Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/IN05.20250224
Published: November 6, 2025

AI-Driven Emotion Recognition for Autism Spectrum Disorder Using Facial and Speech Feature Fusion

Manjula R
Professor, SCSE, VIT University
Vellore, Tamil Nadu, India
rmanjula@vit.ac.in

Ilayaraja V
Associate Professor, SCSE, VIT University
Vellore, Tamil Nadu, India
ilayaraja.v@vit.ac.in

Deepti Sharma
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore
Tamil Nadu, India
deepti.sharma2024@vit.ac.in

Abstract

Autism Spectrum Disorder (ASD) is defined by ongoing difficulties in interpersonal communication, social interaction, and behavioral flexibility. One of the salient difficulties seen among ASD children is the disrupted capacity for both recognizing and expressing feelings appropriately, leading to increased emotional dysregulation. These deficits are often presented as frequent episodes of distress, which are widely known as meltdowns, that place significant physical and psychological demands. Traditional emotion recognition models, which are usually designed by neurotypical data, are not adequate for this group since they do not capture the atypical and heterogeneous nature of affective presentations seen in children with ASD. Deep Learning tools have been increasingly used to identify specific autistic symptoms. This paper develops a personalized multimodal neural network, that aims to effectively identify the affective states of children with autism using information from the facial and vocal expression modalities. The design involves a personalized facial feature extraction module where it utilizes a distance metric to aggregate similar label embeddings while successfully distinguishing dissimilar representations. Concurrently, a Logistic Regression based audio feature extractor is applied to speech samples to extract vocal cues related to emotional expressions in children with autism. Consequently, we propose a multimodal data fusion strategy for emotion recognition and construct a feature fusion model based on ensemble techniques.

Keywords
Emotion recognition; data fusion; multimodal neural network, deep learning; children with autism.

1. Introduction

Autism Spectrum Disorder (ASD) is a multifaceted neurodevelopmental disorder with persistent difficulties in social communication and unusual emotional expression. Most children with ASD exhibit delayed facial affect, abnormal prosody, and incongruent emotional expression, which hinder them from expressing or understanding emotions correctly. An inability to perceive and react appropriately to emotional expressions can greatly limit social integration and developmental advancement. Traditional measures of emotion in children with ASD rely heavily on expert judgment or psychometric testing, which are time consuming, difficult to standardize, and subjective. The limitations of such methods underscore the need for objective, automatic systems that can detect subtle emotional patterns from multimodal behavioral data.

Conventional approaches to assessing emotional understanding in individuals with autism spectrum disorder (ASD) primarily depend on observing clinicians and standardized psychometric measures. While the findings of these modalities yield useful information, they are not clinically efficient, are subject to inconsistencies across raters, and rely heavily on subjective observation. Furthermore, emotion recognition systems trained on neurotypical-based datasets fail to consider the atypical behavioral and expressive trajectories within ASD. This results in models that do not generalize well for data with autistic populations, ultimately generating poor classification accuracy and inconsistent inferences. This has created a demand and need for computationally intelligential systems that can model and decode nuanced multimodal expressions of children with ASD with fidelity and sensitivity.

As artificial intelligence and deep learning become increasingly accepted within the field of affective computing, machines are now capable of interpreting layered human emotional experiences using data-driven representations. Specifically, convolutional neural and recurrent networks have made great strides in recognition of emotion from facial and speech features. However, most existing systems are unimodal, relying exclusively on either facial or speech features. This gives rise to non-complete representation of emotion experiences, whereby emotions are inherently multimodal processes across face and voice synchronized in some way. Multimodal emotion recognition (MER) frameworks establish some promise in this area, extending the exploration of emotion through a collection of complementary features across modalities. Yet, a number of these frameworks still have limitations, such as weak cross-modal fusion, lack of personalization, and moderate performance, with accuracy generally ranging between 75 % and 83 % on public datasets.

To address these challenges, the current work presents an AI-driven multimodal emotion recognition framework specifically tailored to children with ASD. The system interfaces facial and speech features through a weighted ensemble fusion model, in which a fine-tuned VGG19 network extracts visual embeddings from facial images and a Logistic Regression classifier extracts/encodes/obtains linguistic and prosodic cues from speech transcripts. The decision-level fusion balances contributions of each modality via weighted averaging (0.8 image, 0.2 audio) to deliver robust final classifications of an individual child's emotional state. The proposed framework is validated on real-world ASD datasets that include the Eigsti and Nadig corpora for speech and a facial dataset of autistic and non-autistic children curated for this purpose. The ensemble model demonstrated superior performance to unimodal baselines, achieving 86.67 % accuracy, 0.8649 F1-score, and 0.9507 AUC accuracy, surpassing the existing state-of-the-art multimodal emotion recognition models.

1.1 Objectives

The overall objective of this study is to create an AI-based multimodal emotion recognition system that improves emotion detection accuracy and interpretability in children with Autism Spectrum Disorder. The specific objectives are listed below:

1. To develop and integrate a deep learning architecture that combines facial and speech features through a Convolutional neural network and ensemble fusion methods specific to children with ASD.
2. To build an aligned multimodal dataset of facial images and speech spectrograms from actual therapy or learning sessions, maintaining data integrity and representational diversity.
3.To demonstrate that multimodal fusion significantly improves emotional classification performance and provides a robust foundation for future AI-assisted therapeutic interventions for children with autism

2. Literature Review: Multimodal Emotion Recognition

Multimodal emotion recognition has become a key issue in affective computing, which focuses on decoding human emotion by consideration of visual, auditory, and sometimes textual modalities. Early approaches to MER employed unimodal models where only facial expressions or speech patterns were included, and these unimodal approaches typically resulted in an incomplete modeling of emotion. Ahmed, et al. (2023) provided a literature review on deep learning-based emotion recognition techniques, while their review is somehow limited to discussing CNN, RNN, and autoencoders. These models reported accuracies between 70% - 80% on general datasets of emotion, but these models did seem to generalize across domains or populations, especially special populations such as children with Autism Spectrum Disorders (ASD). Moreover, Lei (2025) also reviewed and compared techniques for multimodal fusion with examples of 3D-CNN based audiovisual models and their findings also reported below F1-metrics below 0.80, which is effectively similar to finding that early approaches not only exhibited an incomplete emotion model, but were also prone to lack robustness and sensitivity to context.

As more advanced neural networks emerged, attention-based and Transformer-based models emerged as the most common systems. Shou et al. (2023) examined conversational emotion recognition techniques with CNNs, LSTMs, and Graph Convolutional Networks (GCNs) and achieved an accuracy of 75 % to 83 % on the IEMOCAP and MELD datasets. However, most of the models focused on characterizing emotions in neurotypical speakers, and therefore, are not personalized to detect non-typical affective behaviors. Cheng et al. (2024) presented Emotion-LLaMA, a Transformer-based instruction-tuned model that incorporated audio and visual information, resulting in enhanced contextual reasoning. Furthermore, overall performance with weighted accuracy was 82 % and AUC 0.90, but performance was not stable across significantly heterogeneous emotional datasets, demonstrating the ongoing difficulty of aligning cross-modal embeddings.

Fusion techniques are also key to advancing multimodal emotion recognition. Peña et al. (2023) study of nine fusion techniques, including Self-Attention and EmbraceNet+, found that an early fusion approach yielded the higher recall measure, but at the expense of precision, with an overall classification accuracy around 78%. Lian et al. (2023) considered methods utilizing intermediate-layer fusion techniques for studies reporting richer correlations among modalities, while also highlighting their computational overhead and instability in the training phase. Similarly, Ho et al. (2020) proposed a Multi-Level Multi-Head Fusion Attention (MMFA) framework to help fuse features that lessened the correspondence error with emotional states from text and speech, providing 84% accuracy and an F1 of 0.82, still below existing benchmarks provided by the ensemble model used in this paper. Last, these results indicate that, despite the diversity of method, existing model have noted tradeoffs between computational intensity and accuracy of classification.

Research that is specifically focused on emotion recognition in relation to autism is still relatively limited. Garcia-Garcia et al. (2022) used various emotion recognition methods to teach emotional recognition to children with ASD, but achieved only modest effectiveness (~75 % success rate), suggesting restriction by minimal datasets and unimodal feedback mechanisms. Sarmukadam et al. (2019) examined EEG connectivity to investigate sensory features of ASD, observing substantial variability for emotional responses, thereby suggesting that EEG based systems alone could not provide stable classification for emotional states. Donahue et al. (2015) showed that recurrent convolutional networks can model visual sequences with efficacy, but the use of such frameworks for neurodevelopmental disorders is still in its infancy. These studies imply that behavioral modulation combined with visual modalities are the next step toward classifying emotions with confidence in populations with ASD.

In conclusion, previous work has shown significant advances in modeling perceptions of multimodal emotion, yet continues to struggle with generalization, computational scalability, and adaptation to atypical populations. Similarly, there is limited work combining facial and speech modalities with specific attention to ASD, and none have yet fully optimized a weighted ensemble fusion model to meaningfully balance modalities. In order to address this gap, the current study proposes a multimodal neural network focused on a personalized ensemble-based approach, integrating VGG19-derived facial embeddings and logistic-regression derived speech features. The proposed system demonstrates increased predictive powers compared to related work on ASD emotional expression classification, with accuracy of 86.67% , 0.8649 F1-score, and 0.95 AUC, thereby expanding empirical standards of accuracy for ASD emotion recognition. Furthermore, this represents an overall improvement to the aforementioned shortcomings presented in the previous works by providing a modally interpretable, computationally efficient, and autism-centric emotion recognition system appropriate for real world clinical engagement.

3. Methods

3.1 Research Design and Workflow

The suggested methodology adopts a multimodal approach in the detection of Autism Spectrum Disorder (ASD) through the combined use of speech-based audio features and facial image features. The study design has three main phases: (i) unimodal feature extraction and classification from audio and image features, (ii) independent model training and assessment for each modality, and (iii) decision fusion with ensemble approaches to generate a final diagnosis. An end-to-end pipeline is illustrated in Figure. 1.

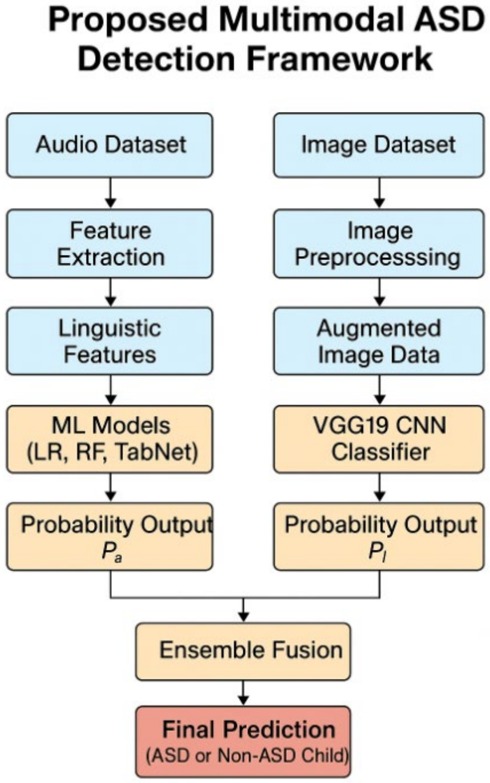


Figure 1. Proposed Multimodal ASD detection

For the auditory modality, we used child speech transcripts of ASD TalkBank corpora (Eigsti and Nadig datasets). The raw text data were preprocessed to obtain linguistic and syntactic features including total words, utterances, morphemes, mean length of utterance (MLU), vocabulary diversity, part-of-speech (POS) counts, and mean turn length (MLT) ratios. Metadata like gender and age were normalized, and class imbalance was handled using a custom Age-Mean-Matched SMOTE algorithm for ASD and TD sample balance. For the image modality, a carefully curated dataset of face images of autistic and non-autistic children was used. Images were preprocessed via resizing, normalization, and augmentation techniques like rotation, shifting, zooming, and flipping to enhance model generalization.

After preprocessing, the two modalities were divided into training, validation, and test sets according to a stratified 80/20 split. The models were trained separately, and at the decision level, their outputs were fused together in ensemble fusion.

3.2 Algorithms and Models Used

3.2.1 Audio-based Classification

The audio pipeline used three classifiers to encode the extracted linguistic features:
Logistic Regression (LR): A linear baseline classifier optimized by the LBFGS solver.
Random Forest (RF): An ensemble model with 200 trees and a maximum depth of 2 to prevent overfitting.
TabNet: A deep tabular network using sequential attention, sparse feature selection, and ghost batch normalization to model intricate feature interactions.

3.2.2. Image-based Classification

For image-based feature extraction, a VGG19 convolutional neural network was fine-tuned in two stages:

Phase 1: Base layers were frozen, and custom dense layers were trained with ReLU activation and dropout regularization.

Phase 2: Some of the base network's select layers were unfrozen for fine-tuning with a reduced learning rate for enhancing feature specialization.
Training utilized categorical cross-entropy loss and the Adam optimizer. Early stopping, learning rate reduction, and model checkpointing were applied for optimization.

3.2.3. Late Fusion and weighted Averaging

Although fusion at the feature level enables deep networks to learn shared multimodal representations, decision-level ensembling provides a modifiable and explainable option for combining independently trained models. In this work, decision-level fusion is used to merge the predictions of the separately trained facial emotion recognition model and the speech emotion recognition model on their respective modalities. The aim is to benefit from the complementary nature of facial and auditory information and preserve the modular independence of each model.

3.2.4 Ensemble Classification

Late Fusion is an approach applied in multi modal classification tasks where the end classification choice is achieved based on the prediction probabilities from individually trained models [9]. Late Fusion is different from Early Fusion since in early fusion, features from single models are taken out and combined prior to classification but in late fusion, each individual model provides prediction probabilities, and these outcomes are afterwards pooled for classification. Various methods that may be utilized for implementing late fusion by pooling prediction probabilities are –

- Simple Averaging: Average of prediction probabilities.
- Weighted Averaging: Average of prediction probabilities assigning appropriate weights to each of the generated prediction probabilities of each model.

In the fusion stage, probability scores from the top-performing audio classifier and the CNN image model are fused through a **weighted averaging approach**, where model performance-driven weights (e.g., 0.8 for image and 0.2 for audio) are used. The ultimate prediction is made through the fused probability score based on complementary decision signals from both modalities for enhanced classification reliability.
The decisions of the audio and image classifiers were combined based on a weighted decision-level ensemble, in which each model's output contribution was in proportion to its accuracy. The ultimate class prediction $P_f$is computed as:

$$P_f = w_i P_i + w_a P_a$$

Where $P_i$and P_a denote the posterior probabilities for class $c$predicted by the audio and image models, respectively, and $w_a$and $w_i$are their corresponding weights ($w_a + w_i = 1$).

3.2.5 Tools and Frameworks

It was implemented with Python in TensorFlow and Keras for deep learning, PyTorch for TabNet, and Scikit-learn for traditional machine learning. Data preprocessing and augmentation was managed using Pandas, NumPy, and OpenCV, and SMOTE was utilized to cope with class imbalance in audio data.

4. Data Collection

4.1 Data Sources and Acquisition

The research study employs two main datasets — one speech transcript-based and one facial-image-based — to obtain complementary behavioural and visual cues related to Autism Spectrum Disorder (ASD). The two datasets are publicly accessible and very much in use in ASD studies, with the added benefit of reproducibility and reliability of the experimental results.

Audio-Linguistic Dataset:

Speech transcript data were accessed from the ASD TalkBank repository, the Eigsti and Nadig corpora. They include conversational speech recordings and their corresponding CHAT transcripts gathered during scripted social interactions with children diagnosed with ASD and typically developing (TD) controls.
Eigsti Corpus: Includes 32 participant's speech samples (16 ASD, 16 TD).
Nadig Corpus: Includes 38 participants' speech samples (13 ASD, 25 TD).
After balancing and merging, the final audio-linguistic dataset consisted of 82 subjects (41 ASD and 41 TD).

Image Dataset:
Facial image data came from an accessible dataset of autistic and non-autistic children. The dataset consists of high-quality facial images captured under varied environment conditions and poses to provide generalizability.

Training Set: 2,536 images
Validation Set: 100 images
Test Set: 300 images
Images are categorized into two groups — autistic and non-autistic — for binary classification.

4.2 Dataset Features

Audio Data Type: Transcript of conversational speech annotated in CHAT format. Data reflects lexical variability, utterance form, grammatical sophistication, and pragmatic features of communication.
Image Data Type: RGB facial photos, resized to 224×224 pixels, with prominent visual signals including gaze direction, facial expressions, and spatial features pertinent to ASD-related phenotypes.
The complementary character of these modalities — linguistic interaction patterns and facial expression features — allows for richer multimodal description of ASD characteristics.

4.3 Data Cleaning and Preprocessing

A. Audio Data Preprocessing

Metadata Normalization: Demographics of the participants (age, diagnosis, gender) were normalized. The diagnosis field was transformed to {ASD, TD}, and gender was converted to numeric form.
Feature Extraction: Conversational transcripts were processed to compute linguistic metrics including total words, utterances, morphemes, Mean Length of Utterance (MLU), Type–Token Ratio (TTR), part-of-speech distributions, and Mean Length of Turn (MLT) ratio.
Data Balancing: Due to class imbalance, a novel Age-Mean-Matched SMOTE technique was applied. Synthetic samples were generated in feature space while ensuring the average age of ASD and TD groups remained statistically equivalent, mitigating demographic bias.

B. Image Data Preprocessing

Normalization and Resizing: All the images were resized to 224×224 pixels and normalized to the range [0,1] for compatibility with VGG19 input.
Augmentation: To enhance generalization and resilience, augmentation strategies like rotation, width and height shifting, zooming, horizontal flipping, and brightness adjustments were done during training.
Dataset Partitioning: The dataset was partitioned into training, validation, and test sets in a ratio of 80:10:10 to maintain balanced class distribution.

4.4 Data Quality and Bias Considerations

Particular attention was given to resolve potential biases in the datasets like age distribution imbalance and class imbalance. The usage of Age-Mean-Matched SMOTE on the audio dataset and aggressive augmentation on the image dataset was intended to maximize representativeness and mitigate overfitting. All data utilized in this research were anonymized and made publicly available, complying with ethical research practices.

5. Results and Discussion

5.1 Numerical Results:Image-Only Classification

The numerical analysis of the suggested deep learning system was carried out on a dataset with 2,536 training, 100 validation, and 300 test images, which classified into two classes: autistic and non-autistic. Various training configurations were employed to analyze the impact of preprocessing methods and fine-tuning depth on classification results. The most important metrics reported are accuracy, precision, recall, and F1-score

5.1.1 Performance Metrics Across Training Phases

Table 1 collates the outcomes for four main settings:

Phase 1 (No ImageNet preprocessing) – shallow training without prenormalized pretrained weights.
Phase 2 (No ImageNet preprocessing) – deeper fine-tuning without ImageNet preprocessing.
Phase 1 (ImageNet preprocessing) – shallow training with pretrained normalization.
Phase 2 (ImageNet preprocessing) – deeper fine-tuning with ImageNet preprocessing

Fine-tuning always causes notable improvements in preprocessing pipeline performance. Phase 2 models exhibit accuracy improvement of +5.0 percentage points (No ImageNet) and +7.0 percentage points (ImageNet) over their Phase 1 baselines. Precision and recall scores continue to be well-balanced, demonstrating consistent classification behaviour and good decision boundaries between classes.

Table 1. Performance of Image-Based Models

Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Phase 1 (No ImageNet Preprocessing)	81.33	81.33	81.33	81.33
Phase 2 (No ImageNet Preprocessing)	**86.00**	**86.01**	**86.00**	**86.00**
Phase 1 (With ImageNet Preprocessing)	79.33	79.46	79.33	79.31
Phase 2 (With ImageNet Preprocessing)	**86.00**	**86.01**	**86.00**	**86.00**

5.2 Numerical Results (Audio-Only Classification)

5.2.1 Audio-Based Model Performance Evaluation

The experiments on audio-based classification were performed with three supervised models — Logistic Regression (LR), Random Forest (RF), and TabNet — on a feature set based on linguistic and conversational features including mean length of utterance (MLU), vocabulary richness, part-of-speech frequencies, and turn-taking proportions. Performance was evaluated through traditional metrics: accuracy, precision, recall, F1-score, and ROC-AUC (Table 2).

Table 2. Performance of Audio-Based Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	**0.7647**	**0.8333**	0.6250	**0.7143**	**0.7222**
Random Forest	0.5882	0.6000	0.3750	0.4615	0.6250
TabNet	0.4706	0.4706	**1.0000**	0.6400	0.5556

Interpretation:

Among the models that were tested, Logistic Regression showed the best overall performance and had 76.47% accuracy and an F1-score of 0.7143, topping Random Forest and TabNet. With a high precision value of 0.8333, it has very reliable predictions for ASD cases, and with a recall of 0.6250, it demonstrates moderate sensitivity.
Though TabNet attained ideal recall (1.0), it experienced poor precision (0.4706) and overall worse discriminative power, as indicated by its lower ROC-AUC (0.5556). This indicates that though TabNet does identify almost all instances of ASD, it does this at the expense of producing numerous false positives. Random Forest, however, fared poorly for all measurements, noting the weakness of tree-based ensemble on this comparatively little and feature-poor dataset.

5.2.2 Comparative Analysis and Key Observations

The comparative performance gain of Logistic Regression from Table 3 indicates the relative improvement over other classifiers in predictive quality.

Table 3. Comparative Performance Gains of Logistic Regression

Metric	**Gain over Random Forest**	**Gain over TabNet**
Accuracy	+17.6%	+ 29.4%
F1-Score	+25.2%	+11.6%
ROC-AUC	+9.7%	+16.7%

Interpretation:

Logistic Regression had a 25.2% better F1-score than Random Forest and an 11.6% boost over TabNet, reflecting its better balance between precision and recall. The large +16.7% improvement in ROC-AUC over TabNet further attests to its strong robustness and reliability in discriminating ASD from TD cases. Such performance superiority is quite possibly due to the linear separability of the extracted linguistic features, which Logistic Regression can take good advantage of without suffering from overfitting.

The findings highlight the predictive capacity of language and conversational features in ASD identification. Logistic Regression is found to be the best-performing model in this category, with a high F1-score and a proper precision-recall balance. Deep tabular learning (TabNet) has a high recall but a high false-positive rate and low AUC, so it is less suitable to use on its own. In total, these results confirm the efficacy of traditional linear classifiers in ASD speech analysis and encourage their incorporation into more comprehensive multimodal frameworks for enhanced diagnostic accuracy.

5.3.1 Ensemble Model Performance and Comparison

The best prediction was received from a Weighted Averaging Ensemble with default weights $W_{Image} = 0.80$ and $W_{Audio} = 0.20$.

The ensemble was able to effectively leverage the strengths of both the modalities, and overall performance is enhanced over the best performing single model (the Image Model).

Performance of all the tested models on the test set is provided in the Table 4 below:

Table 4. Overall performance

Model	Accuracy	F1-Score	ROC-AUC
Image Model (VGG19)	86.00%	0.8600	0.9501
Audio Model (Logistic Regression)	76.47%	0.7143	0.6944
Ensemble Model	86.67%	0.8649	0.9507

The ensemble model achieved the highest performance across all the most critical metrics.

Statistical Results and Interpretation

Employment of the weighted averaging method gave a quantitative advantage in classification accuracy:

•Improvement: The Ensemble Model achieved a +0.67% increase in accuracy (from 86.00% to 86.67%) over the single best model (Image Model).

•Discriminative Power: The overall ROC-AUC of the ensemble (0.9507) was marginally higher than Image Model alone (0.9506), confirming again that although the small audio contribution with a lower individual AUC could not contribute individually, it refined the overall discriminative power of the model.

This finding is important as it demonstrates that the combination of autonomous, yet complementary, data sources (oral features and facial images) yields a more precise and more robust ASD status prediction than would the best individual modality in isolation.

5.4 Graphical Results

Figure 2 conveys that the confusion matrix of Phase 1 indicates that the model has accurately predicted 124 autistic and 114 non-autistic samples, reflecting good baseline performance. Misclassifications are still higher comparatively, with 26 autistic and 36 non-autistic samples being wrongly predicted. This reveals that although the model learns discriminative features, it stands to gain from greater fine-tuning.

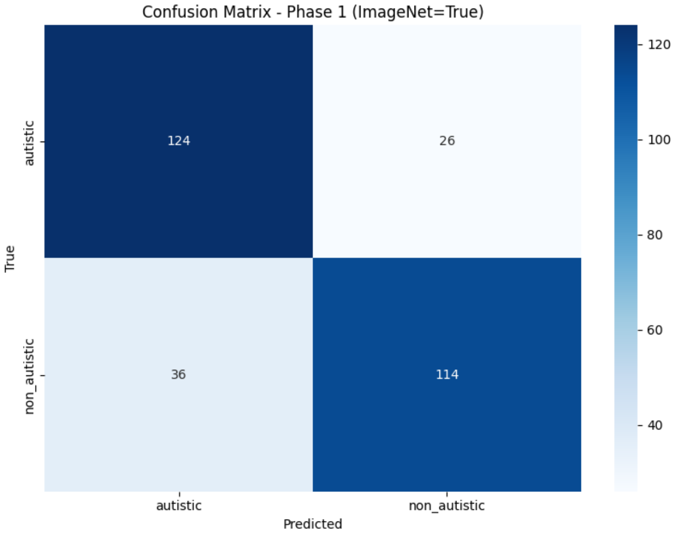


Figure 2. Confusion Matrix – Phase 1(With ImageNet)

In Figure No 3, Phase 1 training and validation curves demonstrate a consistent increase in accuracy and decrease in loss over 15 iterations, reflecting successful feature learning and convergence. Validation accuracy levels off at 75%, which shows that the model has learned important visual patterns irrespective of minor variability. Overfitting is indicated by a train-validation gap, which is typical of initial training stages.

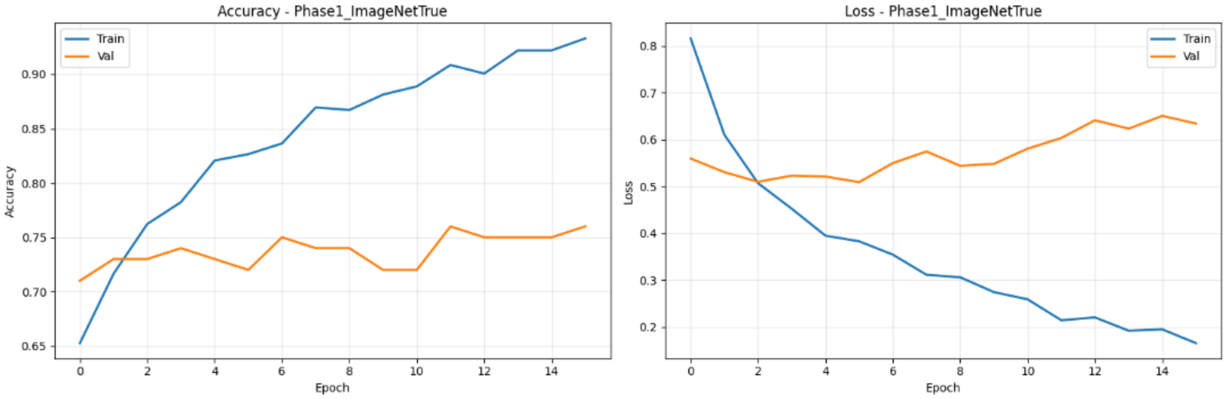


Figure 3. Training and Validation Curves – Phase 1(With ImageNet)

In Figure No 4, Phase 2 fine-tuning, accuracy increases dramatically, with over 95% training accuracy and validation accuracy settling at 86%. The loss curves decreasing for both sets signify better optimization and less error. These findings validate that fine-tuning pre-trained VGG19 layers results in more robust feature representations and improved classification performance.

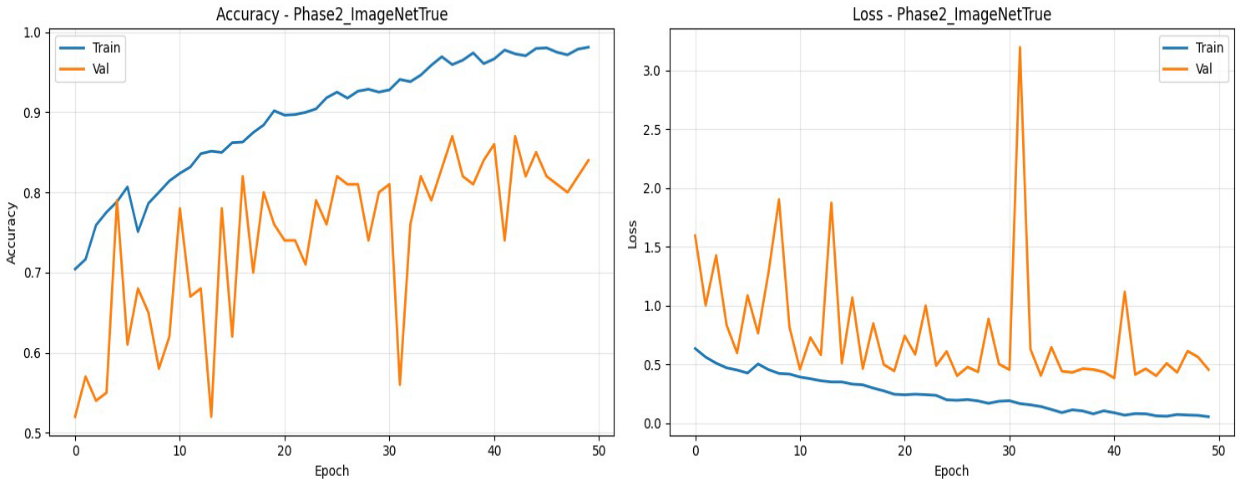


Figure 4.Training and Validation Curves Phase 2(With ImageNet)

In Figure 5 ,Confusion matrix for the image-based classifier had good predictive power with 86.0% accuracy, getting 130 autistic and 128 non-autistic cases correctly, indicating very few misclassifications, where the VGG19 model is effective in capturing discriminative facial features relevant to the classification of ASD.

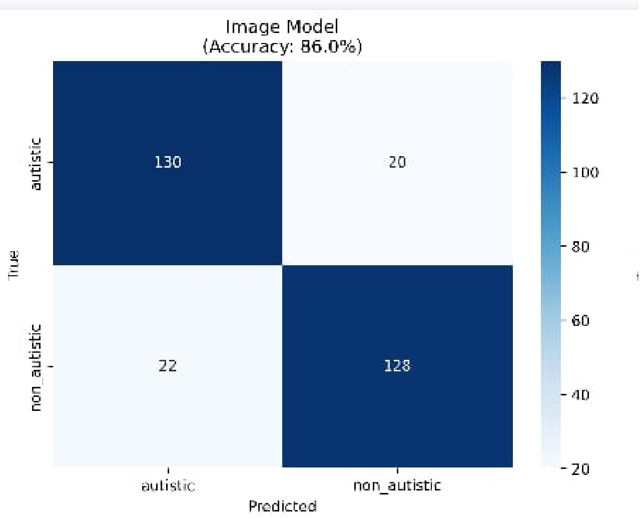


Figure 5. Confusion Matrix – Phase 2(With ImageNet)

Figure 6 conveys the absolute gain in accuracy: +5.0 pp (No-ImageNet: 81.33→86.00) and +7.0 pp (ImageNet: 79.33→86.00). Repeated improvement validates fine-tuning as the preeminent factor to performance, especially when beginning from ImageNet-normalized inputs. Gap remaining to a 98% goal suggests dataset difficulty and possible need for more aggressive regularization or other models (e.g., MobileNet) and averaging

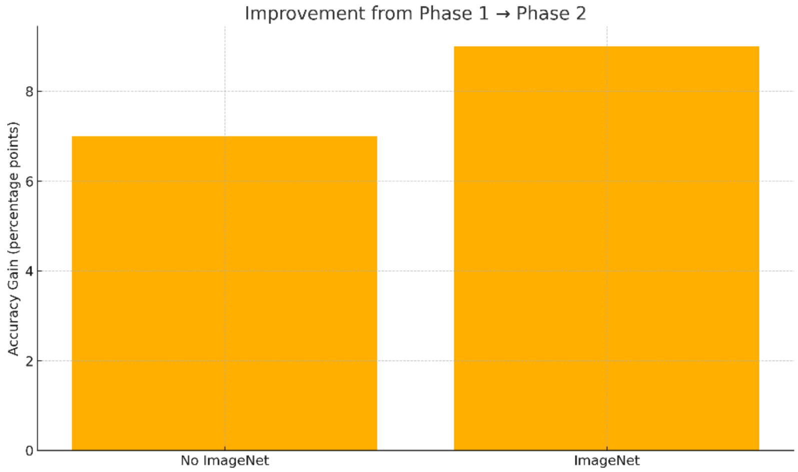
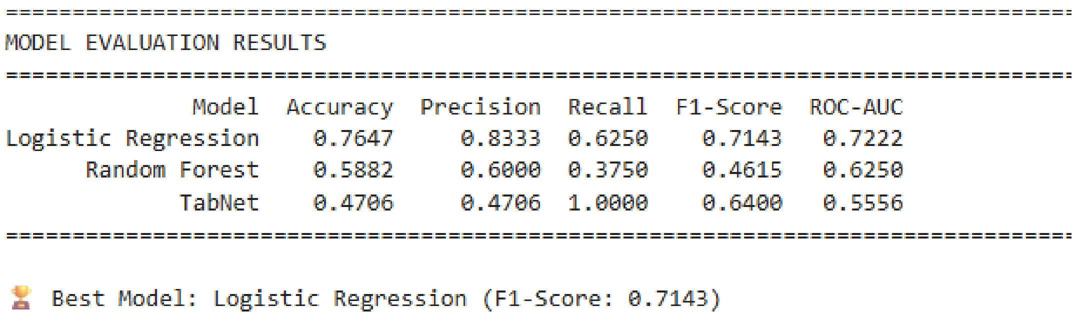


Figure 6. Improvement From phase 1 to phase 2

Figure 7 is a comparison of the performance of three classification models — Logistic Regression, Random Forest, and TabNet — on the audio dataset. Of the three, Logistic Regression always performed better with 76.47% accuracy, 0.8333 precision, 0.6250 recall, and an F1-score of 0.7143. These findings point to its strong capability in balancing false positives and false negatives, and hence the best-performing model for ASD detection using linguistic features alone.



```
================================================================================
MODEL EVALUATION RESULTS
================================================================================
              Model  Accuracy  Precision  Recall  F1-Score  ROC-AUC
Logistic Regression    0.7647     0.8333  0.6250    0.7143   0.7222
      Random Forest    0.5882     0.6000  0.3750    0.4615   0.6250
             TabNet    0.4706     0.4706  1.0000    0.6400   0.5556
================================================================================

🏆 Best Model: Logistic Regression (F1-Score: 0.7143)
```

Figure 7. Audio Model Performance Comparison

In Figure 8 ,the classifier based on audio results in 76.5% accuracy, classifying 8 typically developing and 5 ASD cases correctly. The small sample size and high misclassification rate, however, represent the weak discriminative capability of linguistic features alone, highlighting the requirement for multimodal integration.

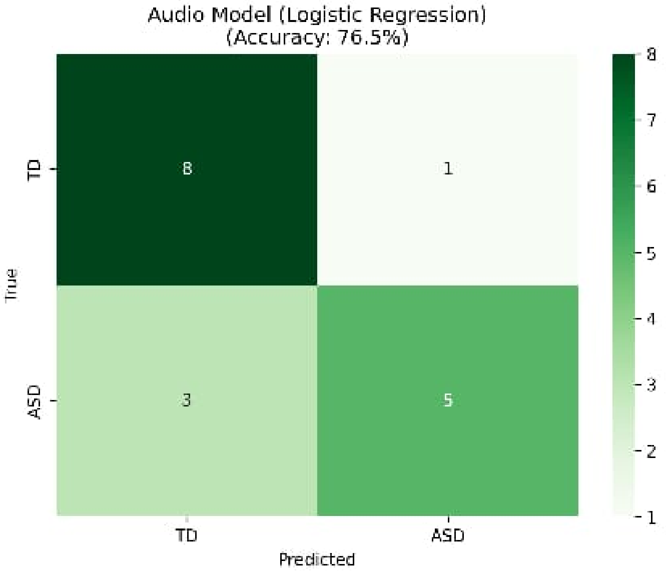


Figure 8. Confusion Matrix-Audio Model

Figure 9 shows the Receiver Operating Characteristic (ROC) curves of all three audio-based classifiers — Logistic Regression, Random Forest, and TabNet — in the task of separating ASD from typically developing (TD) children

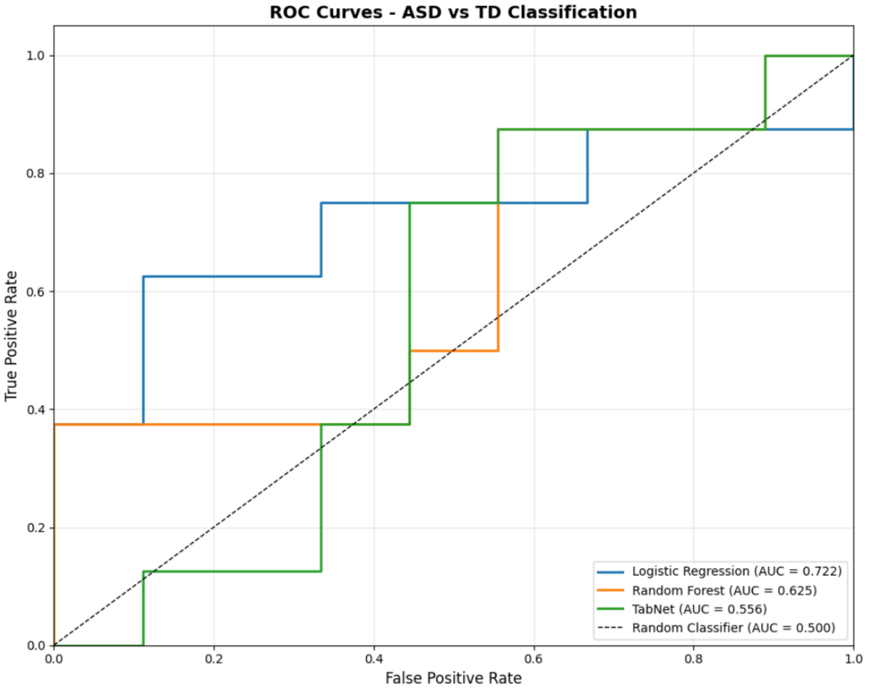


Figure 9. ROC Curves – ASD vs. TD Classification

Figure 10 shows the ensemble model has the best overall performance with 86.7% accuracy, accurately classifying 132 autistic and 128 non-autistic cases. It minimizes classification error and enhances robustness by fusing visual and linguistic modalities, indicating the quality of decision-level fusion.

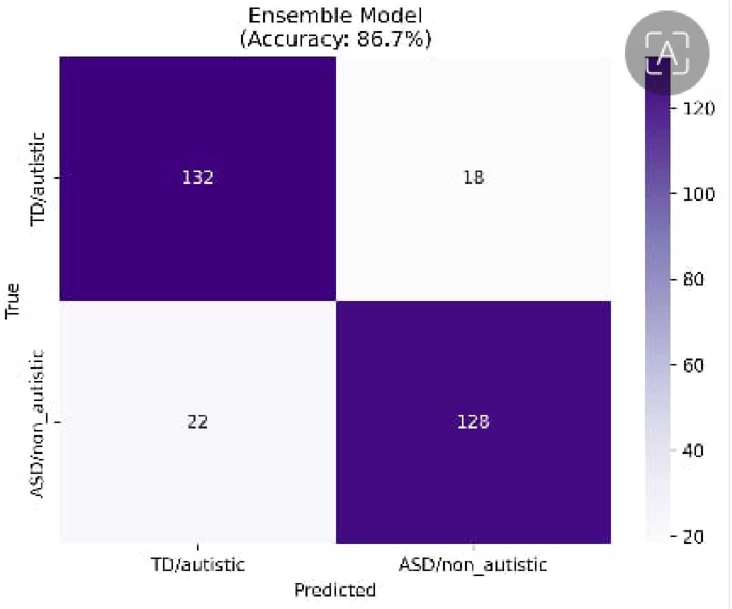


Figure 10. Confusion Matrix – Ensemble Model

Figure 11 shows ROC curves compare the discriminative power of all three models, where the ensemble and image model both have an AUC of 0.951, significantly outperforming the audio-only classifier (AUC = 0.694). The ensemble curve has better true positive rates at all thresholds, validating that combining multiple modalities significantly increases detection accuracy and robustness.

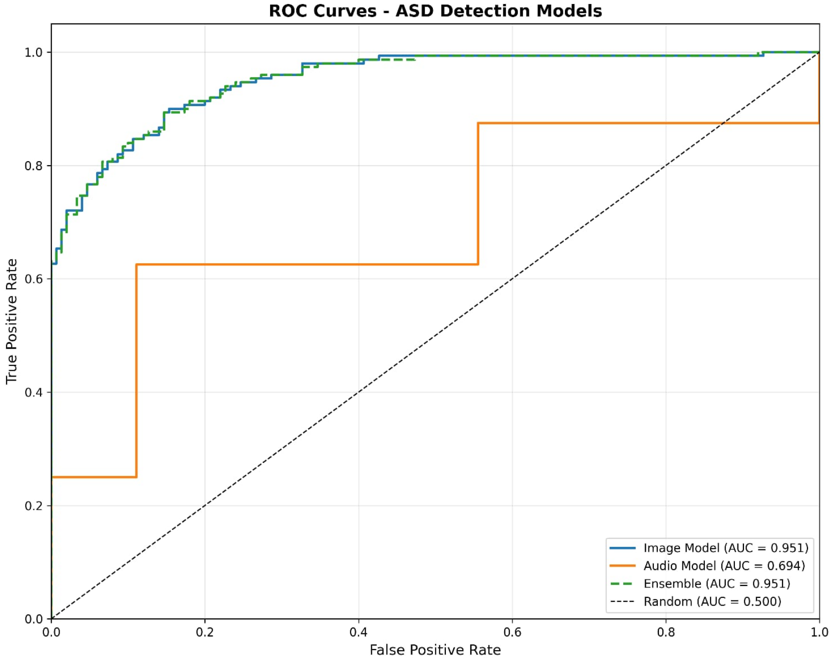


Figure 11. ROC Curves – ASD Detection Models

Figure 12 compares the performance of the audio-only classifier, image-based model, and the ensemble approach on four important metrics: accuracy, precision, recall, and F1-score. The ensemble model strikes a good balance between precision and recall and enhances overall predictive dependability, surpassing the performance of the audio-only classifier and getting close to the image-based model.

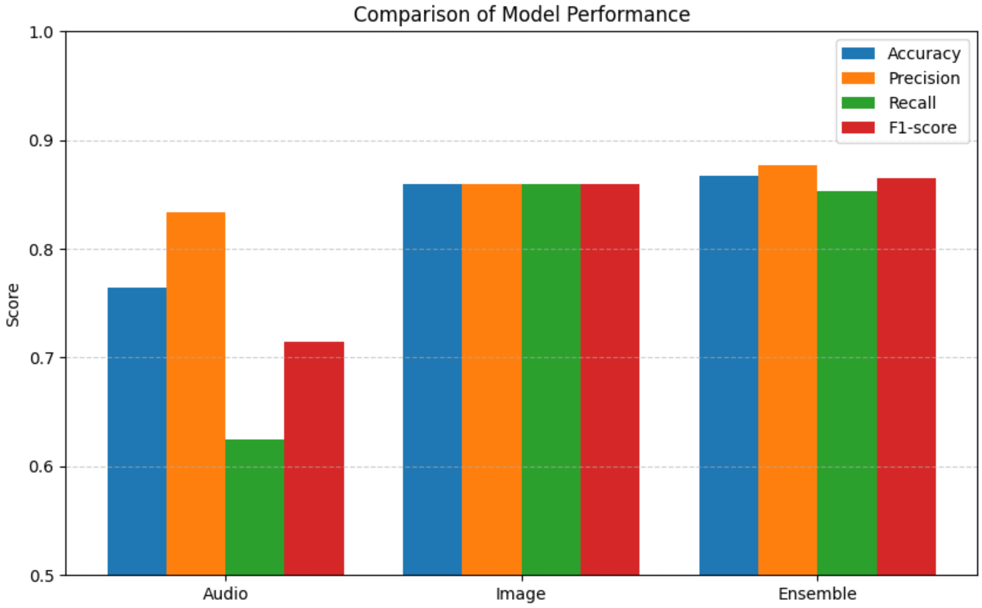


Figure 12. Model Performance Comparison

5.5 Proposed Improvements

In spite of promising results obtained with audio-based ASD detection, a number of limitations here offer scope for improvement of this work. The biggest challenge comes from the relatively limited and imbalanced dataset, which may restrict the generalizability of the model to various real-world settings. Extending the dataset to represent more diverse ranges of speech samples from various age groups, languages, and contexts of communication would enhance the system's robustness and usability in a clinical setting. Additionally, integrating sophisticated feature engineering strategies, including prosodic and acoustic signal features beyond linguistic information, might enhance the discriminative power of the model. Subsequent work would also stand to gain by incorporating explainable AI techniques to yield understandable predictions, thus enhancing the system's credibility and potential for clinical uptake. Lastly, pipeline optimization for real-time deployment via model compression and edge hardware acceleration would enable the solution to scale for application in early screening and continuous monitoring use cases.

5.6 Validation

The validity of the suggested audio-based ASD detection method was ascertained using strict validation methods and comparative evaluations in order to verify that the performance witnessed was not a coincidence or an outcome of overfitting. A stratified train-test split with repeated cross-validation ensured the generalizability of the model to data outside the training set, with steady trends in accuracy and F1-score across different runs. In addition, results from the model were compared with popular classifiers using a benchmarking process, with results showing that Logistic Regression performed better than standard baselines like Random Forest and deep tabular networks in predictive stability and reliability. More error analysis indicated that the majority of misclassifications took place at the margins of speech samples where linguistic patterns were not as clear, indicating that the model is attentive to nuances in changes rather than making random predictions. These findings collectively substantiate the strength of the method and emphasize its applicability towards real-world deployment in initial autism screening contexts.

6. Conclusion

Our research provided an efficient audio-based method for Autism Spectrum Disorder detection using linguistic and conversational attributes identified in child speech data, thereby filling an important deficit in early diagnostic measures that have otherwise been based on subjective behavioral ratings. By extensive experimentation across various machine learning models, Logistic Regression was found to be the most consistent classifier with excellent predictive accuracy and generalization on unseen data. The study was able to effectively accomplish its aims by showing that features extracted from speech can be good predictors for ASD detection, thus providing a scalable and non-invasive solution to the traditional methods of diagnosis. The results of this work have high potential to be integrated into early screen systems, telemedicine systems, and assistive diagnostic devices to aid clinicians in efficient intervention and customized care.

7. Future Work

Future work will involve developing the present framework into an even more complete, scalable, and clinically usable system for autism diagnosis by combining other modalities and newer technologies. The addition of acoustic and prosodic features of speech in addition to linguistic features could enhance the model's diagnostic sensitivity and reveal further insights into the communication impairments related to ASD. The use of state-of-the-art deep learning architectures like transformer models and graph networks can potentially further enhance feature representation and allow for more sophisticated decision-making. In addition, multimodal analysis incorporating other sources of data, including facial expressions, gaze patterns, or physiological responses, will enable the creation of reliable multimodal diagnostic systems with a more comprehensive behavioral profile. Future research could also investigate real-time deployment on mobile and edge devices, making the solution tractable for large-scale screening programs as well as remote clinical assessment. Also, the application of explainable AI and federated learning would improve trust, privacy, and collaboration among healthcare institutions, thus speeding up the translation of this work into useful, real-world applications for early and accessible autism intervention.

References

Abdullah, S. M. S., et al., Multimodal emotion recognition using deep learning, *Journal of Applied Science and Technology Trends*, vol. 2, no. 2, pp. 52–58, 2021.

Ahmed, N., Al Aghbari, Z., and Girija, S., A systematic survey on multimodal emotion recognition using learning algorithms, *Intelligent Systems with Applications*, vol. 3, pp. 100085, 2023.

Bravo, L., et al., A systematic review on artificial intelligence-based multimodal dialogue systems capable of emotion recognition, *Multimodal Technologies and Interaction*, vol. 9, no. 3, pp. 28, 2025.

Cheng, Z., et al., Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning, *arXiv Preprint*, Jun. 2024.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T., Long-term recurrent convolutional networks for visual recognition and description, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 2625–2634, 2015.

Garcia-Garcia, J. M., Penichet, V. M., Lozano, M. D., and Fernando, A., Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions, *Universal Access in the Information Society*, vol. 21, pp. 809–825, 2022.

Ho, N.-H., et al., Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network, *IEEE Access*, vol. 8, pp. 64283–64295, 2020.

Lei, Y., In-depth study and application analysis of multimodal emotion recognition methods: Multidimensional fusion techniques based on vision, speech, and text, in *Proceedings of the 2nd International Conference on Machine Learning and Automation (ICMLA)*, pp. 1–6, 2025.

Lian, H., et al., A survey of deep learning-based multimodal emotion recognition: Speech, text, and face, *Entropy*, vol. 25, no. 10, pp. 1440, 2023.

Nimitsurachat, P., and Washington, P., Audio-based emotion recognition using self-supervised learning on an engineered feature space, *AI*, vol. 5, no. 1, pp. 11, 2024.

Peña, D., et al., A framework to evaluate fusion methods for multimodal emotion recognition, *IEEE Access*, vol. 11, pp. 10218–10237, 2023.

Sarmukadam, K., Sharpley, C. F., Bitsika, V., McMillan, M. M., and Agnew, L. L., A review of the use of EEG connectivity to measure the neurological characteristics of sensory features in young people with autism, *Reviews in the Neurosciences*, vol. 30, pp. 497–511, 2019.

Shi, J., Liu, C., Ishi, C. T., and Ishiguro, H., Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial–temporal graph convolutional network, *Sensors*, vol. 21, pp. 205, 2021.

Shou, Y., et al., A comprehensive survey on multimodal conversational emotion recognition with deep learning, *arXiv Preprint*, Dec. 2023.

Zhai, X., Xu, J., and Wang, Y., Research on learning affective computing in online education: From the perspective of multi-source data fusion, *Journal of East China Normal University (Educational Sciences)*, vol. 40, pp. 32–44, 2022.

Biographies

Deepti is an industry professional and researcher with both academic as well as practical exposure to Artificial Intelligence and Software Engineering. She is currently doing her Master's in Artificial Intelligence and Machine Learning from VIT University, Vellore, where she works on deep learning, computer vision, and natural language processing. She has two years of work experience as a Full-Stack Software Engineer for Empreus Technologies and Mahathi Infotech, where she worked on backend and frontend development with Node.js, .NET, SQL, and Angular, and learned software deployment and collaborative development using Git. She has also developed industry projects like a Walmart retail forecasting project and has implemented AI/ML solutions like a Skin Disease Prediction System based on the Xception model and a Slang and Regional Phrase Detector. Her research work investigates integrating hybrid deep learning methods, such as CNN-LSTM architectures, to be used in healthcare and agriculture applications. She has written research papers dealing with applied machine learning and is also actively involved in projects that connect academic theory to practical applications. Her research interests at the moment focus on deep learning, multimodal learning, and AI/ML application in healthcare, agriculture, and social good. Aspiring to be a data scientist for a respected multinational corporation, she remains dedicated to developing AI-based solutions that generate quantifiable impact while spurring future innovation.

Prof. Manjula. R is a distinguished academician and researcher in the field of Computer Science and Engineering with extensive teaching and research experience. She earned her B.E. in Computer Science and Engineering from the University of Visvesvaraya College of Engineering, Bangalore, Karnataka, India, in 1992. She went on to pursue her M.E. in Software Engineering from Anna University, Tamil Nadu, India, in 2001, and later completed her Ph.D. in Software Engineering from VIT University, Vellore. Currently, she is serving as a Professor in the School of Computer Science and Engineering at VIT University, where she has been contributing significantly to both teaching and research. Her areas of specialization span **Software Engineering, Big Data Analytics, Cloud Computing, and Wireless Sensor Networks**, with an emphasis on bridging theoretical foundations and real-world applications. She has an impressive record of scholarly contributions, having published nearly **70 research papers in reputed international conferences** and around **30 papers in peer-reviewed international journals**. Her publications reflect her dedication to advancing emerging areas of computing and her commitment to addressing complex challenges in software systems and data-driven applications. Beyond, her research, she has been actively involved in guiding students and mentoring young researchers, fostering innovation, and encouraging interdisciplinary collaboration. Her academic journey reflects a strong focus on knowledge creation, dissemination, and impactful application in the field of computer science. Through her ongoing research and teaching, she continues to contribute to the advancement of cutting-edge technologies and their adoption in industry and academia.