

TinyML for Emotion Detection in Voice Signals: Evaluating and Proposing Algorithms for IoT Wearable Devices

Zain Ahmed
Dept. of CSE
BRAC University
Dhaka, Bangladesh
zain.ahmed@g.bracu.ac.bd

Hasibul Hasan Ahmed
Dept. of CS
BRAC University
Dhaka, Bangladesh
hasibul.hasan.ahmed@g.bracu.ac.bd

Tshewang Choden
Dept. CSE
BRAC University
Dhaka, Bangladesh
tshewang.choden@g.bracu.ac.bd

Nutan Chaudhary
Dept. of CSE
BRAC University
Dhaka, Bangladesh
nutan.chaudhary@g.bracu.ac.bd

Fahim Ul Islam
Dept. of CSE
BRAC University
Dhaka, Bangladesh
fahim.islam@bracu.ac.bd

Amitabha Chakrabarty
Dept. of CSE
BRAC University
Dhaka, Bangladesh
amitabha@bracu.ac.bd

Abstract—Voice emotion recognition is critical for applications such as intelligent tutoring, audio mining, security, telecommunication, HCI, and human-machine interactions. With the advent of IoT and wearable technology, there are new opportunities for real-time, remote emotion detection through voice. This thesis explores Tiny Machine Learning (TinyML) for voice emotion recognition, particularly in IoT wearables. We evaluated Bidirectional-LSTM (BiLSTM) and CNN on vector quantization and raw data, achieving accuracies of 88%, 80%, 85%, and 81%, respectively. Additionally, LSTM and GRU on raw data showed accuracy rates of 86% and 82% respectively, using a composite dataset that includes RAVDESS, CREMA-D, TESS, and SAVEE. We implemented 3 other traditional ML models as well, but didn't proceed with it for hardware implementation as DNN models matched our criteria better. The best-performing models were implemented in the TinyML framework using TensorFlow Lite. Benchmarking highlighted that RNN-based models performed best, notably BiLSTM, LSTM, and GRU, alongside CNN. Hardware validation on Raspberry Pi 4 confirmed that the BiLSTM model is most suitable for speech emotion recognition in the TinyML domain, demonstrating reliable performance within resource and power constraints. These findings contribute to advancing voice emotion recognition, TinyML, and IoT, enhancing human-machine interactions across various applications

Index Terms—Tiny Machine Learning (TinyML); Emotion Recognition; SER; Voice Signals; Wearable IoT Devices; BiLSTM; CNNs; LSTM; GRU;

I. INTRODUCTION

The demand for emotion identification has surged due to the increased use of video conferencing platforms during the epidemic. Voice analysis emerges as a versatile and covert alternative to video-based emotion recognition, which faces privacy issues and requires an unobstructed camera view. This study explores speech recognition algorithms to identify and evaluate auditory cues corresponding to different emotional states. We examine various TinyML models suitable for

resource-constrained environments of wearable devices. Recurrent neural networks (RNNs), particularly BiLSTM, LSTM, and GRU, are highlighted for their proficiency in processing sequential and temporal data like voice signals. The study emphasizes the BiLSTM model for its potential in speech emotion recognition. Furthermore, we explore integrating these TinyML models into Internet of Things (IoT) frameworks for real-time emotion detection. Prototyping uses commonly available micro-controllers like Arduino Uno, Raspberry Pi, ESP32, and Arduino Nano BLE 33. A comparative analysis of different models and algorithms is performed, evaluating their performance based on accuracy, F1 score, confusion matrix, precision, and recall. The goal is to design a system that achieves high accuracy in emotion detection while optimizing power usage, and enhancing the utility and sustainability of IoT wearable devices.

A. Motivation

We intend to work on voice-based emotion recognition with Tiny Machine Learning (TinyML) to create a supplemental tool that improves human-computer interactions, particularly in wearable IoT devices. The exponential expansion of voice assistants and the limits of video-based emotional identification highlights the need for a versatile and discreet alternative, such as voice analysis. We hope to develop systems that aid in real-time emotional recognition, which is critical for applications in healthcare, mental health, and social interaction, using TinyML's ability to execute efficient on-device processing. With institutional funding, we hope to increase accessibility through mobile apps and websites, raising awareness and contributing to the emerging field of emotion-sensitive AI technology.

B. RESEARCH CONTRIBUTION

1) *Problem statement:* The growing reliance on voice assistants, combined with the limitations of video-based emotional identification, highlights the need for a more versatile and discrete technique of emotion detection. Current systems frequently encounter obstacles such as privacy concerns and the need for unobstructed visual inputs, which add unnecessary complexity to IoT applications. The optimal solution is to directly detect emotions from speech using only audio data, thereby reducing input dimensions. There is an urgent need for effective real-time emotion identification algorithms that can seamlessly integrate into wearable IoT devices and use voice signals exclusively to predict emotions. This study aims to address these challenges by using Tiny Machine Learning (TinyML) to create efficient, on-device voice-based emotion recognition systems. The goal is to develop a supplementary tool that enhances human-computer interactions, particularly in healthcare, mental health, and social interaction applications, and to explore methods to increase accessibility via mobile apps and websites.

2) *Solutions:* Our work contributes to the field of real-time voice-based emotion recognition in wearable IoT devices by employing a comprehensive research methodology to develop and evaluate Tiny Machine Learning (TinyML) models. We begin by collecting and preprocessing diverse speech datasets, including various dialects and accents, where techniques such as normalizing audio signals, extracting Mel-Frequency Cepstral Coefficients (MFCC), and augmenting the dataset enhance model resilience. Various machine learning models and classifiers, including Bidirectional Long Short-Term Memory (BiLSTM) networks, Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, are selected and implemented for processing sequential and temporal data within the constraints of wearable devices. Supervised learning with labeled datasets enables model training to recognize and classify emotional states based on audio inputs, with performance optimization through hyperparameter tuning and cross-validation. Evaluation metrics such as accuracy, precision, recall, and F1-score assess model performance, complemented by confusion matrices and classification reports. Integration of these models into IoT frameworks, utilizing Raspberry Pi 4B for real-time emotional state detection, is investigated with comparative analysis considering computing complexity and power consumption for continuous operation in resource-constrained environments. Our research aims to develop an energy-efficient and effective emotion recognition tool for healthcare, mental health, and interpersonal relationships, examining TinyML model resilience and flexibility in real-world scenarios to optimize power usage while maintaining accuracy and efficiency.

II. LITERATURE REVIEW

Several papers contribute significantly to the field of real-time voice-based emotion recognition, leveraging deep learning methodologies and diverse datasets. In paper [5] employs CNNs to extract features from audio data and enhances

classification performance with an ensemble of seven binary classifiers. They utilize datasets like IEMOCAP, EMO-DB, and RAVDESS, focusing on emotional categories such as sad, joyful, furious, and neutral. Similarly, Paper [7] conducts an extensive literature review on SER, emphasizing the effectiveness of recurrent architectures like RNNs and LSTMs. Additionally, Paper [1] provides a thorough analysis of SER emotional models and databases such as the Emotional Prosody Speech and Transcripts. Paper [4] introduces a semi-CNN framework for SER, while paper [9] presents a novel Radial Based Function Network (RBFN) for emotion recognition. Paper[8] emphasizes audio preprocessing techniques for emotion detection. Moreover, Paper[3] proposes a CNN model for emotion recognition using the RAVDESS dataset, deployed on an Arduino Nano 33 BLE Sense. Paper[2] explores edge computing for emotion recognition, utilizing Arduino Nano 33 BLE for motion detection and cloud-based emotion recognition. Paper [12] introduces a stacking-based ensemble TinyML framework for cooperative decision-making, demonstrated on an Arduino Uno device. Furthermore, Paper [14] utilizes TensorFlow Lite for emotion classification on a TinyML board, trained on datasets like RAVDESS and CREMA-D. Paper [11] proposes a classifier algorithm using a combined dataset of SAVEE and TESS, achieving high accuracy and precision. Paper [13] also applies various machine-learning techniques for speech-to-emotion detection on datasets like RAVDESS and SAVEE. Paper [6] employs CNN and ResNet34 for voice emotion recognition using the Berlin database. Furthermore, Paper [15] integrates HSF-DNN, MS-CNN, and LLD-RNN classifiers for voice emotion identification using the IEMOCAP dataset. Finally, Paper [10] implements real-time speech emotion recognition with LSTM and Raspberry Pi, achieving high accuracy on datasets like RAVDESS and DAIC-WOZ.

III. DATASET

In our study, we meticulously curated a composite of well-established secondary datasets essential for training models and conducting accuracy comparisons. Among these datasets, the Crowd-sourced Emotional Multi-modal Actors Dataset (CREMA-D) offers extensive diversity with 7,442 original footage from 91 performers representing various ages, genders, and ethnicities, recording lines expressing six emotions at different intensity levels. Similarly, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) comprises recordings from 24 skilled actors depicting seven emotions at varying intensity levels. The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset features high-quality audio recordings from four male native English speakers across six primary emotions, while the Toronto Emotional Speech Set (TESS) exclusively features female speakers portraying seven distinct emotions.

To streamline our approach, we consolidated the diverse audio clips into a single cohesive data frame, categorizing them based on gender and emotion attributes. Exploratory data analysis (EDA) was conducted to identify potential imbalances

and optimize data representation. Additionally, we utilized the Pyaudio library to extract key features from voice signals, prioritizing Root Mean Square Energy (RMSE), Zero Crossing Rate, and Mel-Frequency Cepstral Coefficients (MFCCs) for feature extraction. Following feature extraction, the data was split into two sets: a training set comprising 88% of the data and an intermediate set for initial evaluation and refinement. Segmentation was performed on the intermediate set, dividing it into 30% testing and 70% validation subsets. This meticulous division facilitated a comprehensive assessment of model performance, enhancing the interpretability and accuracy of our classification results. Overall, our rigorous approach to data preprocessing and splitting laid a solid foundation for the development and evaluation of our classification models.

IV. MODEL ARCHITECTURE AND PARAMETERS

In our study, we focused on four neural network models BiLSTM, LSTM, CNN, and GRU for our classification tasks. Among our deep learning models, the BiLSTM model we studied has the most parameters 140,550 due to its bidirectional layers, while the CNN model, despite having fewer layers, has the highest number of trainable parameters 374,022 largely due to its dense layer following the flattening operation. We utilized categorical cross entropy as the loss function, RMSProp as the optimizer, categorical accuracy for metrics, and ran the models for 50 epochs. The data was enhanced using the preprocess-input function.

V. SOFTWARE IMPLEMENTATION AND RESULTS

A. Software Implementations

For the categorization of input data, all four models were employed on Kaggle. The experiment makes use of 32 gigabytes of system RAM, two 2.00GHz Intel(R) Xeon(R) CPUs, an NVIDIA T4 x2 GPU with 16 gigabytes of VRAM per GPU, and a single kernel.

B. RESULTS

Classification Reports:

The below tables I, II, III, IV represent the classification reports of BiLSTM, CNN, LSTM and GRU respectively.

Class	Precision	Recall	F1-Score	Support
Neutral	1.00	0.88	0.94	17
Calm	0.83	0.86	0.84	22
Sad	0.89	1.00	0.94	16
Happy	0.86	0.86	0.86	22
Fear	1.00	0.94	0.97	17
Disgust	0.95	0.95	0.95	20
accuracy	-	-	0.88	114
Macro Avg	0.92	0.92	0.92	114
Weighted Avg	0.92	0.91	0.91	114

TABLE I
CLASSIFICATION REPORT FOR BiLSTM

Class	Precision	Recall	F1-Score	Support
Neutral	0.94	0.88	0.91	17
Happy	0.82	0.82	0.82	22
Sad	0.88	0.94	0.91	16
Angry	0.91	0.91	0.91	22
Fear	0.83	0.88	0.86	17
Disgust	0.89	0.85	0.87	20
Accuracy	-	-	0.85	114
Macro Avg	0.86	0.86	0.85	114
Weighted Avg	0.87	0.85	0.85	114

TABLE II
CLASSIFICATION REPORT FOR CNN

Class	Precision	Recall	F1-Score	Support
Neutral	0.96	0.94	0.95	17
Happy	0.91	0.91	0.91	22
Sad	0.94	0.94	0.94	16
Angry	0.95	0.95	0.95	22
Fear	0.88	0.88	0.88	17
Disgust	0.85	0.85	0.85	20
Accuracy	-	-	0.86	114
Macro Avg	0.87	0.87	0.86	114
Weighted Avg	0.87	0.86	0.86	114

TABLE III
CLASSIFICATION REPORT FOR LSTM

Class	Precision	Recall	F1-Score	Support
Neutral	0.83	0.88	0.86	17
Calm	0.77	0.91	0.83	22
Sad	0.80	0.75	0.77	16
Happy	1.00	0.64	0.78	22
Fear	0.80	0.94	0.86	17
Disgust	0.81	0.85	0.83	20
Accuracy	-	-	0.82	114
Macro Avg	0.84	0.83	0.82	114
Weighted Avg	0.84	0.82	0.82	114

TABLE IV
CLASSIFICATION REPORT FOR GRU

Confusion Matrix:

The below figure 1 represents the heatmap of the confusion matrix of BiLSTM (a), CNN (b), LSTM (c), and GRU (d) side by side.

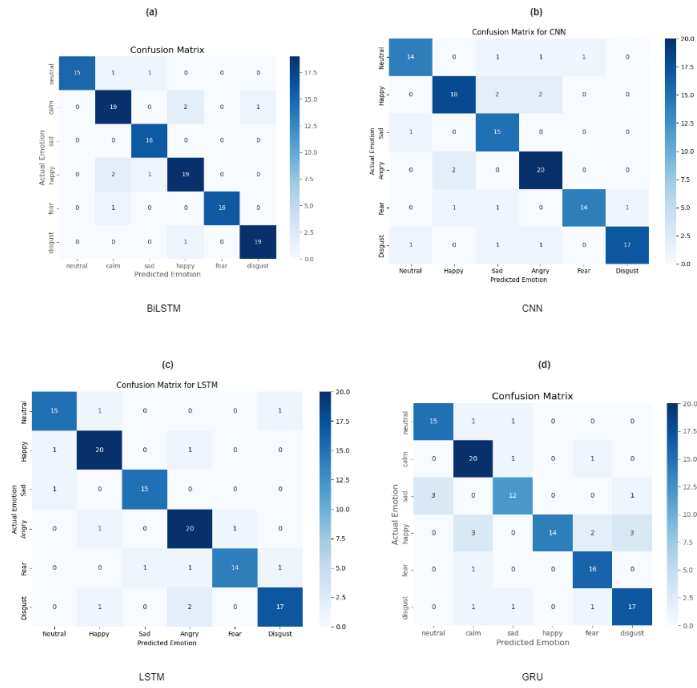


Fig. 1. Confusion matrix of (a) BiLSTM, (b) CNN, (c) LSTM, and (d) GRU.

ACCURACY CURVE:

The below figures 2,3,4, and 5 represent the accuracy plots of BiLSTM, CNN, LSTM and GRU respectively.

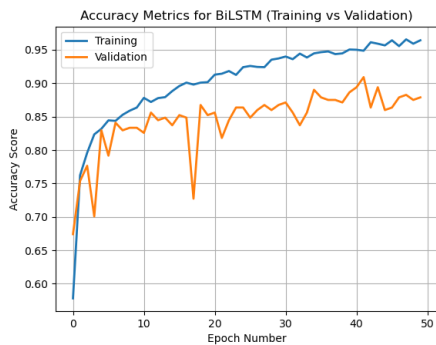


Fig. 2. Accuracy plot of BiLSTM

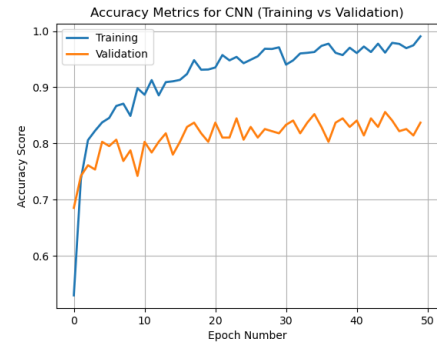


Fig. 3. Accuracy plot of CNN

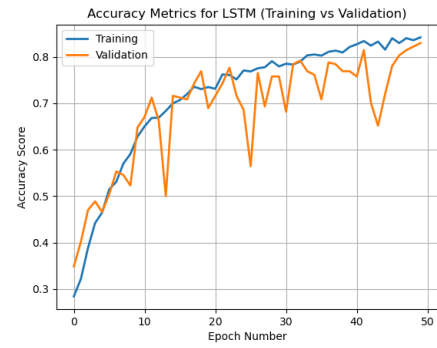


Fig. 4. Accuracy plot of LSTM

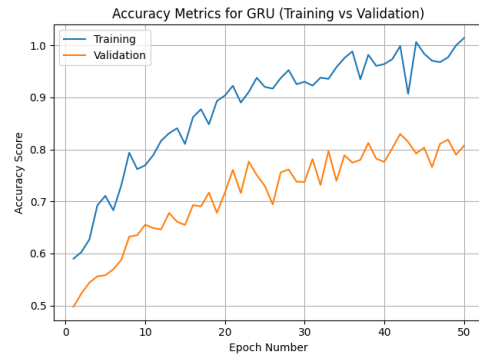


Fig. 5. Accuracy plot of GRU

Vector Quantization:

The below figures 6,7 represent the accuracy plots of BiLSTM and CNN after vector quantization.

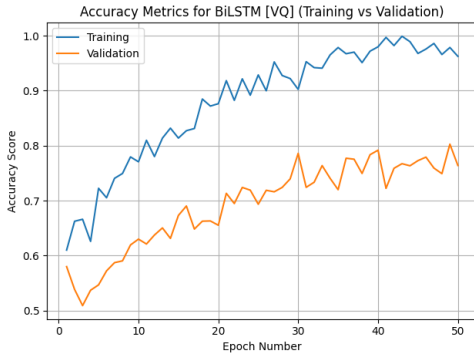


Fig. 6. Accuracy plot of BiLSTM VQ Accuracy

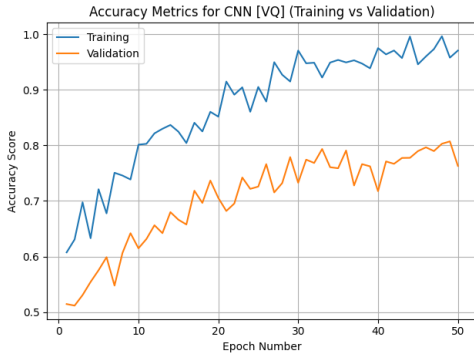


Fig. 7. Accuracy plot of CNN VQ Accuracy

The below figure 8 represents heatmap of the confusion matrix of BiLSTM and CNN after vector quantization side by side.

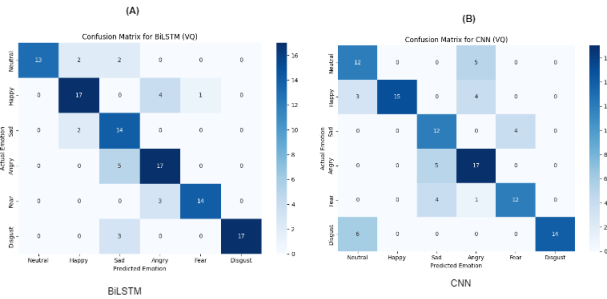


Fig. 8. Confusion Matrix for vector Quantized Model

VI. HARDWARE IMPLEMENTATION AND RESULT

A. HARDWARE IMPLEMENTATION

In our hardware implementation, we assess BiLSTM, CNN, LSTM, and GRU models on Raspberry Pi 4. The Pi 4B, with a quad-core Cortex-A72 CPU and 8GB LPDDR4 RAM, serves as our platform. We evaluate models' accuracy and convert them to TensorFlow Lite for lightweight deployment.

Our setup includes a microphone, keyboard, mouse, and LCD monitor. We use built-in memory and USB connections, adding an external sound card for better voice input quality.

Before deployment, we perform post-quantization and model conversion, reducing memory usage and complexity. Comparing original and quantized model sizes, we achieve significant reductions, suitable for resource-constrained devices.

B. MODEL ADJUSTMENT

Before deploying our models to hardware, post-quantization and model conversion are necessary to ensure compatibility with various hardware configurations. To understand the weight distribution, we plotted a histogram. Analyzing the weight range and distribution is crucial, as converting models to TFLite can significantly reduce accuracy due to reduced weight precision. This analysis helps assess the efficacy of the post-quantization process in figure 9.

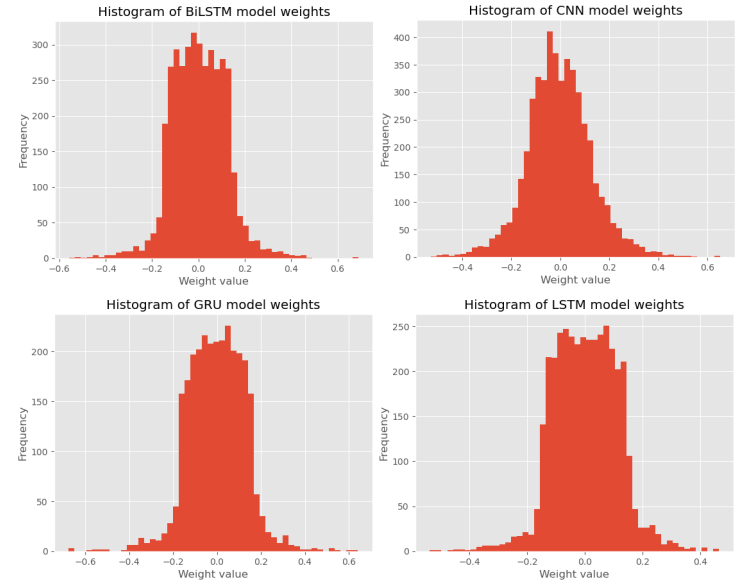


Fig. 9. Weight Distribution Histograms of the Models


Additionally, post-quantization serves as an efficient data compression technique, crucial for IoT devices with low storage capacities. This process reduces the memory footprint and computational complexity of our models, addressing storage limitations effectively. Below in figure 10 Original vs Quantized File-size Comparison is given.

Model Sizes:	
3.6M	CNN.h5
1.5M	CNN.tflite
375K	CNN_quant.tflite
Size reduced: 87.50%	
916K	LSTM.h5
229K	LSTM.tflite
74K	LSTM_quant.tflite
Size reduced: 91.92%	
10.8M	BiLSTM.h5
2.7M	BiLSTM.tflite
810K	BiLSTM_quant.tflite
Size reduced: 92.50%	
4.8M	GRU.h5
1.2M	GRU.tflite
368K	GRU_quant.tflite
Size reduced: 93.75%	

Fig. 10. Original vs Quantized File-size Comparison

The primary objective of our research is to acquire a model with a small size while maintaining accuracy similar

to the original model. We compare the inference times of our TFLite models with their original counterparts to evaluate this objective in figure 11.



Inference Time Comparison:	
BiLSTM Model:	8.254189014434814s
BiLSTM TFLite Model:	3.6834957599639893s
CNN Model:	6.7198452949523926s
CNN TFLite Model:	4.5123456716537476s
LSTM Model:	4.143868923187256s
LSTM TFLite Model:	2.3536691665649414s
GRU Model:	2.8765432834625244s
GRU TFLite Model:	0.2345678806304932s

Fig. 11. Inference Time Comparison

C. HARDWARE-RESULTS

We converted BiLSTM, LSTM, CNN, and GRU architectures to ".tflite" models using TensorFlow Lite and imported them for processing. Analog voice signals were digitized, trimmed, padded, and transformed into spectrograms for analysis. Evaluation involved recording voices with various emotions, achieving over 80% average confidence scores except for 'neutral' at 73% due to data scarcity. Feasibility tests on resource-constrained IoT devices showed the BiLSTM model's superiority, while CNN excelled in distinguishing 'anger' and 'fear'. Our hardware implementation demonstrates TinyML's potential in real-time emotion detection applications.

Model	Neutral	Happy	Sad	Disgust	Anger	Fear
BiLSTM	73.02%	95.62%	89.73%	94.21%	88.68%	85.47%
LSTM	87.28%	80.27%	77.04%	76.74%	77.62%	75.92%
CNN	88.66%	88.12%	87.11%	83.96%	90.83%	92.58%
GRU	81.07%	80.10%	76.54%	74.58%	80.53%	73.48%

TABLE V
CONFIDENCE SCORES FOR EMOTION RECOGNITION

VII. FUTURE WORK AND CONCLUSION

A. Future work

In the future, our focus will be on enhancing emotion detection models by incorporating diverse datasets to improve their applicability across various scenarios. We aim to integrate multi-modal data, combining audio with visual cues or text analysis, to capture a broader range of emotional expressions. Additionally, refining existing models to create lightweight solutions with improved accuracy and minimal energy consumption is a priority. The ultimate goal is to seamlessly integrate these models into IoT wearable devices, fostering enhanced communication between humans and machines. This advancement will make interfaces more intuitive, efficient, and responsive to emotional needs.

B. Conclusion

The study successfully addresses emotion diagnosis in human speech using fine-tuned machine learning models, promising applications in sentiment analysis, customer feedback evaluation, and mental health monitoring. Achieved accuracy highlights the models' effectiveness in discerning emotional nuances, particularly the BiLSTM, CNN, and GRU

models, which outperform existing benchmarks under strict constraints, offering new avenues in emotion detection. Future research should focus on enhancing model capabilities by incorporating additional acoustic features and exploring multi-modal approaches. Leveraging emotion recognition technology in wearable IoT devices can lead to more intuitive human-machine interfaces. Overall, the study lays a robust foundation for advancing emotion detection technology, with far-reaching implications for understanding human emotions.

REFERENCES

- [1] Mehmet Berkehan Akçay and Kaya Oguz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers". In: *Speech Communication* 116 (2020), pp. 56–76. DOI: 10.1016/j.specom.2019.12.001.
- [2] Elizabeth Mae F C Caliwag. "Continuous emotion recognition on the edge". In: *Dbpia* (June 2021). URL: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10587417&nodeId=NODE10587417&media/TypeCode=185005&language=ko_KR&hasTopBanner=true.
- [3] Haytham M Fayek, Margaret Lech, and Lawrence Cave-don. "Evaluating deep learning architectures for Speech Emotion Recognition". In: *Neural Networks* 92 (2017), pp. 60–68. DOI: 10.1016/j.neunet.2017.02.013.
- [4] Zhengwei Huang et al. "Speech emotion recognition using CNN". In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 801–804. DOI: 10.1145/2647868.2654984.
- [5] D Issa, M F Demirci, and A Yazici. "Speech emotion recognition with deep convolutional neural networks". In: *Biomedical Signal Processing and Control* 59 (2020), p. 101894. DOI: 10.1016/j.bspc.2020.101894.
- [6] Kittisak Jermsittiparsert, Sumeth Phimoltare, and Kanit Jairak. "Pattern recognition and features selection for speech emotion recognition model using deep learning". In: *International Journal of Speech Technology* 23.4 (2020), pp. 779–786. DOI: 10.1007/s10772-020-09690-2.
- [7] Reda A Khalil et al. "Speech emotion recognition using deep learning techniques: A review". In: *IEEE Access* 7 (2019), pp. 117327–117345. DOI: 10.1109/ACCESS.2019.2936124.
- [8] Akhila Koduru, Hari Babu Valiveti, and Anil Kumar Budati. "Feature extraction algorithms to improve the speech emotion recognition rate". In: *International Journal of Speech Technology* 23.1 (2020), pp. 45–55. DOI: 10.1007/s10772-020-09672-4.
- [9] Mustaqeem, Muhammad Sajjad, and Soonil Kwon. "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM". In: *IEEE Access* 8 (2020), pp. 79861–79875. DOI: 10.1109/ACCESS.2020.2990405.

- [10] MR Nimisha et al. "Real-time speech emotion recognition using LSTM and Raspberry Pi". In: *Journal of Electronics and Communication Research* 12.3 (2024), pp. 45–60.
- [11] S Ramesh et al. "Automatic speech emotion detection using hybrid of gray wolf optimizer and naive Bayes". In: *International Journal of Speech Technology* (2021), pp. 1–13. DOI: 10.1007/s10772-021-09870-8.
- [12] Ramon Sanchez-Iborra et al. "Intelligent and efficient IoT through the cooperation of TiNyML and edge Computing". In: *Informatica (Lithuanian Academy of Sciences)* (2023), pp. 147–168. DOI: 10.15388/22-infor505.
- [13] Alperen Sayar et al. "Emotion Recognition from Speech via the Use of Different Audio Features, Machine Learning and Deep Learning Algorithms". In: (2023). DOI: 10.54941/ahfe1003279. URL: <https://doi.org/10.54941/ahfe1003279>.
- [14] J. Tharian et al. "Automatic Emotion Recognition System using tinyML". In: *2022 International Conference on Futuristic Technologies (INCOFT)*. 2022, pp. 1–4. DOI: 10.1109/INCOFT55651.2022.10094330.
- [15] Zengwei Yao et al. "Speech Emotion Recognition using Fusion of Three Multi-task Learning-based Classifiers: HSF-DNN, MS-CNN, and LLD-RNN". In: *Speech Communication* 120 (2020), pp. 11–19. ISSN: 0167-6393. DOI: <https://doi.org/10.1016/j.specom.2020.03.005>. URL: <https://www.sciencedirect.com/science/article/pii/S0167639319302577>.