences, along with pre-extracted 25-dimensional acoustic descriptors and 5 cross-modal synchrony features, were then combined into a multimodal feature vector of 755 dimensions per video, which was then reduced to 195 features using a two-stage process of Mutual Information and RFECV. The results showed that the sequence-based models performed worse than classical models, due to the small training set of 414 videos in these models, and the stacking ensemble with MLP, CatBoost, and Random Forest classifiers with a Logistic Regression meta-learner performed best, thanks to its learned weighting mechanism and better generalization. The held-out test fold (149 videos) performance of the stacking (MLP+CB+RF→LR) ensemble was 71.81%, 0.7529, 0.7124, 0.7722, 0.7529, and 0.6719 for accuracy, precision, macro F1 score, AUC-ROC, depression recall, and normal recall, respectively, and a 2.23% drop on the held-out test fold

compared to the validation fold, indicating that the ensemble performed well and was very robust in comparison to all other combinations of ensembles that were tested.

6.2 Contributions to the Field

In this thesis, the problem fields of facial expression recognition and automatic depression detection are further advanced by the contributions of five theses; three of which are novel in their respective problem fields.

The most technical and novel of them is an auxiliary training objective, called Custom Neutral-Sad Separation Loss, which was specifically designed for the clinical domain of depression detection. This type of loss function has not been used in the published literature with the OAHEGA data. The loss explicitly punishes uncertainty in the neutral-sad decision region during training, while in reality, flat and blunted affect, the defining features of depressive symptomology, occur exactly at the neutral-sad decision region. Empirical findings show that the loss decreases neutral-sad confusion across all the architectures tested (CoAtNet-2, ConvNeXt-Base, Swin-Small, and ViT-Base/16), and is fully architecture-agnostic with no restrictions for any specific architecture, nor does it rely on specific components of the FER training pipeline; it can additionally be applied to any FER training pipeline where the neutral-sad boundary is clinically or practically significant.

The second contribution is a design and empirical test of a hybrid FER ensemble specially tuned for neutral-sad discrimination rather than overall accuracy. The ensemble, which combines CNN, CNN-attention, and pure transformer models, performs better than all its constituent models with a lower confusion count of 32 neutral-37 sad on OAHEGA and test accuracy of 93.13%. This provides a principled and clinically driven design framework for FER ensembles, where the architecture of the ensembles is jointly determined by the context of each ensemble and the accuracy ranked model selection, which is shown to favor better boundary discrimination than the model selection alone.

The third contribution is the creation of a 755-dimensional multimodal feature set, based on five features of cross-modal synchrony derived from Pearson correlations between per-frame emotion probability sequences and acoustic signals, which reflects the temporal co-variation between facial and vocal behaviour. The D-Vlog dataset has not been used in a previous study as input to a multi-architecture FER ensemble, and no previous study has incorporated cross-modal synchrony features based on the ensemble-generated emotion probability sequences. The feature set also contains clinically relevant visual descriptors that are obtained directly from the probability output of the FER ensemble: anhedonia ratio, flat affect proportion, blunted affect ratio, and entropy slope, linking the computational emotion recognition with formal symptomology of depression.

The fourth contribution, in fact, is an empirical example that shows that classical stacking ensembles with principled two-stage feature selection (Mutual Information followed by RFECV) can achieve competitive depression detection performance on the limited accessible subset of the D-Vlog data set, outperforming the original baseline of Yoon et al. (2022) and even matching the performance of more complex temporal deep learning techniques that utilize the full data set and more comprehensive modalities. This result is important in research environments in which there is limited data, which makes sequence

models impractical to use.

The fifth contribution is a pipeline that preprocesses the OAHEGA and D-Vlog datasets, which is fully documented and validated on the problem domain, applying MTCNN face alignment, CLAHE illumination normalization, Difference of Gaussians shading removal and MediaPipe ROI masking with the detection parameters optimized for vlog-style natural images. More importantly, the D-Vlog phase of the pipeline includes a content validation system based on the depression, which goes beyond simple image quality filtering. A manual examination of the original 961 videos found that a significant number of videos were not suitable for the depression detection task for several reasons, such as no visible face in the video, occlusion of the face in the in-distribution videos that prevents the in-distribution FER from being considered, and videos too short to provide meaningful temporal features for depression. These cases are systematically identified and eliminated by the pipeline using two principles: video where no valid preprocessed frames are found are excluded, and temporal features such as autocorrelation, statistics computed for each segment, and progression markers are not considered statistically meaningful if the number of frames is less than 10. This 667-video subset is thus neither merely a largest available subset of D-Vlog, but a quality-validated subset of videos that meet strict criteria for facial visibility, FER in-distribution, and temporal sufficiency. This content-based validation framework is not based on data availability but rather on the specific computational and clinical needs of depression detection—thereby highlighting and offering a reproducible quality standard for future research on this corpus.

6.3 Recommendations for Future Work

There are several avenues for developing and enhancing the system presented in this thesis.

The first direction is an extension of the depression classifier to predict depression severity instead of binary depressed or normal tags. The AVEC benchmark series and models by Li et al. (2020) [38] (DRR DepressionNet) and Zhou et al. (2020) [60] (DepressNet) show that continuous severity prediction is more clinically meaningful than binary classification, and the model is capable of monitoring changes in depression severity over time instead of generating a binary screen. The multimodal feature set created in this study is well equipped for a regression-based severity prediction system with minimal architectural changes, especially in terms of the temporal progression features and the statistics per segment.

The second direction is related to the study of the system’s predictions concerning demographic equity. There is a significant imbalance between female and male participants in the D-Vlog dataset (639 female and 322 male), and as mentioned by Gaya-Morey et al. [17], FER datasets are usually small in size, particularly for children and elderly people. Future work needs to test the performance of this system with gender, age, and cultural subgroups to detect and target any systematic differences in the accuracy of depression detection before deployment in the real world.

A final direction is the clinical validation of the full pipeline over a prospective dataset of clinically verified depression diagnoses, ideally with the ground-truth label being a standardized instrument like a PHQ-9 or BDI questionnaire, rather than being self-reported in

a vlog. Although the D-Vlog annotation protocol was carefully designed, it is not based on clinical diagnosis, which will help address labelling uncertainty. Any such use of the system as a real-world depression screening tool would have to be preceded by such validation.

Bibliography

- [1] G. Agrawal, Unnati, and R. Bidwe, "Automatic facial expression recognition using advanced transfer learning," in *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, 2023, pp. 450–458. DOI: 10.1145/3607947.3608047.
- [2] A. Ahmed et al., "Machine learning models to detect anxiety and depression through social media: A scoping review," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100 066, 2022. DOI: 10.1016/j.cmpbup.2022.100066.
- [3] A. Ahmed, R. Sultana, M. T. R. Ullas, M. Begom, M. M. I. Rahi, and M. A. Alam, "A machine learning approach to detect depression and anxiety using supervised learning," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering*, 2020, pp. 1–6. DOI: 10.1109/CSDE50874.2020.9411642.
- [4] A. Ashraf, T. S. Gunawan, B. S. Riza, E. V. Haryanto, and Z. Janin, "On the review of image and video-based depression detection using machine learning," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, pp. 1677–1684, 2020. DOI: 10.11591/ijeecs.v19.i3.pp1677-1684.
- [5] T. Aslam, S. Qadri, M. Shehzad, S. F. Qadri, A. Razzaq, S. S. Muhammad, S. Ahmad, S. A. Nawaz, and N. Ahmad, "Emotion based facial expression detection using machine learning," *Life Science Journal*, vol. 17, no. 8, pp. 35–43, 2020. DOI: 10.7537/marslsj170820.06.
- [6] A. R. Babu, Mohebbanaaz, S. Chandrika, G. Rajeswari, and U. V. Pavani, "Face emotion recognition using deep neural network," in *2025 IEEE 14th International Conference on Communication Systems and Network Technologies*, 2025, pp. 547–551. DOI: 10.1109/CSNT64827.2025.10967857.
- [7] A. Badhan, P. Kaur, V. Kumar, Priyanshu, and S. Sharma, "Implementing human emotion detection using customized CNN model in machine learning," in *2025 Third International Conference on Augmented Intelligence and Sustainable Systems*, 2025, pp. 1783–1787. DOI: 10.1109/ICAISS61471.2025.11042127.
- [8] B. Bakariya, K. K. Mohbey, A. Singh, H. Singh, P. Raju, and R. Rajpoot, "An efficient model for facial expression recognition with music recommendation," *National Academy Science Letters*, vol. 47, no. 3, pp. 267–270, 2024. DOI: 10.1007/s40009-023-01346-4.
- [9] B. Bakariya, A. Singh, H. Singh, P. Raju, R. Rajpoot, and K. K. Mohbey, "Facial emotion recognition and music recommendation system using CNN-based deep learning techniques," *Evolving Systems*, vol. 15, pp. 641–658, 2024. DOI: 10.1007/s12530-023-09506-z.

- [10] V. S. Bakkialakshmi, N. Kar, and V. Kumar, "The digital mirror: Reflecting human emotions through machine learning-based facial gesture recognition," in *2024 1st International Conference on Cognitive, Green and Ubiquitous Computing*, 2024, pp. 1–6. DOI: 10.1109/IC-CGU58078.2024.10530705.
- [11] M. Bhagyalakshmi, *Mental health detection using facial emotion recognition*, Volume 9, Issue 2, p. 555, 2024. [Online]. Available: <https://www.ijnrd.org/papers/IJNRD2402169.pdf>.
- [12] D. Choi, G. Zhang, D. E. Kim, and J. Jung, "Depression diagnosis algorithm based on 2-stream CNN using facial image," in *2023 IEEE/ACIS 23rd International Conference on Computer and Information Science*, 2023, pp. 43–47. DOI: 10.1109/ICIS57766.2023.10210234.
- [13] A.-L. Cîrneanu, D. Popescu, and D. Iordache, "New trends in emotion recognition using image analysis by neural networks: A systematic review," *Sensors*, vol. 23, p. 7092, 2023. DOI: 10.3390/s23167092.
- [14] L. R. Demenescu, R. Kortekaas, J. A. den Boer, and A. Aleman, "Impaired attribution of emotion to facial expressions in anxiety and major depression," *PLOS ONE*, vol. 5, no. 12, e15058, 2010. DOI: 10.1371/journal.pone.0015058.
- [15] K. R. Gavhale and U. H. Gawande, "An analysis of recommendation system based on multimodal emotion detection using facial expression and voice modulation," in *2024 International Conference on Emerging Trends in Engineering and Management Sciences*, 2024. DOI: 10.1109/ICETEMS64039.2024.10965013.
- [16] M. Gavrilescu and N. Vizireanu, "Predicting depression, anxiety, and stress levels from videos using the facial action coding system," *Sensors*, vol. 19, no. 17, p. 3693, 2019. DOI: 10.3390/s19173693.
- [17] F. X. Gaya-Morey, C. Manresa-Yee, C. Martinie, and J. M. Buades-Rubio, "Evaluating facial expression recognition datasets for deep learning: A benchmark study with novel similarity metrics," *arXiv preprint*, 2025. DOI: 10.48550/arXiv.2503.20428.
- [18] G. Gilanie et al., "An automated and real-time approach of depression detection from facial micro-expressions," *Computers, Materials & Continua*, vol. 73, no. 2, 2022. DOI: 10.32604/cmc.2022.028229.
- [19] D. Gimeno-Gomez, A.-M. Bucur, A. Cosma, C.-D. Martínez-Hinarejos, and P. Rosso, "Reading between the frames: Multi-modal depression detection in videos from non-verbal cues," *arXiv preprint*, 2024. DOI: 10.48550/arXiv.2401.02746.
- [20] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013, pp. 1–8. DOI: 10.1109/FG.2013.6553748.
- [21] A. K. Goel, A. Jain, C. Saini, Ashutosh, R. Das, and A. Deep, "Implementation of AI/ML for human emotion detection using facial recognition," in *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications*, Retracted, 2022, pp. 511–515. DOI: 10.1109/ICCCMLA56841.2022.9989091.

- [22] A. Gupta, “Using machine learning to interpret and identify facial expressions and emotions,” *International Journal of Innovative Science and Research Technology*, vol. 9, no. 11, pp. 1413–1417, 2024. DOI: 10.38124/ijisrt/IJISRT24NOV819.
- [23] P. Gupta, S. K. Shukla, B. Yadav, R. Bhatia, H. R. Dereddy, and R. C. Joshi, “ECA-ResNet: An enhanced channel-spatial attention residual network for facial emotion recognition in mental health screening,” in *2026 IEEE Madhya Pradesh Section Conference*, Gwalior, India, 2026, pp. 895–901. DOI: 10.1109/MPCON69668.2026.11508582.
- [24] H. Hadjar, B. Vu, and M. Hemmje, “TheraSense: Deep learning for facial emotion analysis in mental health teleconsultation,” *Electronics*, vol. 14, no. 3, p. 422, 2025. DOI: 10.3390/electronics14030422.
- [25] M. R. Haque, M. M. Islam, S. M. T. U. Raju, H. Altaheri, L. Nassar, and F. Karray, “MMFformer: Multimodal fusion transformer network for depression detection,” in *2025 IEEE International Conference on Systems, Man, and Cybernetics*, Vienna, Austria, 2025, pp. 5912–5919. DOI: 10.1109/SMC58881.2025.11342694.
- [26] M. R. Haque, M. M. Islam, S. M. T. U. Raju, H. Altaheri, L. Nassar, and F. Karray, “MDD-Net: Multimodal depression detection through mutual transformer,” in *2025 IEEE International Conference on Systems, Man, and Cybernetics*, Vienna, Austria, 2025, pp. 6592–6599. DOI: 10.1109/SMC58881.2025.11342588.
- [27] A. Jain and R. Kumar, “Machine learning based anxiety detection using physiological signals and context features,” in *2024 2nd International Conference on Advancement in Computation and Computer Technologies*, 2024, pp. 116–121. DOI: 10.1109/InCACCT61598.2024.10551003.
- [28] M. L. Joshi and N. Kanoongo, “Depression detection using emotional artificial intelligence and machine learning: A closer review,” *Journal of Mental Health and Human Behaviour*, vol. 27, no. 1, pp. 10–18, 2022. DOI: 10.4103/jmhbb.jmhbb_22_21.
- [29] A. R. Khan, “Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis, and remaining challenges,” *Information*, vol. 13, no. 6, p. 268, 2022. DOI: 10.3390/info13060268.
- [30] X. Kong, Y. Yao, C. Wang, Y. Wang, J. Teng, and X. Qi, “Automatic identification of depression using facial images with deep convolutional neural network,” *Medical Science Monitor*, vol. 28, e936409, 2022. DOI: 10.12659/MSM.936409.
- [31] V. Kovenko and V. Shevchuk, “OAHEGA: From research to production of mobile expression detection application,” *Research Square*, 2022, Preprint. DOI: 10.21203/rs.3.rs-1461348/v1.
- [32] V. Kovenko and V. Shevchuk, *OAHEGA: Emotion recognition dataset*, version V1, 2021. DOI: 10.17632/5ck5zz6f2c.1.
- [33] W. L. Ku and H. Min, “Evaluating machine learning stability in predicting depression and anxiety amidst subjective response errors,” *Healthcare*, vol. 12, no. 6, p. 625, 2024. DOI: 10.3390/healthcare12060625.
- [34] G. Kumar, T. Das, and K. Singh, “Early detection of depression through facial expression recognition and electroencephalogram-based artificial intelligence-assisted graphical user interface,” *Neural Computing and Applications*, vol. 36, pp. 6937–6954, 2024. DOI: 10.1007/s00521-024-09437-z.

- [35] K. J. Lakshmi Bai, A. Harshita, T. Gaddam, S. Mirtipati, and D. Peyyala, “Enhancing facial emotion detection for depression detection using image integration by hybrid dataset approach,” in *2024 IEEE 4th International Conference on ICT in Business Industry and Government*, 2024, pp. 1–6. DOI: 10.1109/ICTBIG64922.2024.10911294.
- [36] Y.-S. Lee and W.-H. Park, “Diagnosis of depressive disorder model on facial expression based on fast R-CNN,” *Diagnostics*, vol. 12, no. 2, p. 317, 2022. DOI: 10.3390/diagnostics12020317.
- [37] C. Li, S. Chen, Q. Jin, and A. Hauptmann, “Multimodal spatiotemporal representation for automatic depression level detection,” in *Proceedings of the 9th International Audio/Visual Emotion Challenge*, 2019, pp. 65–72. DOI: 10.1145/3347320.3357682.
- [38] X. Li, W. Guo, and H. Yang, “Depression severity prediction from facial expression based on the DRR DepressionNet network,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine*, 2020, pp. 2757–2764. DOI: 10.1109/BIBM49941.2020.9313597.
- [39] Y. Mahayossanunt, N. Nupairoj, S. Hemrungronj, and P. Vateekul, “Explainable depression detection based on facial expression using LSTM on attentional intermediate feature fusion with label smoothing,” *Sensors*, vol. 23, no. 23, p. 9402, 2023. DOI: 10.3390/s23239402.
- [40] W. C. de Melo, E. Granger, and A. Hadid, “Combining global and local convolutional 3D networks for detecting depression from facial expressions,” in *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition*, 2019, pp. 1–8. DOI: 10.1109/FG.2019.8756568.
- [41] P. Mishra, A. S. Verma, P. Chaudhary, and A. Dutta, “Emotion recognition from facial expression using deep learning techniques,” in *2024 IEEE 9th International Conference for Convergence in Technology*, 2024, pp. 1–6. DOI: 10.1109/I2CT61223.2024.10543313.
- [42] B. Moghe, M. Kachhara, et al., “Multimodal emotion analysis for depression detection: Integrating facial expression and speech recognition,” in *2024 Second International Conference on Inventive Computing and Informatics*, 2024, pp. 37–42. DOI: 10.1109/ICICI62254.2024.00015.
- [43] M. Mohan, A. K. Abhinav, A. Ashok, A. V. Akhil, and P. R. Achinth, “Depression detection using facial expression and sentiment analysis,” in *2021 Asian Conference on Innovation in Technology*, 2021, pp. 1–6. DOI: 10.1109/ASIANCON51346.2021.9544819.
- [44] P. Moon and P. Bhattacharyya, “We care: Multimodal depression detection and knowledge infused mental health therapeutic response generation,” *arXiv preprint*, 2024. DOI: 10.48550/arXiv.2406.10561.
- [45] A. Mustafa, S. Bhatia, M. Hayat, and R. Goecke, “Heart rate estimation from facial videos for depression analysis,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction*, 2017, pp. 498–503. DOI: 10.1109/ACII.2017.8273645.

- [46] R. Nannapaneni. “Human emotion recognition using machine learning,” Accessed: Jun. 7, 2026. [Online]. Available: https://www.academia.edu/90890014/Human_Emotion_Recognition_Using_Machine_Learning?uc-sb-sw=4093683.
- [47] A. Poongodai, V. Bhavitha, D. Hemanth, G. Jahnavi, and S. Huzma, “Human emotion detection through real-time facial expressions using deep learning,” in *2025 6th International Conference on Inventive Research in Computing Applications*, 2025, pp. 1298–1303. DOI: 10.1109/ICIRCA65293.2025.11089747.
- [48] N. Raut, “Facial emotion recognition using machine learning,” M.S. thesis, San José State University, 2018. DOI: 10.31979/etd.w5fs-s8wd.
- [49] Y. Shang et al., “LQGDNet: A local quaternion and global deep network for facial depression recognition,” *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2557–2563, 2021. DOI: 10.1109/TAFFC.2021.3139651.
- [50] A. Sharma, V. Bajaj, and J. Arora, “Machine learning techniques for real-time emotion detection from facial expressions,” in *2023 2nd Edition of IEEE Delhi Section Flagship Conference*, Rajpura, India, 2023, pp. 1–6. DOI: 10.1109/DELCON57910.2023.10127369.
- [51] S. Singh and N. Srivastava, “Real-time depression detection using emotion recognition techniques,” *International Journal of Computer Applications*, vol. 186, no. 66, pp. 40–48, 2025. DOI: 10.5120/ijca2025924467.
- [52] H. Sun, J. Liu, S. Chai, Z. Qiu, L. Lin, X. Huang, and Y. Chen, “Multi-modal adaptive fusion transformer network for the estimation of depression level,” *Sensors*, vol. 21, no. 14, p. 4764, 2021. DOI: 10.3390/s21144764.
- [53] Y. Tao, M. Yang, Y. Wu, K. Lee, A. Kline, and B. Hu, “Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer,” *Digital Communications and Networks*, vol. 10, pp. 577–585, 2024. DOI: 10.1016/j.dcan.2023.03.007.
- [54] M. Tasnim, R. D. Ramos, E. Stroulia, and L. A. Trejo, “A machine-learning model for detecting depression, anxiety, and stress from speech,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 7085–7089. DOI: 10.1109/ICASSP48485.2024.10446567.
- [55] S. Telsang, S. Patil, R. Patil, S. Patil, and S. Patil, “CNN-driven novel facial expression detection and analysis,” in *2024 5th International Conference on Smart Electronics and Communication*, 2024, pp. 1657–1662. DOI: 10.1109/ICOSEC61587.2024.10722627.
- [56] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, *Suppressing uncertainties for large-scale facial expression recognition*, 2020. arXiv: 2002.10392 [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2002.10392>.
- [57] Z. Xu, Y. Gao, F. Wang, and X. Li, “Depression detection methods based on multimodal fusion of voice and text,” *Scientific Reports*, vol. 15, p. 21 907, 2025. DOI: 10.1038/s41598-025-03524-4.
- [58] J. Yoon, C. Kang, S. Kim, and J. Han, “D-Vlog: Multimodal vlog dataset for depression detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 12 226–12 234. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/21483>.

- [59] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, “TAMFN: Time-aware attention multimodal fusion network for depression detection,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 669–679, 2023. DOI: 10.1109/TNSRE.2022.3224135.
- [60] X. Zhou, K. Jin, Y. Shang, and G. Guo, “Visually interpretable representation learning for depression recognition from facial images,” *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 542–552, 2020. DOI: 10.1109/TAFFC.2018.2828819.