Deep Learning-Based Predictive Modeling for Male Depression Detection

R. S. Lakshmi Balaji[1*], Sumetee Jirapattarasakul[2], Kantapat Kwansomkid[3], Sirimonpak Suwannakhun[4], Thaweesak Yingthawornsuk[2*]

[1]Department of Advanced Computing Sciences, Academy of Maritime Education and Training University (AMET), India
[2]Department of Media Technology, King Mongkut's University of Technology Thonburi (KMUTT), Thailand
[3]Department of Computer Engineering, King Mongkut's University of Technology Thonburi (KMUTT), Thailand
[4]Faculty of Industrial Education and Technology, King Mongkut's University of Technology Thonburi (KMUTT), Thailand

[*]**Email:** rslbalaji@proton.me, thaweesak.yin@kmutt.ac.th

Abstract

This project utilizes machine learning techniques to construct a highly precise model for categorizing audio recordings, with a particular focus on male speakers and their mental health conditions. The audio recordings are classified into three distinct categories: Remitted (RMT), Depressed (DPR), and High-risk for suicide (HRK), with special attention to gender-specific nuances. We have conducted an extensive exploration and comparison of diverse machine learning models, including 1D and 2D Convolutional Neural Networks (CNNs), Support Vector Machine (SVM), and Recurrent Neural Network with Long Short-Term Memory (RNN-LSTM). Our primary goal is to identify the most accurate model for classifying these male audio recordings, potentially offering a valuable tool for early detection and intervention in male mental health issues. We eagerly look forward to sharing our research results, aiming to make a substantial contribution to the understanding and treatment of depression among males. In this paper, we present the results of our investigation, comparing the accuracy of audio classification using 25-second and 1-minute speech segmentation.

Keywords

Deep Learning, Predictive, Modeling, Depression, Male speech sample

Submission: 23 December 2024; **Acceptance:** 5 February 2025; **Available online:** February 2025

Introduction

Mental health has garnered increasing global attention in recent years, emerging as a significant concern. An estimated 300 million individuals worldwide contend with depression, a pervasive mental health disorder that ranks as the primary cause of disability. Regrettably, depression can lead to tragic outcomes, with nearly 800,000 lives lost annually, as reported by the World Health Organization.

Swift identification and intervention in mental health disorders play a pivotal role in effectively managing and potentially averting severe consequences. Traditional diagnostic methods often involve clinical interviews and self-reported questionnaires. Nevertheless, these conventional approaches exhibit limitations such as subjectivity, time-intensive processes, and reliance on patients' cooperation and self-awareness.

In the dynamic field of computational psychiatry, there exist promising and objective avenues for diagnosing mental health conditions. A particularly encouraging domain within this field focuses on the application of machine learning techniques to scrutinize speech and voice patterns. The human voice harbors a wealth of information about an individual's emotional and mental state, and deviations in voice characteristics can offer valuable insights into underlying mental health conditions.

Related work

Deep Learning applications have witnessed significant expansion in the realm of mental health research. A pivotal area of investigation within this field involves the analysis of audio signals to discern various emotional states. These studies have consistently demonstrated impressive accuracy when employing Deep Learning techniques. Deep Learning has proven highly valuable in identifying mental health conditions such as depression and assessing suicide risk. For instance, researchers have applied Support Vector Machines (SVMs) to classify audio recordings based on the severity of depression, yielding promising results. Nevertheless, these models have also encountered challenges in distinguishing subtle differences between mild and moderate depression. More recently, convolutional neural networks (CNNs) have been harnessed for audio classification tasks, often surpassing the performance of traditional Machine Learning models. This highlights their potential efficacy in categorizing audio signals based on the speaker's emotional state. Furthermore, the use of recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, has exhibited promise in sequence classification tasks, including the classification of audio signals. Despite these advancements, a research gap persists in integrating these methodologies for gender-specific audio data classification, particularly within the context of depression and suicide risk assessment. This project seeks to address this gap by concentrating exclusively on male subjects and evaluating the performance of Deep Learning models (CNN1D, CNN2D, RNN-LSTM) in categorizing audio recordings into three distinct groups: remitted, depressed, and high-risk for suicide, with a specific focus on excluding data from female Speakers.

Methodology

- **Data Preprocessing and Visualization**

The audio recordings underwent a segmentation process, where they were divided into 2-second segments. This segmentation was vital to maintain a uniform size for subsequent analysis. This standardization of input size is crucial for training Machine Learning models consistently, enabling a fair comparison of their performance. We also kept track of the number of segments generated for each class (Remitted, Depressed, High-Risk of Suicide) and gender (Male, Female). This tracking allowed us to gain insights into the distribution of our data, which is of paramount importance. An imbalance in class or gender distribution could potentially impact the performance of our models. To comprehensively understand the audio data and effectively distinguish between the classes, we performed feature extraction on the audio recordings.

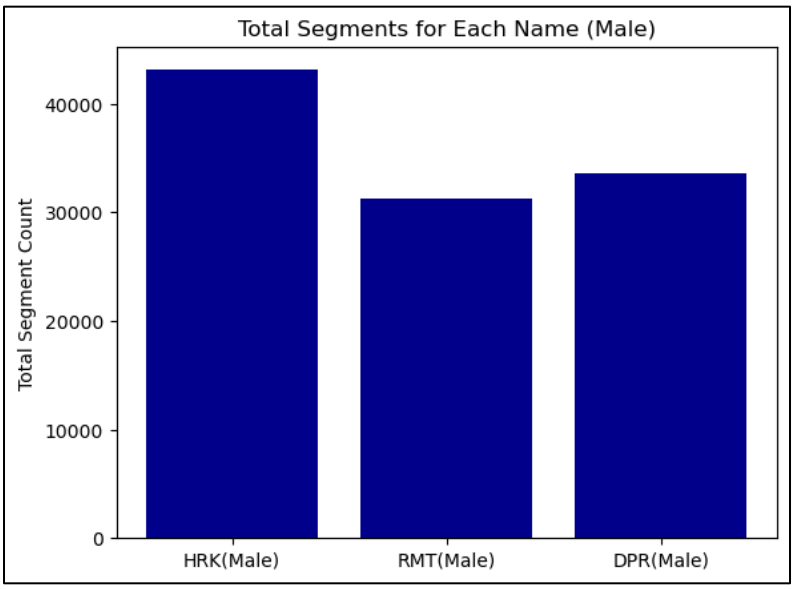

Figure 1. Total Segments for each class

This process encompassed a range of audio features, including Waveplot, Zero-Crossing Rate (ZRC), Spectral Centroids, Spectral Rolloff, Mel Frequency Cepstral Coefficients (MFCCs), and Mel Spectrogram. Each of these features captures distinct facets of the audio signal, offering unique insights into the underlying patterns that may differentiate the various classes The selection of these specific features was guided by their widespread application in audio and speech analysis, as well as their established capability to capture pertinent information from audio signals.

- **Feature Extraction**

After conducting an extensive analysis of the various extracted features, we have decided to utilize Mel Frequency Cepstral Coefficients (MFCCs) as the primary features for our model training. This decision is based on the MFCCs' capability to capture essential characteristics of the audio signals, which have the potential to differentiate between different classes.

We have extracted with 128 sufficient MFCC features from each audio segment, as this number represents the maximum feasible features obtainable during the extraction process. MFCC features

offer a robust and compact representation of the power spectrum of audio signals, making them an ideal choice for our classification task.

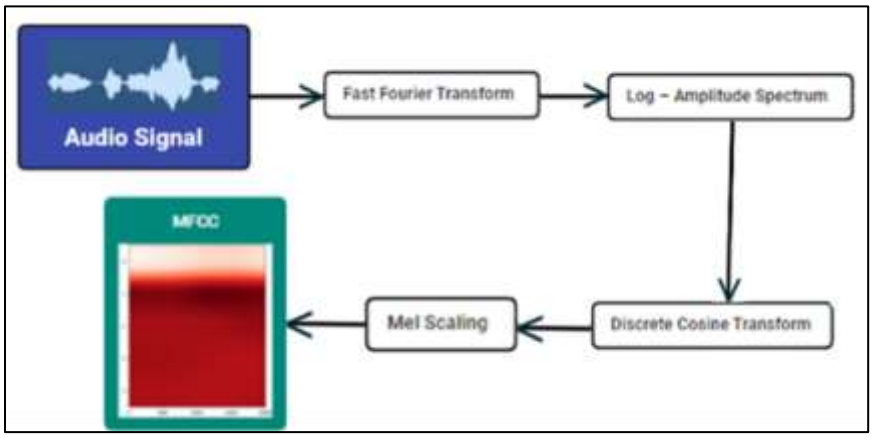

Figure 2. MFCCS Extraction

To further investigate the utility of these features, we conducted a Principal Component Analysis (PCA). The PCA provided a visual representation of the MFCC features representing audio samples for each categorized class. This visual inspection revealed significant differences between the classes, which further reinforces our choice of using MFCCs as the primary features for validating our models. The resulting PCA plots clearly illustrate the distinctions in MFCC features among the classes, providing strong visual evidence in support of our decision to use MFCCs as the primary features in our analysis.

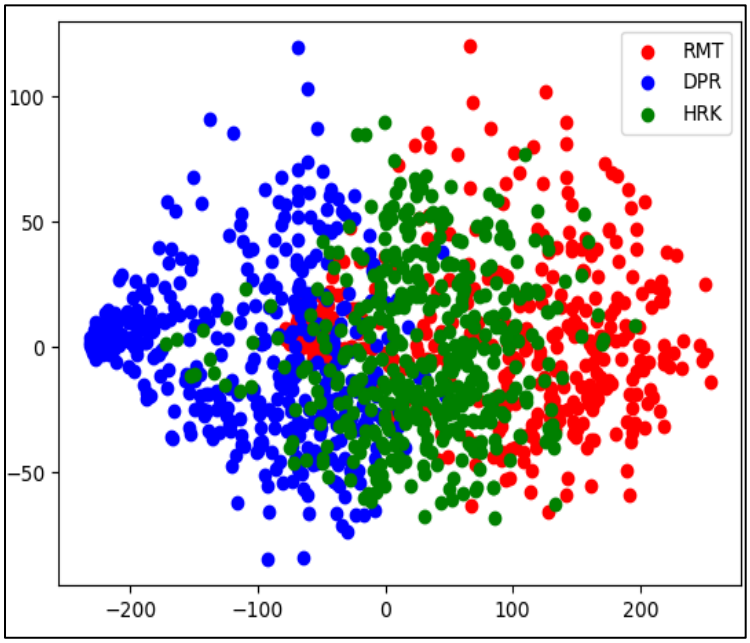

Figure 3. PCA Plotting

Implementation

All figures and tables should be placed after their first mention in the text. Place figure captions below the figures and centered within a column. Place table titles above the tables and flush left. Should figures have two parts, include labels "(a)" and "(b)" as part of the artwork. Please verify that the figures and tables mentioned in the text actually exist.
When citing a figure(s) in the text, they should be referred as, for example, "Fig. 1" or "Figs. 1 – 3." Use "Figure" or "Figures" when it starts the sentence and also do not abbreviate "Table". There should be one line of space above the figure and one line of space below the caption before the text continues. This is also applied to the table.

- **Model Selection**

In this research, we employed four distinct machine learning models, specifically chosen for their suitability in the realm of audio classification. These models encompass:

- One-Dimensional Convolutional Neural Network (1D-CNN): The efficacy of 1D-CNNs in capturing temporal attributes within audio signals through the extraction of pertinent local 1D subsequences is widely acknowledged. This property renders them highly suitable for tasks involving audio classification.

- Two-Dimensional Convolutional Neural Network (2D-CNN): 2D-CNNs exhibit remarkable proficiency in handling spectrograms as 2D images, thereby enabling the effective capture of both spectral and temporal features.

- Long Short-Term Memory (LSTM): LSTMs, a category of Recurrent Neural Networks, excel in learning and retaining information over extended sequences. This characteristic proves particularly advantageous in the domain of audio signal processing, where extended sequences of sound samples are inherent.

- Support Vector Machine (SVM): SVM, a versatile machine learning model capable of both linear and nonlinear classification, was chosen for its robustness in high-dimensional spaces. It presents a practical and well-suited choice for the objectives of this study.

- **Model Training**

The training process for each model was conducted in two distinct study cases. In the first case, the models were trained to differentiate between 'Remitted' and 'Depressed' conditions, and in the second stage, they were trained to distinguish 'Remitted' and 'HRK' conditions. This two-case procedure was specifically applied to male voices. For the neural network-based models (1D-CNN, 2D-CNN, and LSTM), we used the Adam optimizer to iteratively update network weights based on the training data. To prevent overfitting, we incorporated Dropout layers into our network architecture as a regularization technique. Additionally, we employed Early Stopping as another form of regularization. Early Stopping halts the training process if the model's performance on the validation set does not show improvement after a certain number of training epochs reached.

Model Construction:

- 1D-CNN: We built the 1D-CNN model with a structure comprising two convolutional layers, two max pooling layers, and two dense layers. One of the dense layers incorporates dropout. Relu activation functions were used in all layers except the final one, which employed sigmoid activation for binary classification.

- 2D-CNN: The 2D-CNN model followed a similar design philosophy, involving convolutional, max pooling, and dense layers. To align with the 2D nature of this model, the data was appropriately reshaped.

- LSTM: For the LSTM model, we employed two LSTM layers followed by a dense output layer with SoftMax activation for binary classification. To accommodate the sequential input required by LSTM networks, the data underwent reshaping.

- SVM: The SVM model was trained using a linear kernel. Before being input into the model, the data was reshaped and scaled as necessary.

- **Evaluation Metrics**

In this section, we focus on the evaluation metrics used to assess model performance. The primary metrics considered include precision, recall, and the F1-score, which are pivotal for evaluating predictive quality. Precision quantifies the proportion of true positive identifications among all positive
predictions. The F1-score, a balanced metric, factors in both precision and recall to offer a comprehensive evaluation of model performance. Given the medical context of our study, special emphasis is placed on recall. Recall measures the proportion of actual positives correctly identified, assessing the model's ability to minimize false negatives. In the context of depression, avoiding false negatives—misclassifying a depressed patient as non-depressed—is crucial to ensure patients receive the necessary care.
These operations elucidate the computation of precision, recall, and the F1-score. They demonstrate how True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) are utilized for these calculations. Precision represents the ratio of TP to the sum of TP and FP, while recall is the ratio of TP to the sum of TP and FN. The F1-score, a harmonic mean of precision and recall, provides a comprehensive measure of model performance. To gain a more detailed understanding of the model's performance, we employ a confusion matrix, also known as an error matrix. This two-by-two table reports the counts of false positives, false negatives, true positives, and true negatives. Unlike accuracy, which can be misleading, especially in skewed datasets, a confusion matrix offers a more thorough analysis of the model's effectiveness.

Results and discussion

In the confusion matrix displayed in the accompanying image, Class 0 is attributed to 188 true negatives, while Class 1 is associated with 88 true positives, demonstrating high performance in accurately classifying both classes.

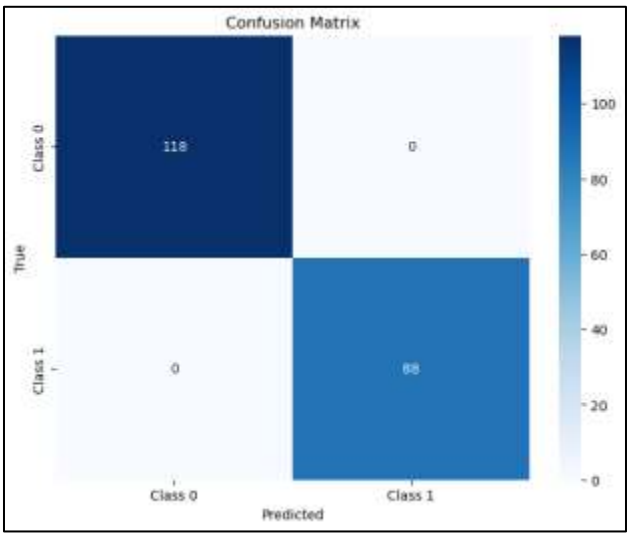

Figure 4. Confusion Matrix

During the training of a neural network for audio classification, significant progress was observed over five epochs. The model's accuracy steadily increased, reaching 100% accuracy on the validation data by the final epoch. This promising trend is visually depicted in the following graph, which shows a clear and consistent rise in accuracy over the training epochs, indicating successful learning and the potential application in classifying audio segments.

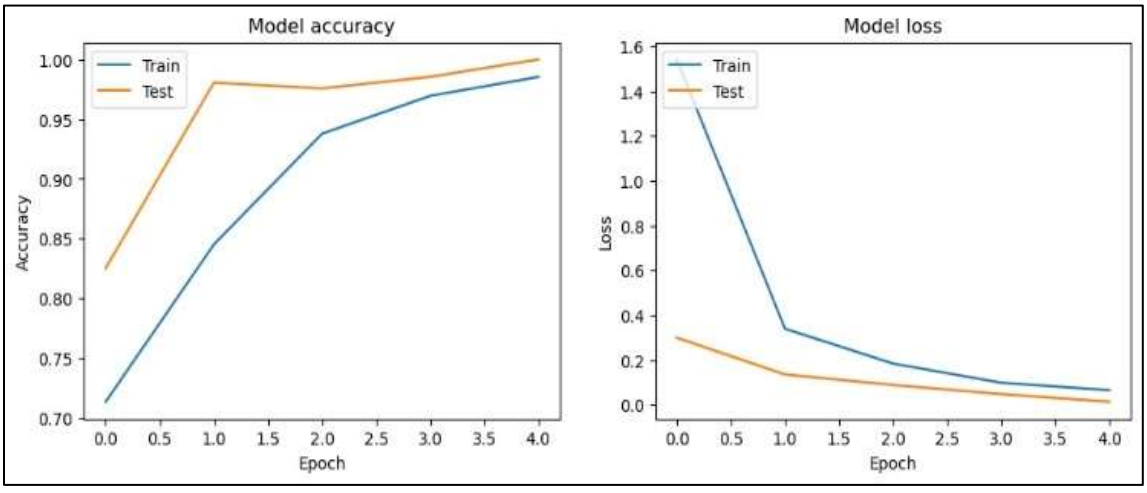

Figure 5. Model Graph

- Data Segmented in 25-Seconds
  - Remitted/Depressed

Table 1. PERFORMANCE OF DIFFERENT MODELS OF MALE (RMT/DPR) – 25 SECONDS SEGMENTED

Model	Accuracy	Precision	Recall	F1
CNN1D	0.90	0.92	0.92	0.91
CNN2D	0.93	0.93	0.93	0.93
LSTM	0.50	0.54	0.26	0.50
SVM	0.92	0.92	0.94	0.93

- Remitted / High Risk for Suicide

Table 11. PERFORMANCE OF DIFFERENT MODELS OF MALE (RMT/HRK) – 25 SECONDS SEGMENTED

Model	Accuracy	Precision	Recall	F1
CNN1D	0.94	0.90	0.96	0.93
CNN2D	0.90	0.90	0.90	0.90
LSTM	0.49	0.52	0.28	0.49
SVM	0.95	0.92	0.93	0.93

The results of model performance for distinct categories, specifically "Remitted/Depressed" (RMT/DPR) and "Remitted/High Risk for Suicide" (RMT/HRK), utilizing 25-second segmented data, are presented in Tables I and II, respectively. In the "Remitted/Depressed" category, the CNN1D model achieved an accuracy of 0.90, accompanied by robust precision and recall scores of 0.92 each, leading to F1 score of 0.91.

The CNN2D model performed even better with an accuracy of 0.93, along with equally commendable precision, recall, and F1 scores of 0.93. However, the LSTM model struggled in this category, managing only an accuracy of 0.50, with a precision of 0.54, recall of 0.26, and F1 score of 0.50. Conversely, the SVM model excelled with an accuracy of 0.92, boasting a precision of 0.92, recall of 0.94, and F1 score of 0.93.

Shifting to the "Remitted/High Risk for Suicide" category, the CNN1D model displayed a robust performance, achieving an accuracy of 0.94, precision of 0.90, recall of 0.96, and F1 score of 0.93. The CNN2D model secured an accuracy of 0.90, maintaining well-balanced precision, recall, and F1 scores of 0.90.

However, similar to the previous category, the LSTM model encountered challenges, recording an accuracy of just 0.49, with precision of 0.52, recall of 0.28, and F1 score of 0.49. In contrast, the SVM model performed yet again in this category, achieving accuracy of 0.95, precision of 0.92, recall of 0.93, and F1 score of 0.93.

- Data Segmented in 1 Minute

  - Remitted/Depressed

Table 111. PERFORMANCE OF DIFFERENT MODELS (RMT/DPR) – 1 MINUTE SEGMENTED

Model	Accuracy	Precision	Recall	F1
CNN1D	1.00	1.00	1.00	1.00
CNN2D	0.91	0.93	0.90	0.90
LSTM	0.50	0.54	0.72	0.59
SVM	0.96	0.96	0.99	0.98

- Remitted / High Risk for Suicide

Table 1V. PERFORMANCE OF DIFFERENT MODELS (RMT/HRK) – 1 MINUTE SEGMENTED

Model	Accuracy	Precision	Recall	F1
CNN1D	0.99	0.99	0.99	0.99
CNN2D	0.91	0.93	0.89	0.91
LSTM	0.49	0.50	0.71	0.58
SVM	0.99	0.99	0.97	0.98

The results of model performance in the categories of Remitted/Depressed (RMT/DPR) and Remitted/High Risk for Suicide (RMT/HRK), using 1-minute segmented data, have been summarized in Tables III and IV, respectively. In the Remitted/Depressed category, the CNN1D model demonstrated outstanding performance, achieving accuracy of 1.0, accompanied by high precision, recall, and F1 scores of 1.0. The CNN2D model also delivered commendable results with accuracy of 0.91, and notable precision and recall scores of 0.93 and 0.90, resulting in F1 score of 0.90. Conversely, the LSTM model faced persistent challenges, recording an accuracy of 0.50, with precision of 0.54, a significant recall of 0.72, and F1 score of 0.59. Once again, the SVM model performed impressively, achieving accuracy of 0.96, and featuring remarkable precision, recall, and F1 scores of 0.96, 0.99, and 0.98, respectively. Transitioning to the Remitted/High Risk for Suicide category, the CNN1D model showcased robust performance, attaining an accuracy of 0.99, along with exceptional precision, recall, and F1 scores of 0.99. The CNN2D model maintained solid performance with accuracy of 0.91 and commendable precision, recall, and F1 scores of 0.93, 0.89, and 0.91, respectively. However, the LSTM model continued to face challenges, registering accuracy of 0.49, with precision of 0.50, a notable recall of 0.71, and F1 score of 0.58. In contrast, the SVM model excelled yet again in this category, achieving accuracy of 0.99, with precision and recall scores of 0.99 and 0.97, leading to F1 score of 0.98.

Discussion

In our comprehensive research investigation, we conducted a thorough evaluation of various machine learning and deep learning models aimed at classifying individuals into two significant mental health categories: Remitted/Depressed (RMT/DPR) and Remitted/High Risk for Suicide (RMT/HRK). Our study involved the analysis of two different temporal resolutions: 25-second and 1-minute segmented data, which provided valuable insights into the performance of these models. For the 25-second segmented data, the CNN1D deep learning model consistently demonstrated strong performance, achieving accuracy scores of 0.90 for RMT/DPR and 0.94 for

RMT/HRK. Similarly, the CNN2D deep learning model exhibited robust performance with accuracy scores of 0.93 for RMT/DPR and 0.90 for RMT/HRK. In contrast, the LSTM model, another deep learning architecture, faced challenges, especially in the RMT/DPR category, where it achieved a noticeably lower accuracy of 0.50. On the other hand, the SVM model, a conventional machine learning approach, consistently displayed impressive accuracy, precision, recall, and F1 scores for both categories, underscoring its reliability. When analysing the 1-minute segmented data, the deep learning model CNN1D continued to excel with a perfect accuracy score of 1.0 for RMT/DPR and an accuracy score of 0.99 for RMT/HRK. The CNN2D model, another deep learning approach, maintained commendable results, achieving accuracy scores of 0.91 for both categories. However, the LSTM model's performance remained suboptimal, particularly in terms of accuracy and F1 scores. Meanwhile, the SVM model, a traditional machine learning method, consistently outperformed all other models, boasting impressive accuracy, precision, recall, and F1 scores consistently exceeding 0.96. In summary, our research findings highlight the critical importance of selecting the most suitable machine learning or deep learning model for mental health classification tasks. Deep learning models, particularly CNN1D and CNN2D, consistently exhibited robust and reliable performance, positioning them as strong candidates for applications in this domain. The SVM model also showed promise, while the LSTM model faced challenges in achieving competitive performance, especially in the RMT/DPR category. These results emphasize the potential of deep learning models, in conjunction with traditional machine learning techniques, for accurately identifying individuals with various mental health conditions. They also underscore the need to adapt model choices to specific data characteristics to ensure precise predictions in the field of mental health classification.

Conclusion

In conclusion, this study conducted an extensive evaluation of deep learning models, including 1D Convolutional Neural Network (CNN1D), 2D Convolutional Neural Network (CNN2D), Long Short-Term Memory (LSTM), and Support Vector Machine (SVM), for the binary classification of depression in male subjects. The results consistently demonstrated that the Convolutional Neural Networks (CNN1D and CNN2D) outperformed the other models across various evaluation metrics, including accuracy, precision, recall, F1 scores. Their ability to minimize false results is particularly crucial in a medical diagnosis context. The outstanding performance of CNN models underscores their potential as powerful tools for classifying depression among male patients, enabling early detection and treatment. Consequently, we have decided to further develop and adapt the CNN1D and CNN2D models for a more complex three-category classification task. The robust performance of these models in previous experiments instils optimism regarding their adaptability and potential success in future endeavors. Additionally, this study revealed a significant discrepancy in accuracy between 25-second and 1-minute audio data segments, highlighting the importance of data segmentation in influencing classification accuracy. Future research will focus on optimizing the segmentation approach to enhance model performance.

Future directions

Considering the promising results achieved in this study, there exist promising avenues for future research in the domain of depression detection among male subjects using deep learning models. Researchers may consider the integration of multi-modal data, combining audio with other relevant sources, to enhance overall model performance. Advanced feature engineering and model fine-tuning present opportunities for further model optimization. It is essential to validate the robustness and generalization of selected models through cross-dataset evaluations. Additionally, exploring transfer learning from pre-trained models could enhance classification accuracy. Developing real-time monitoring systems and fostering explainable AI will facilitate practical clinical applications, advancing early detection and treatment of depression in males.

Acknowledgement

This work has been collaborated between Department of Media Technology, King Mongkut's University of Technology Thonburi, Bangkok Thailand and Department of Advanced Computing Sciences, AMET (Academy of Maritime Education and Training) University, Chennai India.

References

Cummins, N., Epps, J., Sethu, V., Breakspear, M., & Goecke, R. (2013). Modeling spectral variability for the classification of depressed speech. *Interspeech* (pp. 857-861). https://www.researchgate.net/publication/279888238_Modeling_spectral_variability_for_the_classification_of_depressed_speech

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

Gupta, S., & Sharma, R. (2019). An ensemble of deep learning models for depression detection from speech with MFCC features. Frontiers in Psychiatry.

Kanoujia, S., & Karuppanan, P. (2024, March). Depression Detection in Speech Using ML and DL Algorithm. In 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI) (Vol. 2, pp. 1-5). https://doi.org/10.1109/IATMSI60426.2024.10503510

Lee, M., & Kim, S. (2019). MFCC-based depression detection in natural conversations with attention mechanisms. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.

Li, K., & Zhang, Q. (2021). Exploring multi-modal data for depression detection: A deep learning approach with MFCC features. Multimedia Tools and Applications.

Nguyen, T., & Nguyen, K. (2020). Depression detection using MFCC and bidirectional LSTM in social media audio. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., & Othmani, A. (2022). MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomedical Signal Processing and Control, 71, 103107. https://doi.org/10.1016/j.bspc.2021.103107

Scibelli, F., Roffo, G., Tayarani, M., Bartoli, L., De Mattia, G., Esposito, A., & Vinciarelli, A. (2018). Depression speaks: Automatic discrimination between depressed and non-depressed speakers based on nonverbal speech features. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 6842-6846. IEEE. https://doi.org/10.1109/ICASSP.2018.8461858

Singh, P., & Das, A. (2020). Depression detection from audio data using transfer learning and MFCC features. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems.

Suparatpinyo, S., & Soonthornphisaj, N. (2023). Smart voice recognition based on deep learning for depression diagnosis. Artificial Life and Robotics, 28(2), 332-342. https://doi.org/10.1007/s10015-023-00852-4

Verma, A., Jain, P., & Kumar, T. (2023). An Effective Depression Diagnostic System Using Speech Signal Analysis Through Deep Learning Methods. *International Journal on Artificial Intelligence Tools*, *32*(02), 2340004. https://doi.org/10.1142/S0218213023400043

Wang, H., & Liu, X. (2018). Depression detection using MFCC features and convolutional recurrent neural networks. Proceedings of the International Conference on Pattern Recognition.

Wu, G., & Zhang, Y. (2021). A deep learning framework for real-time depression detection from speech signals using MFCC features. Computers in Biology and Medicine.

Yingthawornsuk, T., & Thanawattano, C. (2010). Characterizing sub-band spectral entropy based acoustics as assessment of vocal correlate of depression. ICCAS 2010, 1179-1183. https://doi.org/10.1109/ICCAS.2010.5669862

Yingthawornsuk, T. (2016). Spectral entropy in speech for classification of depressed speakers. 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 679-682. https://doi.org/10.1109/SITIS.2016.113