International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume 9, Issue 1 (April '2026)

CROSS-MODAL TRANSFORMER AND GRAPH NEURAL NETWORK FRAMEWORK FOR ROBUST CLASSIFICATION AND ENHANCEMENT OF PATHOLOGICAL SPEECH

R. Keerthigadevi,
PG Scholar,
Department of Computer Science Engineering,
Paavai Engineering College,
Namakkal,Tamilnadu,India.
keerthigadevi@gmail.com

K. Santhanalakshmi,
Professor,
Department of Computer Science and Engineering,
Paavai Engineering College,
Namakkal, Tamilnadu,India.
santhana11lakshmi@gmail.com

Abstract: *Communication may greatly be affected by speech and language disorders and therefore, early and proper diagnosis of the condition is the key to effective clinical treatment. The current paper introduces a single artificial intelligence system used in pathological speech classification and improvement through the combination of cross-modal learning and graph-based relational modeling. These two modalities are combined in the proposed system with the help of a cross-modal transformer which identifies the contextual dependencies between the modalities with the help of attention mechanisms. The model uses self-supervised pre training to enhance the quality of representation in a situation where limited labeled data are available and thus it is able to learn both generalized and discriminative speech features. Besides that, a graph neural network is used to model structural dependencies between speech segments, with nodes and edges respectively modeling feature embeddings and temporal continuity and phonetic similarity. This two-sided representation enables the framework to share the analysis of local variations and international speech patterns linked to such disorders like dysarthria and Parkinsonian speech. Moreover, a speech enhancement module is presented that is lightweight and helps to reduce noise interference and to enhance the quality of input signals, thus, boosting the performance of the downstream classification. Both cross-modal transformers and graph neural networks coupled to each other lead to a higher level of robustness, feature discrimination, and interpretability. According to the experimental evidence, the given approach is more effective than the traditional ones, especially in noisy and low-resource conditions.*

Keywords: Speech disorder classification, cross-modal transformer, graph neural networks, self-supervised learning, speech enhancement, multimodal learning, dysarthria, acoustic features, linguistic features, pathological speech analysis.

I.INTRODUCTION

Communication, social interaction and quality of life are greatly affected by speech and language disorders, especially in people with neurological disorders who have dysarthria and Parkinson disease. This automatic recognition of pathological speech has thus been an issue of considerable study in the creation of aids diagnostic and treatment methods. Conventional methods of identifying a speech disorder are based on signal processing methods and the classical machine learning algorithms. As an example, first approaches were based on Mel-Frequency Cepstral Coefficients (MFCC) with Dynamic Time Warping (DTW) and Support Vector Machines (SVM) to determine the similarity between speech samples and categorise disordered and normal speech patterns, which showed a decent level of performance in controlled settings [1]. Nevertheless, these methods tend to have a problem in generalization where they are used in real world complex speech variations. As the deep learning progressed, hybrid models that use convolutional neural network (CNN) and optimization models have been proposed in order to enhance the performance of the classifications. Other techniques like FOA-SCNet use feature extraction and evolutionary optimization techniques to improve convergence and accuracy of prediction in pathological speech detection [2]. In spite of these advances, these models continue to emphasize the acoustic characteristics and they do not pay much attention to the incorporation of complementary linguistic or contextual information. Recently, research points to the shift to more advanced forms of artificial intelligence, such as multimodal learning, or deep neural architecture, where acoustic, prosodic and linguistic information is used to achieve better classification results. Extensive surveys have shown that hybrid and multimodal models are more robust and accurate than the traditional single-modality methods especially when used together with speech enhancement methods which enhance the quality of signal in noisy environments [3]. Also, other speech preprocessing problems like segmentation are also important as proper temporal segmentation has a strong influence on further feature extraction and classification. It has been demonstrated that both synthetic and emulated speech variations are data augmentation techniques that enhance the performance of models in low pathological speech dataset conditions [4]. Moreover, computer programs like digit classification systems of pathological speech show that signal representation methods of spectrogram may be a good match to normal speech, but fail in pathological conditions because of small and skewed datasets [5]. These issues have shown that much stronger, multimodal, structure sensitive frameworks are required which could account of not only local acoustic variations, but also global relational dependencies of speech data. Driven by these constraints, this paper discusses a combined method of cross-modal transformers and graph neural networks to enhance their robustness, feature representation, and classification in pathological speech analysis.

II.RELATED WORKS

Some of the recent studies in automatic pathological speech detection have investigated the use of diverse signal processing and machine learning as well as deep learning methods to enhance the classification performance across various speech conditions. The first methods mostly depended on speech conditions which are controlled and discriminative acoustic features are extracted using identical phonetic content. Nonetheless, the results of such approaches tend to be unable to be extended to real-life spontaneous speech. Sheikh and Kodrasi [6] examine the effect of speech mode on pathological speech recognition and show that the classical machine learning

methods find it difficult to capture speech-based pathology-related information in spontaneous speech, whereas deep learning models are more adaptable and have the capacity to extract more features. This explains why strong models are needed that can be used to deal with variability in speech production. Data inadequacy and inconsistency of dysarthric speech is still a significant issue. To solve this, Zhang et al. [7] suggest long-range and non-stationary variational autoencoder (LNVAE) to synthesize and augment dysarthric speech. Their models represents time-dependent dependence and non-stationary nature of pathological speech, which allows producing high-quality synthetic data. It has been demonstrated experimentally that with such augmentation strategies the recognition performance is greatly improved, avoiding the reliance on large clinical datasets. The use of deep learning in clinical speech analysis has also been attracted to interpretability and transparency of deep learning models. Gimeno-Gomez et al. [8] present an interpretable model that uses cross-attention speech representations that are self-supervised to examine the speech patterns of the Parkinson disease. Their methodology allows embedding-level as well as temporal-level interpretability, which makes the model more applicable to clinical use but without compromising the performance of classification in the model based on various benchmarks.

In pathological speech detection, feature representation learning is very vital. Kaloga et al. [9] suggest Multiview Canonical Correlation Analysis (MCCA) to eliminate irrelevant, unrelated data of spectrogram and embedding representations. Their approach helps to classify the data better and retain the interpretability of the results due to the simultaneous alignment of the various perspectives of the data, and this approach indicates that well-thought-out preprocessing and dimensionality reduction strategies can be used to improve traditional and modern classifiers. Stability to environmental noise is another major problem in the real world use. Amiri and Kodrasi [10] present a test-time adaptation model, which enhances model robustness in cases where there is noise by using noise-only segments to adapt the model. Their technique not only improves generalization without the need to retrain everything afresh, but also is appropriate to realistic deployment contexts with noise conditions whose characteristics are not predictable. New developments in pathological speech detection have involved deep feature extraction, end to end models, as well as foundation model representations, to enhance robustness and classification accuracy. The possible directions include one such prominent direction, which is the application of signal transformation techniques that are not parameter-intensive. Reddy et al. [11] suggest an end-to-end pathological speech detection system grounded on wavelet scattering networks, in which wavelet scattering network feature extraction, which is manually designed but mathematically found, is combined with a multilayer perceptron classifier. Their methodology proves that scattering transforms are more effective than convolutional neural networks in some situations, especially with compressed speech signals, at which they can achieve competitive results with traditional handcrafted feature-based systems.

Pathological speech classification has also been extensively done using deep convolutional neural networks. The Vavrek et al. [12] apply transfer learning using a VGG16-based framework to identify the pathological speech based on spectrogram representations. Their paper emphasizes the efficacy of pre-trained models in the cases of the scarcity of data. Using ensemble learning on vowel subsets, the model can be more accurate showing that deep learning models can be successfully generalized when combined with proper preprocessing strategies and data partitioning strategies. Recently, the development of the large language models (LLMs) has led to new opportunities in the speech-related classification tasks. Amiri et al. [13] test the multimodal LLMs in-context learning abilities like ChatGPT-4o to detect pathological speech. They find that the few-shot prompting strategies are capable of making competitive performance when compared to the traditional machine learning and deep learning practices. This literature indicates that it is possible to use foundation models as something that can be easily and readily integrated with speech analysis without the need to undergo intensive task-specific training. The combination of hybrid signal processing and deep learning methods has also been studied to increase features representation. According to Souli et al. [14], a hybrid system of scattering transform and deep convolutional neural network (ST-DCNN) is suggested to voice recognition of pathological cases. The extraction of robust features is done using the scattering transform, and a DCNN is used to classify the features. According to the experimental findings, the hybrid approach can be used to reach high recognition rates, which proves the usefulness of the combination of the deterministic feature extraction and deep learning classifier to enhance the robustness in clean speech scenarios. Besides that, recent studies have oriented to combining low-level acoustic descriptors and high-level representations of speech foundation models. Ariyanti et al. [15] present VOQANet and VOQANet+, that is, self-supervised embeddings that are used together with acoustic features like jitter, shimmer, and harmonics-to-noise ratio. Their findings indicate the improvement of both vowel and sentence-level testing and enhance the strength in the noisy settings. This shows the need to use complementary types of features in order to obtain more generalization and reliability in the actual pathological speech evaluation.

In general, these papers show the obvious development of the traditional signal processing and handcrafted features to the deep learning, hybrid architecture, and foundation models. Although these developments have been made, majority of the current approaches work on feature extraction, classification, or enhancement individually. The research gap is still on the unified frameworks that are capable of simultaneously combining the multimodal feature fusion, structural modeling, and enhancement of robustness. The proposed cross-modal transformer and graph neural network framework overcome these weaknesses, as both the acoustic-linguistic interactions and the relational dependencies are modeled and the preprocessing enhancement mechanisms are introduced in order to achieve high performance.

III.PROPOSED SYSTEM

The suggested system is aimed to overcome the problem of the pathological speech classification through the combination of the cross-modal representation learning, the graph-based relational modeling and the speech improvement in the framework of the single system. Figure.1 shows a proposed work architecture design. The system receives speech recording as input and processes it using a number of coordinated modules to select, refine and categorize relevant features related to speech disorders.

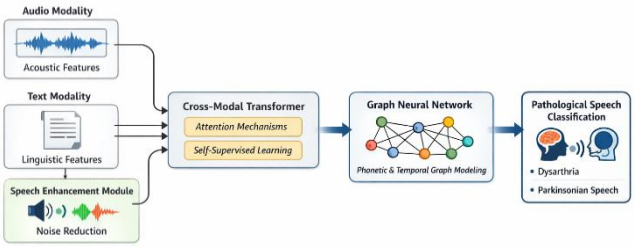

Figure.1 Proposed Work Architecture Diagram

The first step involves preprocessing of the input speech signal in which noise reduction and normalization is done. There is the use of a lightweight speech enhancement module to reduce the background noise and enhance the signal clarity that will result in the extraction of features of higher quality. Acoustic features of Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and spectral descriptors are obtained out of the enhanced signal. Simultaneously, linguistic characteristics are acquired with the help of automatic speech recognition (ASR) or text embeddings based on transcriptions. These cross-modal transformer architecture is then inputted with these acoustic and linguistic modalities. Attention mechanisms are used in the transformer in order to learn modalities-modalities interactions and time-dependent contextual dependencies. To enhance the process of generalization, particularly in small labeled data setting, the model integrates self-supervised pretraining, so that the model can be trained to acquire strong and transferable feature representations using unlabelled or partially labeled data. The learnt transformer embeddings are further represented in a graph. The nodes in this graph are speech segments embeddings and the edges are built on a temporal proximity and similarity of phonemes. This structure is fed through a graph neural network in order to learn relational structure and global tendencies in the speech sequence. The last step is the fusion of the outputs of the cross-modal transformer and the graph neural network and a classification layer. This layer forecasts the nature and existence of speech disorder. The integrated architecture is more robust and discriminates features better and is more interpretable than traditional methods especially in noisy and low-resource settings.

IV.METHODOLOGY

The suggested methodology presents a single framework, which is a combination of cross-modal transformers, graph neural networks, and speech enhancement methods to classify pathological speech robustly. The general pipeline is aimed at successfully acquiring complementary information between acoustic and linguistic modalities along with the modeling of the sequence and relationship among the speech data.

A. Input Acquisition and Preprocessing

The system starts with taking the raw speech signals using available datasets or clinical recording. Preprocessing steps that are applied to these signals include resampling, normalization and noise suppression. A lightweight speech enhancement engine is used in order to decrease the environmental noise and distortions. The step enhances the signal to noise ratio so that the latter feature extraction will give more credible representations.

Let the raw speech signal be represented as $s(t)$. The signal is first subjected to preprocessing and speech enhancement to reduce noise and improve clarity. A general denoising formulation can be expressed as

$$\hat{s}(t) = s(t) - n(t) \qquad (1)$$

where $n(t)$ denotes the estimated noise component and $\hat{s}(t)$ is the enhanced speech signal. From $\hat{s}(t)$, acoustic feature vectors $X_a \in \mathbb{R}^{T \times d_a}$ are extracted, where T represents the number of frames and d_a denotes the acoustic feature dimension. In parallel, linguistic features $X_l \in \mathbb{R}^{T \times d_l}$ are obtained from transcripts using embedding techniques.

B. Feature Extraction

Based on the improved speech cues, acoustic features are derived in order to obtain the physical and spectral characteristics of the speech. The most frequent ones are Mel-Frequency Cepstral Coefficients (MFCCs), spectral contrast, pitch, formants and energy-related descriptors. Simultaneously, linguistic characteristics are obtained through the translations of speech into text with the help of an automatic speech recognition (ASR) system or through the use of pre-trained language embeddings. These characteristics are semantic and syntactic characteristics of what is said.

C. Cross-Modal Transformer Encoding

The features of acoustics and language obtained are matched and fed into a cross-modal transformer network. The transformer uses the multi-head self-attention techniques to acquire intra-modal and inter-modal dependencies. Positional encodings are added in order to maintain time sequence. The model is able to attain contextual associations between speech sounds and the linguistic meanings through attention fusion. Pretraining Self-supervised pretraining is also optionally used to use unlabelled data and enhance quality of representation.

The acoustic and linguistic embeddings are projected into a shared representation space and concatenated as input tokens. The input embedding sequence is given by

$$X = [X_a; X_l] \qquad (2)$$

Self-attention within the transformer is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (3)$$

where $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are the query, key, and value matrices, respectively, and d_k is the key dimension. Multi-head attention is defined as

$$\text{MHA}(X) = \text{Concat}(h_1, h_2, \dots, h_h)W_O \qquad (4)$$

where each head is computed independently using the attention function. The transformer encoder produces contextualized embeddings $Z_T \in \mathbb{R}^{T \times d}$.

D. Graph Construction and Representation Learning

Transformer creates embeddings, which are utilized to make a graph structure. The graph nodes are the speech segment embeddings and the edges are created according to the temporal adjacency as well as the distances like the cosine similarity between the feature vectors. This type of graph formulation allows the explicit modeling of the relationship among the segments that might not be modeled by sequential models.

The transformer embeddings are used to construct a graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a speech segment embedding $z_i \in Z_T$. Edges are formed based on temporal and similarity criteria. The adjacency matrix A is defined as

$$A_{ij} = \begin{cases} 1, & \text{if } \text{sim}(z_i, z_j) \geq \tau \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where similarity is commonly measured using cosine similarity:

$$\text{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{\| z_i \| \| z_j \|} \quad (6)$$

E. Graph Neural Network Processing

Graph Neural Network (GNN) is implemented on the built graph in order to learn graph relationships between nodes. The GNN is able to update nodes representations by means of message passing and aggregation, using information about the neighboring nodes. This enables the model to model the local differences as well as global structural dependencies in the speech signal that is important in the detection of pathological patterns.

Node representations are refined using graph convolution operations. The propagation rule for a Graph Convolutional Network (GCN) layer is given by

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (7)$$

where $\tilde{A} = A + I$ is the adjacency matrix with self-loops, $\tilde{D}$ is the degree matrix, $H^{(l)}$ is the node representation at layer l, $W^{(l)}$ is the trainable weight matrix, and $\sigma(\cdot)$ is a non-linear activation function. The final graph embedding is denoted as Z_G .

F. Feature Fusion and Classification

The cross-modal transformer and the graph neural network outputs are fused with the help of concatenation or attention-based fusion techniques. The resulting combination representation is then subjected to fully connected layers and softmax/sigmoid classifier depending on the formulation of the task. The prediction of the presence and type of speech disorder is done by the classifier.

The outputs from the transformer and GNN are fused to form a joint representation:

$$Z = \left[Z_T^{\text{pool}} \,\|\, Z_G^{\text{pool}}\right] \quad (8)$$

where ‖ denotes concatenation and pooling (e.g., mean or attention pooling) is applied over node/sequence representations. The fused representation is passed through a classifier:

$$\hat{y} = \text{softmax}(W_c Z + b_c) \quad (9)$$

where W_c and b_c are learnable parameters and $\hat{y}$ represents the predicted probability distribution over classes.

G. Training Strategy and Optimization

The whole framework is brought to train through a supervised learning with cross-entropy loss. Gradient based optimization techniques like Adam optimizer are used to optimize it. The process of regularization such as dropout and early stopping are used to avoid overfitting. In case of self-supervised pretraining, the model is initially trained on unlabelled data of large sizes and then it is fine-tuned on labeled pathological speech data. The model is trained using cross-entropy loss defined as

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (10)$$

where C is the number of classes, y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability for class i. Optimization is performed using gradient-based methods to minimize L.

V.RESULT & DISCUSSION

The presented cross-modal transformer and graph neural network model was tested on the basis of definite performance indicators and compared with baseline models. The experiments were to evaluate the classification performance, noise resistance, the contribution of individual modules and computational behavior. The findings always indicate that the suggested hybrid architecture offers better results in the pathological speech classification exercises.

H. Overall Classification Performance

The major measures of evaluation are accuracy, precision, recall and F1-score. The proposed model performs better than the traditional models like CNN-based and LSTM-based models in incorporating multimodal information and contextual as well as relational dependencies.

TABLE I. PERFORMANCE COMPARISON OF MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN-Based Model	85.2	84.7	83.9	84.3
LSTM-Based Model	87.5	86.9	86.2	86.5
Transformer Only	90.3	89.8	89.5	89.6
GNN Only	88.1	87.6	87.2	87.4
Proposed Method	93.7	93.1	92.8	92.9

As it was observed in Table I, the suggested framework demonstrates the best performance according to all measures because of the complementary nature of the interaction between the cross-modal attention and graph-based relational learning.

I. Model Performance Visualization

Graphical representation is also used to depict the trends in the performance. The results of the comparison of accuracy and F1-score of the models by bar charts demonstrates clearly that the proposed method is at the top of the table in terms of performance.

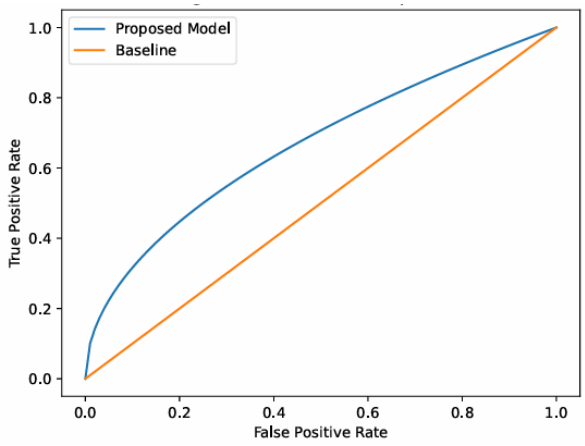

Figure 2. ROC Curve Comparison

Moreover, the training and validation accuracy curve versus the epochs will indicate steady convergence behavior with

minimum divergence and hence good generalization and less overfitting. ROC curve figure.2 indicates the trade-off between the rate of true positive and false positive of each model. The proposed approach has a better area under the curve (AUC), which shows better discriminative power. The curve is still more near the optimal top-left position indicating high sensitivity and specificity in picking pathological and non-pathological speech.

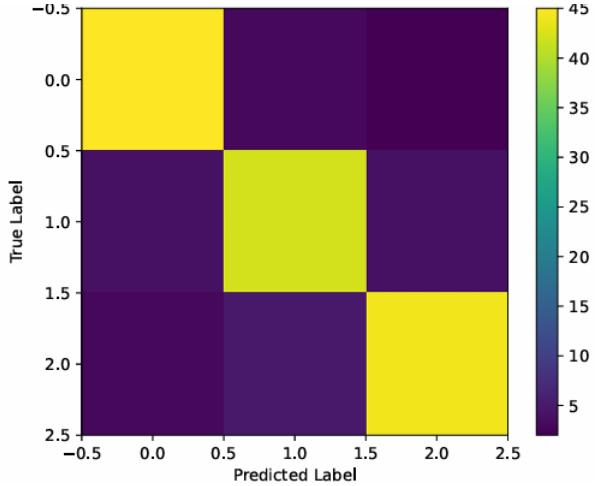

Figure 3. Confusion Matrix Visualization

The confusion matrix figure.3 gives an in-depth information on the classification results of various classes. The diagonal dominance in the matrix means that there were many right predictions and the off-diagonal values are comparatively low. This implies that there is minimal confusion between the close categories of speech disorders using the model.

J. *Noise Robustness Analysis*

The resilience of the suggested system was tested in different conditions of noise. There was also the addition of a speech enhancement module to counter the impact of environmental distortions. The experiment was conducted in low, moderate, and high noise levels to represent a real life situation.

TABLE II. PERFORMANCE UNDER NOISE CONDITIONS

Noise Level	Accuracy Without Enhancement (%)	Accuracy With Enhancement (%)
Low Noise	91.2	93.4
Moderate Noise	86.5	91.0
High Noise	80.3	88.2

The data in Table II show that the speech enhancement module has a great effect of enhancing the performance in all noise levels. This is more so when the noise level is high where the model has been found to perform consistently even in high-noise conditions as opposed to a significant decline in the absence of the enhancement.

K. *Ablation Study*

Ablution study was done to assess the input that each significant part of the framework makes. The effect on the overall performance was studied with the help of selective elimination of the transformer, the graph neural network, or the speech enhancement module.

TABLE III. ABLATION STUDY RESULTS

Configuration	Accuracy (%)	F1-Score (%)
Full Model	93.7	92.9
Without GNN	91.0	90.5
Without Transformer	89.8	89.2
Without Speech Enhancement	90.6	90.1

Table III shows the results that each component is significant in the final performance. Deleting the graph neural network lowers the model on relational dependency making and omitting the transformer influences the understanding of context. On the same note, the performance in noisy conditions when the speech enhancement module is omitted will be poor.

L. *Error Distribution Analysis*

In further analyzing model behavior, prediction errors were analyzed and presented in a form of an error distribution graph. The distribution reveals that the majority of the error in prediction is centered around a small range, which means that the model is going to perform consistently. Bigger errors are quite infrequent and they normally happen when there is overlapping of the characteristics of speech samples among classes.

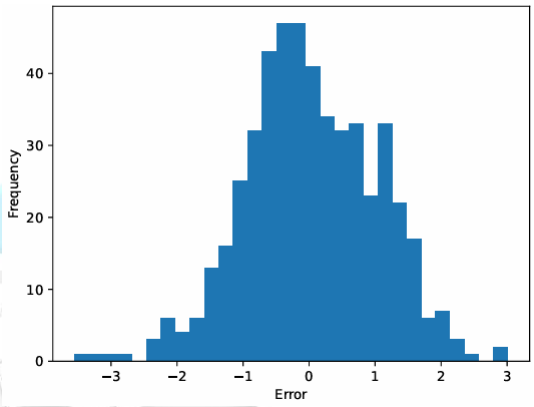

Figure 4. Prediction Error Distribution

The figure.4 plot of error distribution shows that most of the predictions lie within the range of the low error values and this indicates stability and reliability of the proposed framework.

M. *Computational Performance Analysis*

The measure of computational efficiency was calculated based on the training and inference time of varying models.

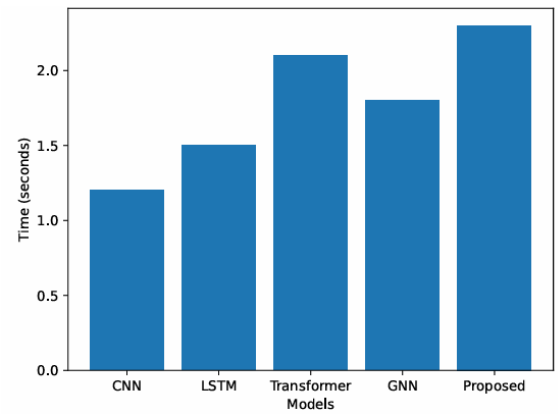

Figure 5. Computational Time Comparison Across Models

Though the suggested framework adds some computational blocks because of the combination of transformer and graph-based modules, the duration of inference is not too long to be used in practice. The comparison (figure.5) of the computational time shows that the proposed model has a marginally longer training time than the baseline models, but inference time is also similar, which makes it appropriate in clinical practice in the context of real-time or close time.

N. Discussion

The experimental analysis proves that the suggested cross-modal transformer and graph neural network architecture is a holistic and efficient approach to pathological speech classification. The combination of acoustic and linguistic modalities by the model leads to the acquisition of complementary information that would have been overlooked by unimodal methods. The cross-modal transformer improves the contextual comprehension with the help of attention mechanisms, whereas the graph neural network reinforces relational reasoning by training to capture the dependency between pieces of speech. The speech enhancement module also enhances the robustness especially in noisy environments, which also guarantees greater reliability in the extraction of features and makes predictions consistent. Comparative analysis reveals that there are steady improvements over baseline approaches in all the evaluation metrics which are accuracy, precision, recall, and F1-score. Also, ablation experiment proves that every part of it plays an important role in the overall performance. There is also high generalization ability and equal distribution of errors in the model which means that it can be used to deal with variability in speech patterns. All in all, the framework is quite appropriate to be practically implemented in clinical and low-resource environments.

VI.CONCLUSION

This paper introduces a single framework of pathological speech classification that combines cross-modal transformers, graph neural networks, and speech enhancement module to enhance robustness and diagnostic accuracy. The results prove that acoustic and linguistic modalities inspire more discriminative representations compared to unimodal ones. The cross-modal transformer is a successful model in capturing the contextual dependencies across modalities and the graph neural network is a successful model in improving the modeling of structural and relational patterns between speech segments. This is further enhanced by the addition of speech enhancement component to enhance the performance on noisy and real-life conditions by enhancing quality of the input signals. The main value of the work is the creation of a hybrid architecture, which integrates multimodal fusion and graph-based reasoning on a single pipeline. The overall results of the experiment indicate a consistent improvement in all standard metrics of evaluation, as well as good generalization and robustness. The significance of each of the modules in the context of the best performance is justified by the ablation studies that reveal the complementary functions of the transformer-based contextual learning and the graph-based structural modeling. The future research may involve investigating the incorporation of multimodal data sets that are larger and more complex pretraining methods to continue to improve representation learning. Also, the explainable AI methods can enhance the model interpretability to be used in clinical practice. Other potential opportunities of further research are the extensions of the framework to real-time deployment and other disorders involving speech.

VII.REFERENCES

[1] J. Liu et al., "Speech Disorders Classification in Phonetic Exams with MFCC and DTW," in Proc. IEEE 7th Int. Conf. Collaboration and Internet Computing (CIC), Atlanta, GA, USA, 2021, pp. 35–40, doi: 10.1109/CIC52973.2021.00015.

[2] K. Sujigarasharma et al., "Detection and Classification of Speech Disorder using FOA-SCNet," in Proc. IEEE 3rd Int. Conf. Computing and Information Technology (ICCIT), Tabuk, Saudi Arabia, 2023, pp. 391–395, doi: 10.1109/ICCIT58132.2023.10273910.

[3] M. S. Remya et al., "Artificial Intelligence for Speech Classification and Enhancement of Speech and Language Disorders: Techniques, Applications, and Future Directions," IEEE Access, vol. 13, pp. 177136–177159, 2025, doi: 10.1109/ACCESS.2025.3620114.

[4] S. A. Naeini et al., "Improving Dysarthric Speech Segmentation With Emulated and Synthetic Augmentation," IEEE Journal of Translational Engineering in Health and Medicine, vol. 12, pp. 382–389, 2024, doi: 10.1109/JTEHM.2024.3375323.

[5] Y. A. Goutham et al., "Digit Classification System for Normal and Pathological Speech," in Proc. Int. Conf. Smart Systems for Applications in Electrical Sciences (ICSSES), Tumakuru, India, 2024, pp. 1–5, doi: 10.1109/ICSSES62373.2024.10561448.

[6] S. A. Sheikh and I. Kodrasi, "Impact of Speech Mode in Automatic Pathological Speech Detection," in Proc. 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 81–85, doi: 10.23919/EUSIPCO63174.2024.10714947.

[7] D. Zhang, H. Zhang, W. Lu, W. Li, J. Wang, and J. Wei, "Long-Range and Non-Stationary Encoding for Dysarthric Speech Data Augmentation," IEEE Journal of Selected Topics in Signal Processing, vol. 19, no. 5, pp. 767–782, July 2025, doi: 10.1109/JSTSP.2025.3562417.

[8] D. Gimeno-Gómez, C. Botelho, A. Pompili, A. Abad, and C.-D. Martínez-Hinarejos, "Unveiling Interpretability in Self-Supervised Speech Representations for Parkinson's Diagnosis," IEEE Journal of Selected Topics in Signal Processing, vol. 19, no. 5, pp. 717–730, July 2025, doi: 10.1109/JSTSP.2025.3539845.

[9] Y. Kaloga, S. A. Sheikh, and I. Kodrasi, "Multiview Canonical Correlation Analysis for Automatic Pathological Speech Detection," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10888902.

[10] M. Amiri and I. Kodrasi, "Test-Time Adaptation for Automatic Pathological Speech Detection in Noisy Environments," in Proc. 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 2024, pp. 86–90, doi: 10.23919/EUSIPCO63174.2024.10715004.

[11] M. K. Reddy, Y. M. Keerthana, and P. Alku, "End-to-End Pathological Speech Detection Using Wavelet Scattering Network," IEEE Signal Processing Letters, vol. 29, pp. 1863–1867, 2022, doi: 10.1109/LSP.2022.3199669.

[12] L. Vavrek, M. Hires, D. Kumar, and P. Drotár, "Deep convolutional neural network for detection of pathological speech," in Proc. IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 2021, pp. 000245–000250, doi: 10.1109/SAMI50585.2021.9378656.

[13] M. Amiri, H. O. Shahreza, and I. Kodrasi, "Exploring In-Context Learning Capabilities of ChatGPT for Pathological Speech Detection," in Speech Communication; 16th ITG Conference, Berlin, Germany, 2025, pp. 46–50.

[14] S. Souli, R. Amami, A. Soltani, and S. B. Yahia, "On the use of Deep Learning and Scattering Transform for Pathological voices recognition," in Proc. 8th Int. Conf. Control, Decision and Information Technologies (CoDIT), Istanbul, Turkey, 2022, pp. 1055–1058, doi: 10.1109/CoDIT55151.2022.9803962.

[15] W. Ariyanti, K.-Y. Chen, S. M. Siniscalchi, H.-M. Wang, and Y. Tsao, "Towards Robust Assessment of Pathological Voices via Combined Low-Level Descriptors and Foundation Model Representations," IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2025.3644692.

