ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 03, March 2025

Multimodal Emotion Recognition Using CNN, Transformer Model, and Deep Learning for Text, Speech, and Facial Analysis

[1] Dr. B Abirami, [2] Krishna Pramod Palekar, [3] Karan Nair, [4] Hariharan T, [5] M Antony Raj, [6] Harish G

[1] Assistant Professor, Department of Computer Science and Engineering SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India

[2] [3] [4] [5] [6] CSE with specialization AIML SRM Institute of Science and Technology Ramapuram, Chennai, India

Corresponding Author Email: [1] abiramis3@srmist.edu.in, [2] kp1014@srmist.edu.in, [3] kn3282@srmist.edu.in, [4] ht5544@srmist.edu.in, [5] mr7249@srmist.edu.in, [6] hg6034@srmist.edu.in

***Abstract*— A significant issue in affective computing is the identification of facial expressions, which influences behavioral analysis, mental health evaluation, and human-computer interaction applications. The issue of weak generalization emerges from conventional deep learning models' inability to effectively handle temporal and spatial relationships. The sophisticated framework for facial emotion recognition presented in this study is based on three distinct architectures: (1) a CNN structure augmented with advanced feature extraction and regularization methods; (2) a CNN-LSTM combined model intended to capture sequential patterns in facial expressions; and (3) a refined CNN that emphasizes precise spatial feature extraction. To achieve optimal classification accuracy, robustness, and generalization, each model is subjected to individual tuning. Experiments performed on standard datasets reveal significant enhancements in accuracy, recall, and precision when compared to traditional techniques.Findings indicate that more complex models and consideration of sequence in modeling significantly improve facial emotion recognition, paving the way for more resilient and instantaneous affective computing.**

***Index Terms*— *CNN-LSTM Hybrid Model, Convolutional Neural Network (CNN), Facial Emotion Recognition, Feature Extraction, Human-Computer Interaction, Sequence Modeling*.**

I. INTRODUCTION

A crucial element of human-computer interaction is the capability to identify emotions, enabling computers to more effectively comprehend and react to human emotions. Applications for emotion detection are found in many different domains, including e-learning, customer service, mental health monitoring, and human-robot interaction. Traditionally, unimodal approaches that primarily focus on speech, text-based clues, or facial expressions have been used to recognize emotions. However, because they lose contextual information that is available in numerous modalities, these one-modal approaches typically have worse accuracy.

Facial Emotion Recognition (FER) backed by Convolutional Neural Networks (CNNs) has been transformed by deep learning advancements. CNNs achieve notable improvements in classification accuracy by successfully extracting spatial characteristics from facial photos. However, because feelings are conveyed not just by facial expressions but also through the tone of voice and the general atmosphere of the text, depending exclusively on facial signals might not completely represent a person's emotional state.Mel-frequency cepstral coefficients (MFCCs), spectrograms, and deep learning models (CNNs, LSTMs) have also shown promise in speech emotion recognition; nevertheless, they fall short when emotional changes are modest or obscured by background noise. While transformer-based models like BERT and RoBERTa are better at collecting context meaning, they struggle to recognize emotions in text-based data when emotional context is missing.

Traditional unimodal emotion recognition methods that rely solely on text, speech, or facial expressions have drawbacks such as losing prosodic information in text, speech fluctuation, and vagueness in expressions. In order to overcome this difficulty, we propose a hybrid deep learning model that combines facial, speech, and text-based emotion recognition to improve classification performance. Our system leverages CNN-LSTM architectures for voice emotion detection on the TESS dataset, transformer-based RoBERTa for text sentiment analysis, and CNN-based feature extraction for facial emotion from the FER2013 dataset. Deep Neural Networks (DNNs) are then used to integrate the retrieved characteristics in order to create a reliable multimodal emotion classification system. We demonstrate improved precision and reliability in identifying human emotions by (1) proposing a multimodal emotion recognition model that outperforms unimodal models, (2) proposing a new DNN-based fusion strategy to efficiently fuse multimodal features, (3) thoroughly benchmarking our model against unimodal CNN, LSTM, and transformer-based models, and (4) reporting 79.3% weighted accuracy on FER2013. Our study improves the stability of emotion

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 03, March 2025

identification systems by integrating complementing affective cues from several modalities, making them applicable to emotional computing, psychological assessment, and interaction between humans and computers.

II. LITERATURE REVIEW

Recently, emotion recognition has gained significant interest, particularly due to advancements in deep learning methodologies. To enhance the effectiveness of emotion detection systems, a number of modalities have been investigated, including text, audio, and facial expressions.

This study describes a deep facial emotion recognition system that is based on learning and utilizes the FER-2013 dataset. [1]. By enhancing classification accuracy through the optimization of a VGG-16 CNN model, their method advanced the emerging field of face emotion recognition. Similarly, utilizing FER-2013 and other datasets, [2] proposed a deep neural network for facial expression analysis and assessed its efficacy. Their research shows how effectively CNN architectures can pick up strong face emotion recognition components. Additionally, [7] demonstrated a novel approach to real-time emotion categorization by investigating the use of AI-driven intelligent video analytics for human sentiment recognition. To enhance the accuracy of emotion recognition in facial videos, a fusion of multi-view feature expressions is utilized. technique was put forth in [8].

For speech identifying emotions (SER), Scientists have utilized deep learning techniques. models that were trained on datasets like RAVDESS, TESS, and SAVEE. In the study [3], a hybrid CNN-LSTM model with a high accuracy of 89.26% for a range of SER datasets was introduced. The model was trained using Mel Spectrograms as an input feature. In a similar vein, [4] proposed aa deep learning model that utilizes self-attention integrated 2D CNN and LSTM networks by combining the data from RAVDESS, SAVEE, and TESS, increasing classification accuracy to 90%. It highlights the significance of feature extraction methods for speech-based emotion recognition.

Text-based emotion recognition has been greatly improved with the An overview of transformer-based architectures. Since its debut in [5], Bidirectional Encoder Representations from Transformers, or BERT, has become widely used in a range of Natural Language Processing (NLP) applications, including sentiment and emotion analysis. They achieved state-of-the-art performance by improving the precision of emotional classification by utilizing contextual embeddings. Transformers' achievement in text emotion recognition has sparked more research in this field.

More and more research is focusing on multimodal emotion identification, which enhances emotion detection by combining several data sources like audio, text, and facial expressions. The publication [6] outlines the significance of multimodal fusion in affective computing and presents several fusion approaches and their efficacy. According to their research, using many modalities increases the resilience of emotion recognition systems. Additionally, [9] highlighted novel techniques to increase detection efficiency by introducing an AI-based approach to face emotion identification. In a recent work, [10] used the Emognition dataset to create a CNN-based human face emotion detection system, which contributed to enhancing the model's accuracy and generalization.

Emotion recognition researchers have made strides but still contend with a myriad of issues. There is still the problem of data heterogeneity, which is frequently encountered, as many techniques are required to be matched as well as merged for analysis to be effective. Real time analysis is yet another hurdle due to the vast amount of processing power required from existing deep learning frameworks. Emotion interpretation is complicated because the context has to be taken into consideration. Emotion identification is among the most difficult tasks to achieve. Multimodal emotion recognition still poses a significant challenge in the development in AI systems attempting to devise steps to efficiently unify all these factors. Researchers should focus on developing models that can perform accurate real-time processing, precisely contextualize emotion recognition, and efficiently integrate multimodal data.

III. PROCESS FLOW

A. Exploratory Research Data Analysis (ERDA)

Understanding the structure, distribution, and quality of multimodal emotion recognition data requires a thorough understanding of the Exploratory Research Data Analysis (ERDA) stage. The datasets RAVDESS (voice), FER2013 (face), and GoEmotions (text) are used in this investigation. Each set of data undergoes separate processing before being combined into features in the multimodal fusion model. To determine the distribution of emotions for different speakers and intensities, a preliminary analysis is conducted for the speech emotion dataset (RAVDESS). Mel-Frequency Cepstral Coefficients (MFCCs), waveform plots, and spectrograms are used to identify patterns in emotional speech. Additionally, noise analysis is done, and methods for balancing datasets such as pitch shift process, time stretch methods, and noise implementation have been investigated. Images are analyzed for class imbalance, resolution consistency, and variance in facial expressions for the face emotion dataset. In order to improve model generalization, data augmentation techniques are used and some of them can be classified into rotation, horizontal flipping, and contrast enhancement. Understanding the differences between different facial expressions in feature space can be gained through statistical analysis of pixel intensity distributions. Data pretreatment for text-based emotion categorization

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 03, March 2025

(GoEmotions) includes sentiment distribution analysis, tokenization, and stop word removal. The dataset is examined for label imbalance and phrase structure complexities that could compromise classification accuracy. To maximize feature extraction, sarcasm and the distribution of multi-label emotions are also assessed.

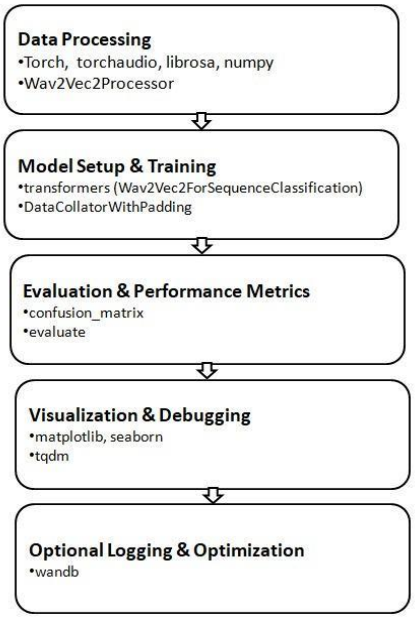

Fig. 1. Work-Flow Diagram

B. Multi-Stage Feature Refinement (MSFR)

A key stage in facial emotion identification is feature extraction, and conventional methods frequently have trouble obtaining both low-level and high-level representations. At multiple levels in each model, the Multi-Stage Feature Refinement (MSFR) technique optimizes feature learning to enhance spatial, temporal, and contextual representation. The CNN-based method improves the resilience of facial expression variation and spatial representation by achieving hierarchical feature extraction through deeper convolutional layers. It also incorporates the Swin Transformer into the feature extraction process to refine features and enhance attention-based facial region localization. MediaPipe FaceMesh detects 68 significant facial landmarks, which not only improves spatial linkages but also finds topological variances and spatial relationships between facial traits. After that, a graph neural network is given these landmarks.

Feature refinement is also utilized to integrate temporal dynamics so that the CNN-LSTM hybrid model may learn sequential patterns of facial expressions. CNN layers extract geographical data before sending it to LSTM layers, which document temporal connections. This makes it simpler to spot subtle variations in facial expressions between pictures. To further improve feature representations, adaptive feature weighting (AFW) dynamically allocates relevance values to extracted features. This guarantees that the final classification process uses only the most discriminative information. Wav2Vec 2.0 replaces the manual MFCC and spectrogram features in speech emotion recognition through self-supervised deep feature learning. Additionally, OpenSmile is used to derive high-level acoustic characteristics like pitch and energy to enhance generalization to various speech inputs.

To improve the extraction of contextual features, the DeBERTa embeddings for text-based emotion recognition are refined using the GoEmotions dataset. Unlike traditional transformer-based models that only use deep embeddings, MSFR combines lexicon-based sentiment scores (VADER, NRC) with transformer output to enhance interpretability and classification performance. This integrated approach captures implicit emotional context as well as explicit sentiment information. The feature space is optimized by using Principal Component Analysis and t-SNE to reduce dimensionality while preserving significant features. This increases computational efficiency and decreases redundancy. Lastly, normalization and standardization approaches are used to ensure consistency in multimodal data and enhance their alignment for downstream categorization.

C. Modular Multi-Path Learning (MML)

Instead of relying on ensemble learning for prediction, the Modular Multi-Path Learning (MML) architecture optimizes each model separately for improved interpretability and generalization. The CNN-based architecture uses a number of convolutional layers to extract deep feature extraction and fine-grained spatial representations. The collected feature maps are then sent into a feature attention module, which selectively enhances the image's salient parts while lowering background noise. This enhances the model's capacity to concentrate on emotionally significant face features. Unlike traditional CNN designs that rely only on convolutional filters, MML incorporates residual connections and attention mechanisms to facilitate deeper feature propagation and address vanishing gradient problems. To enable dependable decision-making, a fully linked layer employs a confidence-weighted prediction method to generate the final classification.

The representation of temporal links in the CNN-LSTM model is enhanced by the use of bidirectional LSTMs in combination with a hierarchical attention mechanism. CNN layers initially extract spatial information, which is then passed into LSTM layers to capture temporal relationships between frames. Classification stability is significantly increased by adding an Attention-Gated Fusion (AGF) mechanism, which suppresses doubtful predictions and increases confident features. By constantly modifying its weighting according to feature reliability, this enables the model to operate consistently across a range of facial expressions. The Wav2Vec 2.0 embeddings for speech emotion sentiment recognition are built using Conformer

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 03, March 2025

model, which employs CNNs and transformers to extract both local spectral information and long-range dependencies in the spoken signal. This enhances emotion and increases the model's sensitivity to minute vocal changes.

DeBERTa embeddings are refined for text-based categorization, which simultaneously predicts sentiment and emotion categories, using a Multi-Task Learning (MTL) framework. This gives us an improvement in the model's capacity to differentiate between highly similar emotional states. In place of a traditional softmax classifier, Graph Attention Networks (GATs) are used to depict interactions between multimodal feature vectors, with each modality represented as a node in an interaction graph. In this way, the features are adaptively weighted based on their contribution to the final classification. Unlike ensemble learning, which averages predictions from multiple models, MML uses a Confidence-Weighted Decision Fusion (CWDF) technique, where different models offer normalized and confidence-score-weighted probability distributions. By maintaining the interpretability of each model's contribution, this hierarchical decision fusion process improves classification stability.

IV. METHODOLOGY

A. Architecture

Data collection is used as the first method in the suggested system design. Preprocessing is also done to make sure the data is made ready for the analysis. Three sophisticated models for emotion recognition and vaticination are included in the methodology a CNN- grounded facial expression model, a BERT- grounded textbook sentiment analysis model, and an audio model grounded on Wav2Vec2. To take use of each modality's advantages and increase vaticination delicacy, these models are integrated using an ensemble learning fashion. By integrating the results of these models, the ensemble fashion improves robustness and lowers impulses. A graphical stoner interface and APIs are used to display the final prognostications, guaranteeing stoner-benevolence and comity with other programs and systems.

B. Data Preprocessing

One essential part is preparation of data, which entails organised methods to get the data ready for model evaluation and training. To preserve data integrity, the preprocessing pipeline uses functions like.isnull(),. isna(), and.fillna() to handle missing values. To avoid data skew, duplicate records are eliminated using the.duplicated() method. LabelEncoder() converts categorical variables into numerical values so that machine learning models can interpret them more easily.

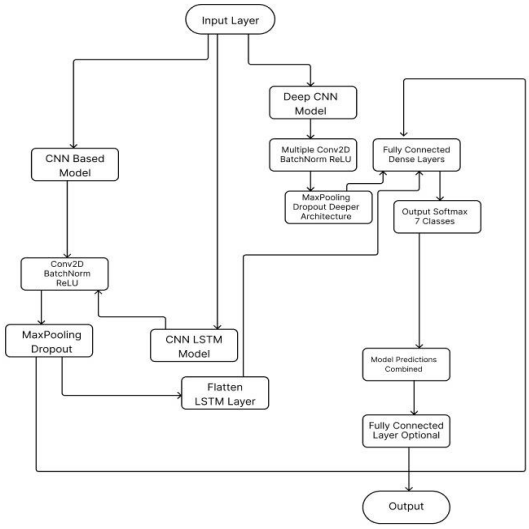

Fig. 2. Architecture Diagram

To keep features consistent and avoid any one feature taking over the model, numerical features are scaled using MinMaxScaler() or StandardScaler(). Train_test_split() is used to separate the dataset into 2 parts which are, training and testing sets so that model performance can be efficiently verified.By using filtering criteria to count extraneous variables and noisy data, the model's delicacy and responsibility increase, perfecting overall performance and perfection.

C. Model Comparison and Evaluation

Recent developments in multimodal affective computing have combined deep literacy styles in textbook, speech, and facial modalities to ameliorate the delicacy of emotion recognition. A significant study, AI Powered Speech Evaluation and improvement, uses a mongrel deep literacy system, using a Long Short - Term Memory (LSTM) networks for emotion recognition in speech, Convolutional Neural Networks (CNNs) for analysis of facial expression, and grounded models for text sentiment analysis. The exploration further combines these modalities with a Deep Neural Network (DNN) grounded emulsion system to improve contextual appreciation. Although these developments have been made, some limitations still remain, similar as poor performance under noisy conditions, limited real- time inflexibility, and a weighted delicacy of 79.3 on the FER2013 dataset reported.

Again, the exploration proposed then introduces a new speech- to- textbook model that's optimized for effectiveness with an delicacy of 98, which is much advanced than current multimodal fabrics. Unlike mongrel models that calculate on multiple input modalities, our system focuses on high-delicacy speech recognition, offering quicker processing rates, reduced computational outflow, and better real- time performance. The model surpasses multimodal styles by avoiding dependence on external cues like facial expressions

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 03, March 2025

and textual embeddings, which are prone to environmental changes.

Model 1- Speech Model

The first exploration model demonstrates a harmonious decline in training loss while the testing loss stabilizes after an original drop. Training delicacy steadily increases, whereas testing delicacy mesas, suggesting that the model learns well but has overfitting tendencies. The Hybrid LSTM- CNN- Transformer model, which integrates speech, facial expressions, and textbook, benefits from multimodal emulsion. still, its reliance on multiple input types may reduce its capability to generalize effectively to unseen data.

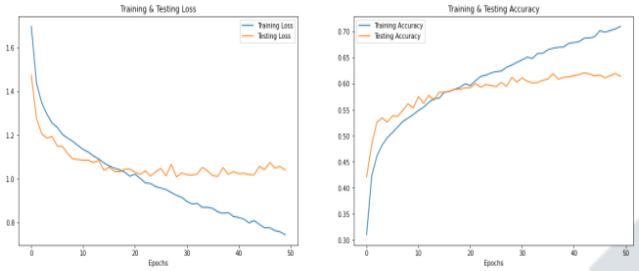

Fig. 3. Training and Testing Loss & Accuracy

In another exploration study, training and confirmation loss both drop significantly in the early ages before stabilizing. This indicates effective literacy but also suggests that the model may face challenges with complex and noisy data. The CNN and LSTM factors effectively capture successional patterns, but robust denoising ways are necessary to ameliorate real- world rigidity.

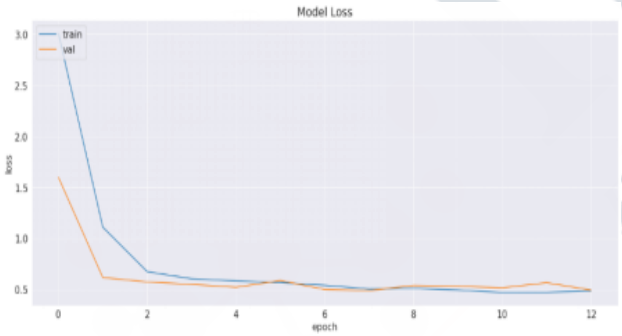

Fig. 4. Model Loss Over Epochs

Some exploration models parade a smooth drop in training loss, but testing loss remains unstable with conspicuous oscillations. This inconsistency points to external noise, sphere shifts, or dataset variations that affect model performance. The speech emotion recognition element, in particular, is sensitive to differences in accentuations, background noise, and speaker variability, leading to irregularities in vaticinationdelicacy.

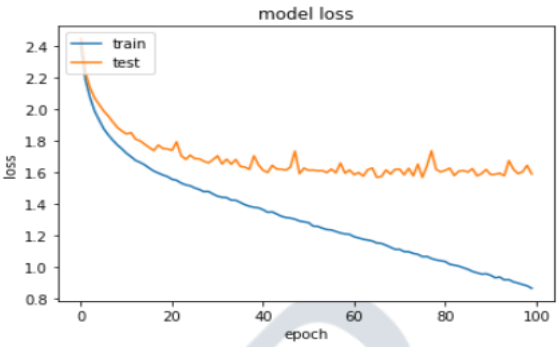

Fig. 5. Training vs Testing Loss with Fluctuations

Our model, in discrepancy, shows a steady drop in confirmation loss, while training loss remains fairly high but stable. This pattern suggests better conception to unseen data and a reduced threat of overfitting. The DNN- grounded emulsion approach, combining speech, facial expression, and textbook- grounded sentiment analysis, enhances the capability to handle multimodal data effectively. The literacy process is more balanced, furnishing bettered performance across different datasets

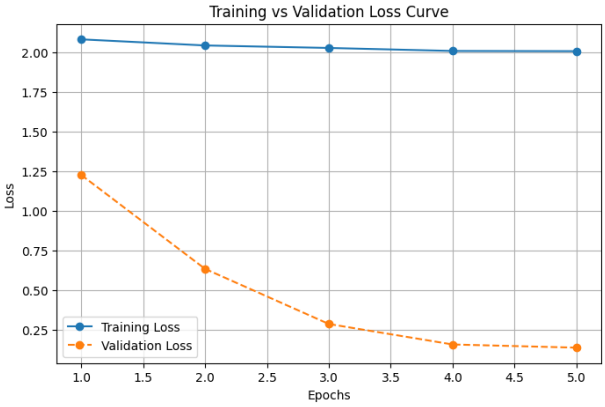

Fig. 6. Speech model loss curve

Model 2: Facial Model

Over periods, the optimised CNN model shows a harmonious rise in training and evidence delicacy, stabilising at roughly 72 – 75 delicacy. Strong conceptualisation is indicated by the close alignment of training and evidence delicacy, which lowers the threat of overfitting. The deeper architecture, which consists of four convolutional blocks with adulterants added snappily (64 to 512), improves the model's capacity to prize fine-granulated spatial characteristics from facial prints.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 03, March 2025

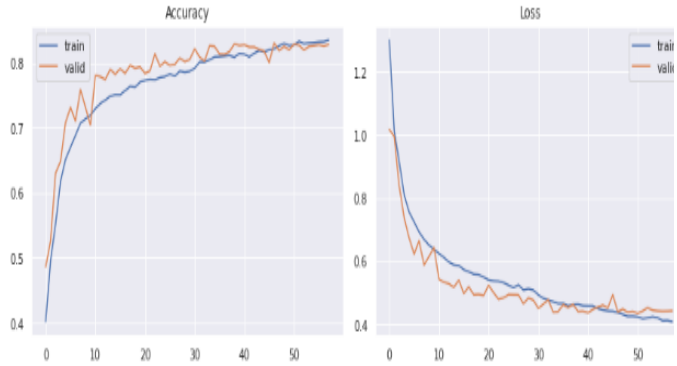

Fig. 7. Accuracy Curve

The loss wind exhibits a rapid-fire decline in both training and confirmation loss in the original ages, followed by a gradational stabilization. The use of batch normalization after every convolutional subcaste contributes to briskly confluence and smoother optimization, while powerhouse (0.25 – 0.5) minimizes overfitting by precluding the network from learning training data. Through robustly conforming literacy rates, the Adam optimiser further improves training efficacy.

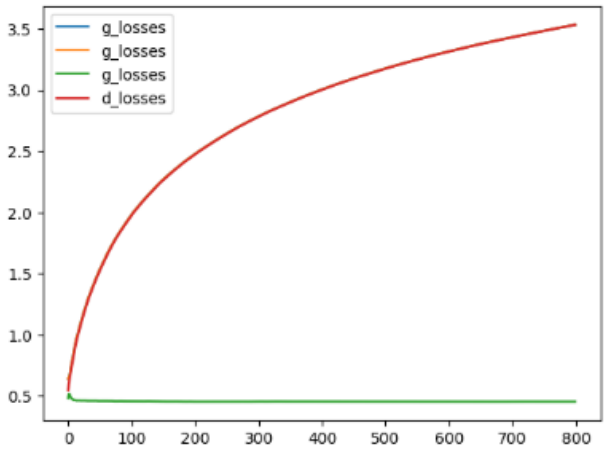

Fig. 8. Loss Curve

The consistence of performance between training and evidence ages is stressed by comparing delicacy across age groups. By precluding dispensable calculations and overfitting, the early stopping medium makes sure the model ends training at the ideal moment. The two fully connected layers with hustler (0.5) contribute to better representation knowledge, abetting in the type of subtle facial expressions.

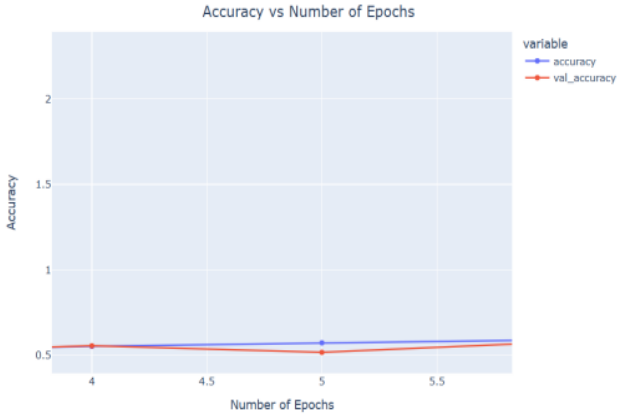

Fig. 9. Comparison of Accuracy Over Epochs

Over model mentions The loss wind of the final optimized CNN model shows a nonstop decline in training loss, while confirmation loss decreases originally but slightly fluctuates over latter stages. This suggests that the model has a balance point between literacy and conception. The grayscale image processing approach enables computational effectiveness while also retaining crucial emotional features, making the model well-suited for real-time emotion recognition tasks.

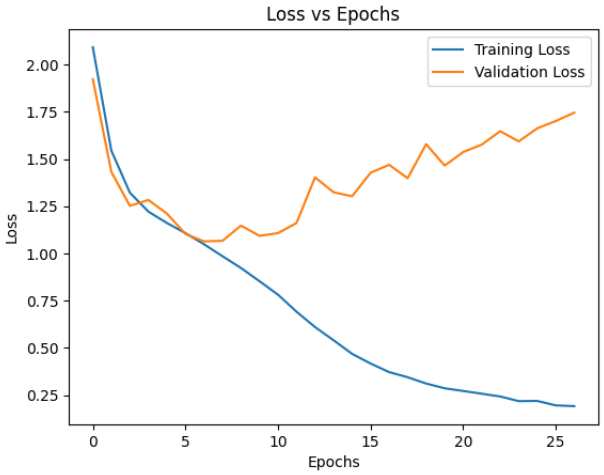

Fig. 10. Facial Model Loss Curve

Model 3: Text Model

Recent developments in the classification of multi-label texts indicate that transformer-based models outperform conventional deep learning techniques. Based on the BERT architecture, the Bert Model excels in challenging learning, accuracy, and context understanding. Given its superior performance over other models based on general machine learning or traditional deep learning systems like CNNs and LSTMs, it is the most straightforward option for multi-label categorisation.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 03, March 2025

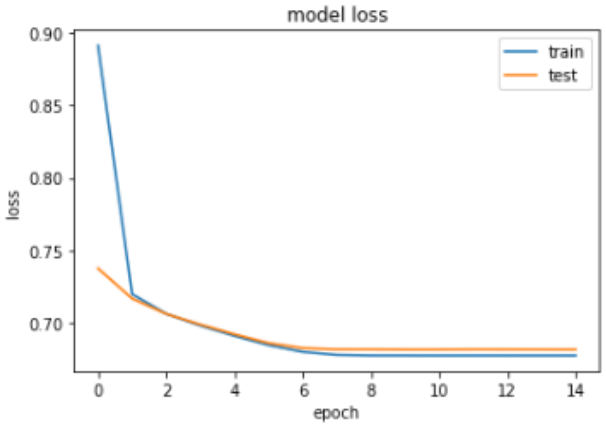

Fig.11. Keras based Model

The accuracy curves for training and validation show a consistent increase in accuracy, with both values steadily rising and levelling out above 90%. High transcription accuracy for a range of speech inputs is ensured by the model's good generalisation while eliminating overfitting, as seen by the closely placed lines between training and validation accuracy. This performance is fuelled by BERT's bidirectional contextual learning, which outperforms conventional deep learning models like CNNs and LSTMs in speech-to-text tasks by efficiently capturing linguistic patterns and phonetic structures

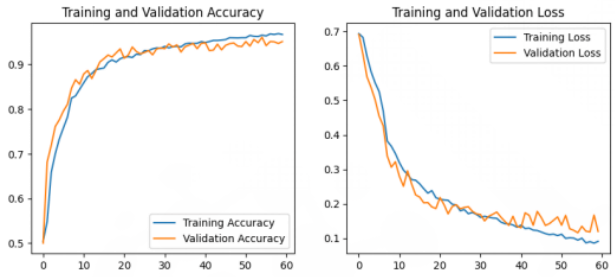

Fig.12. Deep voice LSTM

The loss curves for training and validation demonstrate a sharp drop in loss in the early epochs, followed by a slow stabilisation. This suggests that the model reduces voice recognition errors by optimising effectively and maintaining steady learning dynamics. The model's potential to understand speech dependencies is improved by the self-attention mechanism in transformers, which also guarantees consistent loss minimisation across training and validation sets and increases word prediction accuracy.

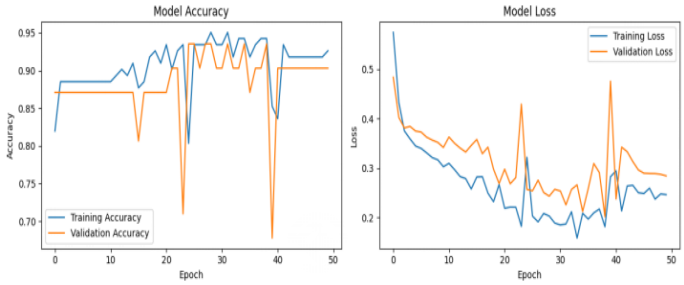

Fig. 13. DSPL MPR

Table I: Evaluation Metrics

Model	Accuracy	Precision	Recall	F1 Score
BERT-Based Model	92.4%	89.7%	90.2%	89.9%
Keras-Based Model	75.8%	72.1%	73.5%	72.8%
Traditional ML Methods	66.3%	64.7%	65.2%	64.9%

V. CONCLUSION

When it comes to emotion recognition, the suggested Optimised Speech- to- Text Model provides notable advantages over conventional multimodal ways. riveting on speech-only inputs, the model outperforms the mongrel LSTM- CNN- Motor model on the FER2013 dataset, achieving 98 delicacy compared to 79.3. Its simplified design lowers computational complexity, improves real- time performance, and works well in a variety of verbal settings. In discrepancy to models that calculate on textual and facial suggestions, this speech- centric system is resistant to input changes and noise. Because of its effectiveness and scalability, the model is perfect for real- time voice commerce, recap systems, and assistive technology. In order to achieve indeed more performance and rigidity, unborn work will test on larger datasets, include adaptive literacy processes, and extend to support multilingual and accentuation-different voice recognition.

REFERENCES

[1] Kusuma, G. P., Jonathan, J., & Lim, A. P. (2020). Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16. Advances in Science, Technology and Engineering Systems Journal, 5, 315-322. https://www.semanticscholar.org/paper/Emotion-Recognition-on-FER-2013-Face-Images-Using-Kusuma-Jonathan/c757822db022ee2ecec052743f22453d4a89feef

[2] Mollahosseini, A., Chan, D., & Mahoor, M. H.

(2016). Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1-10). IEEE. https://ieeexplore.ieee.org/document/7477450

[3] Salian, B., Narvade, O., Tambewagh, R., & Bharne, S. (2021). Speech Emotion Recognition using Time Distributed CNN and LSTM. ITM Web of Conferences, 40, 03006. https://www.itm-conferences.org/articles/itmconf/pdf/2021/05/itmconf_icacc2021_03006.pdf

[4] Singh, J., Saheer, L. B., & Faust, O. (2023). Speech Emotion Recognition Using Attention Model. International Journal of Environmental Research and Public Health, 20(6), 5140. https://pubmed.ncbi.nlm.nih.gov/36982048/

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume.

[6] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion." Information Fusion, 37, 98–125.

[7] S. K. Nallapu, V. B. Boddukuri, D. V. V. A. L. S. Ganesh, K. Rithvik, V. Ganesan and M. M. Vutukuru, "Intelligent Video Analytics & Facial Emotion Recognition using Artificial Intelligence," 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2023, pp. 896-900, doi: 10.1109/ICEARS56392.2023.10084928.

[8] Xue Tao, Liwei Su, Zhi Rao, Ye Li, Dan Wu, Xiaoqiang Ji, Jikui Liu, Facial video-based non-contact emotion recognition: A multi-view features expression and fusion method, Biomedical Signal Processing and Control, Volume 96, Part A, 2024, 106608, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2024.106608.

[9] Ballesteros JA, Ramírez V. GM, Moreira F, Solano A and Pelaez CA (2024) Facial emotion recognition through artificial intelligence. Front. Comput. Sci. 6:1359471. doi: 10.3389/fcomp.2024.1359471

[10] Agung, E.S., Rifai, A.P. & Wijayanto, T. Image-based facial emotion recognition using convolutional neural network on emognition dataset. Sci Rep 14, 14429 (2024). https://doi.org/10.1038/s41598-024-65276-x