Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 13, No.2 pp186–197April-Jun,2025

p-ISSN: 2663-628X
e-ISSN: 2663-6298

AN IMPROVED DEEP LEARNING TECHNIQUE FOR SPEECH EMOTION RECOGNITION AMONG HAUSA SPEAKERS

Martins E. Irhebhude[1**1], Adeola O. Kolawole[2], Mujtaba K. Tahir[3]
[123]Department of Computer Science, Nigerian Defence Academy, Kaduna

Corresponding author email: mirhebhude@nda.edu.ng

Received: 17 Nov 2024 / Accepted:17 Feb., 2025/ Published: 10Apr., 2025. https://doi.org/10.25271/sjuoz.2025.13.2.1433

ABSTRACT:

This research addressed the challenge of recognizing emotions from speech by developing a deep learning-based speech-emotion recognition (SER) system. A key focus of the study is the creation of a new Hausa emotional speech dataset, aimed at addressing the linguistic and cultural imbalance in existing SER datasets, which predominantly feature Western languages. This study captured four emotions: happy, sad, angry, and surprise among native Hausa speakers. The self-captured dataset was recorded in an environment that is devoid of noise to ensure high quality and uniformity in the audio data. A public dataset; RAVDESS was used for benchmarking the proposed technique. CNN and Bi-Directional Long Short-Term Memory (BiLSTM) architectures were combined and used as proposed model for the SER experiment. The developed CNN architecture helped in extracting spatial features, while the BiLSTM without the attention mechanism captured temporal dependencies from the audio data. The approach reduced time complexity and improved performance to 100% and 96% recognition accuracies against 94% and 90% of the benchmark model for the local and benchmark datasets respectively. The results demonstrate the proposed approach's robustness to generalize across linguistic contexts.

KEYWORDS: Emotion Recognition, Hausa Audio Data, Bi-Directional Long-Term Short-Term Memory Network, BILSTM, Cross-Cultural Variations.

1. INTRODUCTION

The task of recognizing emotions in audio data is one of the key components of improving human-computer interaction (HCI) (Selvaraj *et al*., 2016). With the increase in smart gadgets like Apple HomePod and Amazon Echo, voice interaction has become a key component of recent technologies. This innovation has led to a vast amount of audio data that can be further analyzed. There is a need to improve recognition of speech emotion because of the increased reliance on verbal communication between humans and technology (Duttaa *et al*., 2023). Speech Emotion Recognition (SER) has become very important part of these systems by allowing machines to recognise human emotions, improve communication and offer more individualised smiles. This development showcases the increasing demand for technology, that can understand the emotional states that underlie human speech in addition to recognising it (Hossain & Muhammad, 2018). An example is the identifying of driving stress of fatigue by SER in automotive systems, thereby enhancing safety. When this emotional distress is being identified, healthcare industry can easily track mental health. In addition to this, emotion-aware systems in gaming and education advance, their interactions according to user emotions, to produce more engaging settings (Vaaras *et al*., 2023). Although these developments are being put in place, creating trustworthy SER systems are proving difficult since speech analysis of human emotions is complicated (Goncalves *et al*., 2023).

On of the main issue in SER is speech length variability which features extraction and model training. To guarantee fixed input sizes for deep learning models, traditional approaches which entails clipping or padding voice data is being used (Rasheed *et al*., 2024). This procedure however runs the risk of increasing computing costs or losing important emotional information if something goes wrong along the way both of which impair system performance. One other challenge is the local vs global feature conundrum, methods that can be used are the conventional approaches, which include identifying local features including minute variations in tone or pitch, but they find it difficult to capture long-range contextual dependencies over whole speech sequences because longer speeches may hide certain emotional information (Duttaa *et al*., 2023). The development of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) helped eliminate the restrictions on longer speeches. CNNs and RNNs has shown their strength to reconciling the intricacies in audio data (Rasheed *et al*., 2024). These algorithms are perfect when it comes to identifying complicated patterns in a large amount of data. This makes them ideal for examining human speech for emotional clues (Sundar *et al*., 2022). CNN, RNN, hybrid models are prominent architectures that have shown good results for matrix factorization (Hassan, 2025), COVID-19 cases confirmation (Hassan & Ahmed, 2023), web vulnerabilities detection (Ali *et al*., 2025), sign language recognition (Ahmed *et al*., 2024), (Irhebhude *et al*., 2024), and emotion recognition using facial images (Irhebhude *et al*., 2023), (Hasan, 2022) due to their capacity to extract robust information; CNNs also be used for analysing speech signals. CNNs can capture changes in spoken expressions by identifying patterns in the audio frequency that carry emotional information (Kim *et al*., 2015). CNNs can process spectrograms or Mel-Frequency Cepstral Coefficients (MFCCs) in a frequency-based manner, to bring out significant features that aid in emotion recognition. RNNs on the other hand, can be used to analyze speech sequential data because emotions are often expressed temporally (Dixit & Satapathy, 2023). LSTM networks can accurately capture emotional dynamics in speech by recording patterns and dependencies often expressed over time (Jeong *et al*., 2023). LSTMs capture emotional expressions over time by connecting the short-term and long-term dependency in speech data. Ensemble models often outperform other single models in tasks involving emotion recognition

* Corresponding author

(Vaaras *et al*., 2023). CNNs which extract spatial information and RNNs which capture temporal dependencies can be combined and used as ensemble model for emotion identification systems (Tarunika *et al.,* 2018).

Studies have shown that CNNs and LSTM can be combined in hybrid learning architectures. The kind of information captured by CNN and LSTM can be integrated and used for emotion recognition tasks (Sundar *et al*., 2022).

Rasheed *et al*. (2024) used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, a well-known English emotional speech dataset with eight emotion classes—Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, and Surprise—to group emotional data in speech made by human, thereby combining CNN for feature extraction with Bi-Directional Long Short Term Memory (BiLSTM) networks with attention mechanisms. An average of 94% recognition accuracy was achieved. However, the implementation of these methods resulted in additional computational burden, thereby rendering these models less appropriate for implementation on devices with low resources (Abbaschian *et al*., 2021). A 6% error margin can be further decreased in order to allow for greater recognition accuracy. Datasets like IEMOCAP and RAVDESS are being relied on significantly for SER studies which include Western languages and emotional expressions (Barazida, 2021). Hausa which is one of the underrepresented languages is being kept at a disadvantage because of linguistics bias limitation of the generalizability of these systems across culturally diverse population (Saunders, 2023).

Speech Emotion Recognition (SER) has emerged as a crucial component of human-computer interaction (HCI), mental health monitoring, and affective computing (Selvaraj, *et al*., 2016). Despite its growing importance, most SER research has been centered on Western languages, leaving underrepresented linguistic groups, such as Hausa speakers, without adequate representation. This imbalance raises concerns about the generalizability of existing models, as they may struggle to interpret emotional expressions in different cultural and linguistic contexts (Barazida, 2021).

Traditionally, SER studies have often relied on datasets like the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018), the Toronto Emotional Speech Set (TESS), and the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. While these datasets have been instrumental in advancing SER research, they primarily feature Western speakers, limiting their effectiveness for non-Western languages (Barazida, 2021). Researchers such as Abbaschian *et al*., (2021) and Lucas *et al*., (2023) have highlighted the risk of misclassifications when these models encounter speakers from different cultural backgrounds, emphasizing the need for more diverse datasets.

Some efforts have been made to address this limitation by creating datasets for underrepresented languages. The Berlin Database of Emotional Speech (EMO-DB) and the Arabic Speech Emotion Dataset (ASED) have contributed to expanding SER research beyond English and European languages. However, African languages, particularly Hausa, remain largely unexplored in the context of SER. While Ryumina *et al*., (2022) acknowledged the significance of linguistic diversity in emotion recognition, existing resources for African languages are limited, hindering the development of effective models.

Previous research efforts, such as (Ibrahim, 2021) have focused on speech recognition for Hausa speakers rather than emotion detection, which primarily focus on transcribing spoken language rather than identifying emotional cues. Speech recognition datasets capture linguistic content without emphasizing the prosodic features such as pitch, tone, or rhythm, essential for emotion recognition, highlighting the absence of dedicated emotional speech datasets.

Recognizing these gaps, this study takes a significant step toward bridging the divide by developing a Hausa Emotional Speech Dataset tailored for SER. Unlike previous Hausa speech datasets, which primarily address phoneme recognition, this dataset is specifically designed to capture emotional expressions. It is also benchmarked against the widely-used RAVDESS dataset, ensuring a comparative analysis of its effectiveness. By integrating Convolutional Neural Networks (CNN) and Bi-Directional Long Short-Term Memory (BiLSTM) networks, this research enhances the recognition of emotional nuances in Hausa while optimizing computational efficiency by omitting attention mechanisms.

There are still a number of significant shortcomings even with the advancements in deep learning based SER. The absence of culturally inclusive datasets that faithfully capture emotional expressions from underrepresented language has limited the capacity of existing models to generalise across various linguistic contexts. The robustness of SER system is diminished when applied to non-western languages because the majority of existing datasets like the RAVDESS dataset, and the Acted Emotional Speech Dynamic Database (AESDD) were designed with western speakers in mind (Barazida, 2021). Goncalves *et al*. (2023) and Abbaschian *et al*. (2021) noted that the lack of culturally diverse datasets causes biases in emotion detection algorithms, this however, lowers the systems accuracy when they come across storage accents or language. Secondly, attention technique add some computational costs and reduce the models suitability for real-time applications on devices with constrained processing capacity even while they enhance model performance by selectively focusing on important portions of input sequences. This research is focused on two main objectives to address these challenges: (1) The development of a human emotional speech dataset, which fills a critical gap in the SER literature by providing a resource for studying emotional expressions in an underrepresented African language (Saunders, 2023); and (2) the design of a lightweight CNN-BiLSTM model that achieves high accuracy while eliminating the need for attention mechanisms. The proposed model offers a practical solution for real-world applications by balancing accuracy and efficiency with limited computational resources (Abbaschian *et al*., 2021).

The rest of this paper is structured as follows. Section 2 shows the methodology, which involves data collection and model building. Section 3 presents the results of the experiment conducted. The final section concludes and presents future research plans.

SER has a lot of interest from many quarters to improve human-computer interaction (HCI) systems. Studies have shown efforts made using deep learning architectures to enhance the performance of emotion recognition systems. Deep learning architectures eliminate the need for handcrafting of features by automatically learning from speech data. This is the advantage CNN has over the traditional machine-learning approach to classification (Sundar *et al*., 2022). The combination of CNN and LSTM models was motivated to address the challenge of balancing local feature extraction with contextual awareness, among other difficulties (Goncalves *et al*., 2023).

Rasheed *et al.* (2024) in their study trained a hybrid deep learning model that combined CNNs and BiLSTM networks with attention mechanisms for the classification of four emotions using the RAVDESS dataset. The study showed the effectiveness of combining CNNs and BiLSTM, in handling sequential data-related problems by recording an accuracy of 94%. However, the approach has issues such as the complexity of the attention mechanism; despite the excellent accuracy. The raise in the computing requirements made it less appropriate for deployment on low-resource devices or real-time applications. Risks of overfitting; when trained on smaller or less varied datasets, the complex architecture, which includes the attention mechanism-presented a greater danger of overfitting. The model's efficacy is diminished because it was trained on datasets like RAVDESS which mostly reflect Western languages. The lack of application

of the model of different datasets makes it difficult to generalize to other populations.

SaiDhruv *et al*. (2023) recorded a high accuracy of 96.73% SER task using the LSTM model. The study showed the impact of adjusting the model's parameters and how overfitting can be avoided as data complexity rises. The study demonstrated the dominance deep learning has over traditional machine learning in emotion recognition tasks.

Jeong *et al.* (2023) in their study suggested a method for recognizing emotion in conversation data. The approach works by integrating text and audio data using a multimodal fusion network with a pre-trained language model. The strength of the multimodal techniques in capturing a variety of emotional cues was demonstrated when it achieved a new state-of-the-art performance on the KEMD,20, and MELD datasets (Middya *et al*., 2022).

Schonevelda *et al.* (2020) approach to the audio-visual emotion recognition (AVER) task was tested on the RECOLA data (audiovisual recordings of French speakers in a range of emotional states). The proposed approach outperformed the selected state-of-the-art techniques. The suggested approach trained distinct audio and visual deep convolutional neural networks (DCNN) using a pre-trained network. These were used to extract generic emotion identification charactertistics, which were used to train a fusion module. The authors also differentiated the strategy with a number of cutting-edge techniques that employ features derived from deep neural networks or finding, the AVER predicts valence on the RECOLA dataset better than cutting-edge methods. A significant improvement of the concordance correlation coefficient (CCC) for valence prediction as reported by the authors is 0.665 as against the previous state-of-the-art CCC of 0.616.

In order to accomplish precise and almost real-time recognition of emotional states in speech, Abbaschian *et al*. (2021) used deep learning approaches to handle the SER challenge, which surrounds anger, happiness, neutral, disgust, and surprise in HCI. One of the methodologies employed was a thorough analysis of the most recent studies or deep learning strategies for SER. The authors carried act an extensive search of pertinent literature using a variety of scholarly databases and search engines, including google scholar, IEEE Xplore and the ACM Digital Library. Relevant material was chosen and examined based on particular inclusion and exclusion criteria, such as the application of deep learning methods for SER. The accessibility of training and testing datasets was taken into account following the models accuracy documentation. According to the findings, deep learning approaches such as CNNs, RNNs and their variations have outperformed conventional machine learning techniques and demonstrated encouraging outcomes for SER. The quality and availability of datasets for testing and training are essential for the SER models to achieve high accuracy and generalisation. The performance and generalisation of the models may be affected by the present SER database's short size, lack of diversity, and lack of standardisation, among other issues. This performance of transfer learning and data augmentation approaches, which also assisted in overcoming some of the restrictions of the existing SER databases. Some of the future research directions in SER are the creation of more varied and standardised databases, the analysis of multimodal strategies and the study of explainable and interpretable deep learning models.

The application of deep learning methods to enhance emotion recognition across many modalities was covered by Njoku *et al.* (2021). The aim of the study was to compare how well those methods performed in identifying emotions (happy, sadness, surprise, disgust and neutral) from multimodel data. The authors of multimodal fusion and classification investigated three distinct deep learning models in early fusion (EF), hybrid fusion (HF) and multitask learning (MTL). These three models are applied to three distinct modalities. Speech, facial expression and EEG data. The authors compared the model's performance using two dataset, the RAVDESS audiovisual dataset and an EEG dataset for a prior study. According to he study's results, the EFG model performed the best in the audiovisual data, scoring 78.33%, HF and MTL performed worse, with a score of 57.91% and 55.41% respectively. The MTL model had the best accuracy while the HF model recorded the lowest performance for all modalities, all things considered.

Zaman *et al*. (2023) reviewed methods and strategies for employing deep learning models to categorise and identify emotions including sorrow, happiness, neutral, disgust, and surprise in audio datasets. In order to outline the future paths of deep learning in audio classification, additional study of widely used audio dataset was conducted. The study however included a thorough overview of the most recent methods and developments in audio categorisation using deep learning models. To this effect, the findings from the study are as follows: (1) Deep learning models, including CNNs, RNNs, autoencoders, transformers, and hybrid models have greater promise for audio classification tasks. (2) while RNNs are suitable for tasks that require sequential modelling of audio data, CNNs are effective for audio classification tasks that involve spectrogram or mel-spectrogram data. (3) Autoencoder can be used for unsupervised feature learning and dimensionality reduction in audio classification tasks. 4. Transformers are a relatively new deep-learning architecture that has shown promising results in audio classification tasks, and more transformer-based methods are expected to be proposed in the future. 5. Hybrid models that combine different deep learning architectures can achieve better performance than single-models. 6. The choice of dataset is crucial for the performance of deep learning models in audio classification tasks, and there are several commonly used datasets in this field. The authors outlined future directions of deep learning in audio classification to include the use of more complex deep learning architectures, the development of more efficient training algorithms, and the exploration of new audio datasets and applications.

Several gaps remain despite the successes recorded in speech emotion recognition studies. The dearth of diverse datasets in culturally underrepresented languages like Hausa is the first gap. This creates biases in SER models leading to reduced performance a model is supposed to generalize (Abbaschian *et al*., 2021). Another gap lies in the reliance on attention mechanisms in a recent study (Rasheed *et al*., 2024); an approach that introduced computational costs and models overfitting. The need for lightweight architectures that can generalize well and yield high accuracy cannot be overemphasized. Hence, this study addressed the identified gaps by capturing local data on Hausa emotional speech and proposed a modified CNN-BiLSTM model (Rasheed *et al*., 2024) that eliminated the attention mechanisms. The captured dataset provided a valuable resource for studying emotional expressions in a different language. The proposed model offered a solution for real-time applications that maintain high recognition accuracy with reduced computational demand.

2. METHODOLOGY

The approach effectively captures spatial and temporal dependencies within speech signals, guaranteeing accurate emotion classification devoid of computation complexity.

The methodology organized into several stages is comprised of: data acquisition, data preprocessing, model; design, training, and evaluation. Each stage is critical to the model's overall performance and is depicted in the methodology flow diagram in Figure 2.1. A major aspect of this research is comparing the proposed model with existing methods, particularly the work of (Rasheed *et al*., 2024). While Rasheed *et al*. (2024) employed an attention mechanism to enhance their model's performance, this study demonstrates that similar levels of accuracy can be

achieved without the added complexity. By simplifying the model architecture and eliminating the attention mechanism, the proposed approach reduced computational demands. It improves the model's suitability for deployment in real-world scenarios, especially where resources are limited.

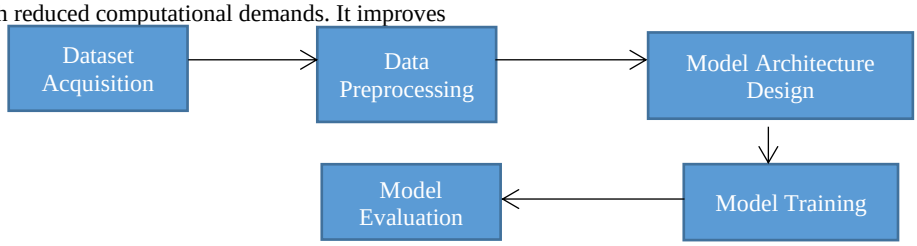


Figure 2.1: Proposed Methodology

Data Acquisition

The datasets used in this study consist of self-captured audio recordings of speech emotions in the Hausa language. The development of the Hausa emotional dataset was necessary due to the absence of any existing emotional speech datasets specific to the Hausa speakers. Although prior research efforts (Ibrahim, 2021) created Hausa speech recognition datasets, with a primary focus on transcribing spoken language rather than identifying emotional cues. Speech recognition datasets capture linguistic content without emphasizing the prosodic features such as pitch, tone, or rhythm, essential for emotion recognition. As a result, existing Hausa Speech datasets (Ibrahim, 2021) are unsuitable for this study, which aims to explore emotional nuances through speech. This limitation motivated the development of a specialized Hausa emotional dataset.

Dataset Description

The dataset used in this study was developed to address the lack of Hausa emotional speech data. It comprises 4,000 speech samples, each lasting approximately five seconds, collected from 1,000 native Hausa speakers. The dataset includes four primary emotion categories: happiness, sadness, anger, and surprise The selection of four emotional categories happiness, sadness, anger, and surprise was guided by their universal relevance and applicability in emotion recognition research as shown in Table 2.1. Pei *et al*., (2024) highlighted these emotions as part of the "basic emotions" framework, which posits that certain emotional expressions are universally recognized across cultures. For this study, these four emotions were chosen because they represent distinct and easily distinguishable emotional states in speech. This simplifies annotation and classification tasks while ensuring that the dataset remains manageable for model training and evaluation.

Participant Demographics

Participants were selected from Hausa-speaking communities in Nigeria, specifically from from Hausa Communities in Gaya, Dala and Rano Local Governments of Kano State, Nigeria. The sample consisted of a diverse group of individuals spanning different age groups, ranging from 15 to 56 years old. The age range of 15 to 56 years was selected to ensure diversity in vocal characteristics and emotional expressions across different life stages. This range captures both younger and middle-aged individuals, who often exhibit clear and dynamic emotional expressions. Both male and female speakers were included to ensure gender diversity and to avoid potential biases arising from gendered speech patterns. This broad demographic representation enhances the dataset's applicability across different age groups and genders.

Recording Conditions

Recordings were conducted in acoustically treated rooms to minimize background noise and interference. Participants were instructed to deliver their emotional expressions based on scripted scenarios and spontaneous speech using a sentence *"To shikenan, nagode",* ensuring that the dataset contained both natural and acted emotional speech samples.

Potential Biases in the Dataset

Despite efforts to ensure diversity, certain biases may exist within the dataset. The dataset primarily drawn participants from both urban and urban areas of Kano State, potentially limiting its applicability to Hausa speakers in other States of Nigeria and Africa (Saunders, 2023), who may have different speech patterns or intonations. Additionally, while the dataset captures four primary emotions, it does not account for more nuanced emotional states such as fear, disgust, or neutrality. Future extensions of the dataset could address these gaps by incorporating a broader range of emotions and recruiting participants from a wider geographic area.

Table 2.1: Classification of Hausa Datasets

S/N	Emotion Class	No. Of Audio Samples
1	Happy	1000
2	Angry	1000
3	Sad	1000
4	Surprise	1000
	TOTAL	**4000**

Data Preprocessing

Data preprocessing is a fundamental step that ensures audio data are transformed into a clean, normalized, and structured form that is suitable for machine learning experiments. Speech data sometimes contains silent intervals, noise, and variations in amplitude. The presence of these imperfections can degrade a model's performance if not handled properly.

Each preprocessing activity ensures that the inputted audio data to the CNN-BiLSTM model is rich in emotion-relevant features. The preprocessing activities ensure that the raw audio data is cleaned, and transformed into a form that is suitable for feature extraction.

Silence Detection and Removal

Human speech sometimes contains pauses or silent intervals between words or phrases; these segments do not carry useful emotional information. Such intervals could lead to inaccurate feature extraction and increased training time (Calzone, 2022). To address this issue, this study employed the Praat Silence Detection Tool (Al-Tamimi, 2022). The technique analyzed the intensity contour in the audio signal to detect and remove silent intervals. The intensity contour $I(t)$ calculated as shown in equation 1 measures the short-term energy of the signal:

$$I(t) = 10\ log_{10}\ \frac{1}{N}\sum_{i=0}^{N-1}(x_i)\ t^2 \qquad (1)$$

where N is the number of contours in each frame, and $(x_i)\ t^2$ represents individual contours within the frame. If the intensity falls below a specified threshold (e.g., −30dB), the corresponding segment will be marked as silence. A minimum quiet interval of 50ms ensures that only meaningful pauses are treated as silence, while a minimum sounding interval of 100ms ensures that short bursts of noise are not misclassified as speech. The result of this

process is a trimmed audio signal containing only relevant speech segments. This step not only improves the signal-to-noise ratio but also reduces computational costs by eliminating non-informative data, making the training process more efficient (Tzirakis *et al*., 2017).

Noise Reduction

Environmental noise is another major factor that can degrade the quality of speech signals. To mitigate this, a combination of bandpass filtering and Wiener filtering was applied. The bandpass filter removes high-frequency noise *H(f)* offsets from the signal by restricting the frequency range to 100 Hz – 8 kHz, the range most relevant to human speech (Gong *et al*., 2021).

The response of the bandpass filter is defined as (equation 2):

$$H(f) = \frac{f}{f + f_{cutoff}} \quad for\ f \in [f_{low} \quad f_{high}] \qquad (2)$$

where f_{low} = 100Hz and f_{high} = 8kHz. This step eliminates frequencies outside the range of normal speech, preserving only the meaningful portions of the signal (Gong *et al*., 2021).
Next, the Wiener filter is applied to further suppress background noise by modeling it from silent segments identified earlier. The Wiener filter estimates the noise spectrum $N(f)$ and subtracts it from the signal spectrum $S(f)$ in (equation 3) as follows:

$$H(f) = \frac{S(f)}{S(f) + N(f)} \qquad (3)$$

Normalization

This adaptive filtering technique ensures that only relevant speech information is retained while minimizing distortions (equation 4).

$$Y_{norm}(t) = \frac{\max(Y) - \min(Y)}{Y(t) - \min(Y)} \qquad (4)$$

where Y(t) is the original signal amplitude, and min(Y) and max(Y) are the minimum and maximum amplitude values, respectively. This step guarantees that the dynamic range of the input signal is uniform across all recordings Gong *et al*. (2021).

Frame Segmentation and Hamming Windowing

Since Speech Is A Non-Stationary Signal, Its properties change over time. To capture these temporal variations, the audio was divided into overlapping frames. Each frame is 50ms long with a 10ms overlap to ensure that transitions between frames are preserved, avoiding the loss of emotional cues at the frame boundaries.

A Hamming window function is applied to each frame to reduce spectral leakage and improve feature extraction. The Hamming window is defined in equation 5 as:

$$w(n) = 0.54 - 0.46\left(\frac{2\pi n}{2N-1}\right) \qquad (5)$$

where N is the frame length, and (n) is the sample index within the frame. The windowed signal is then computed as (equation 6):

$$Y_{windowed}(n) = Y_{frame}(n) \cdot w(n) \qquad (6)$$

This step ensures that the features extracted from each frame are representative of the underlying emotional content (Gong *et al*., 2021).

This detailed and structured preprocessing ensured that the subsequent CNN-BiLSTM model focused on meaningful features, ultimately improving the performance of the emotion recognition system.

Proposed Model Architecture

The core of the model architecture adapted from Rasheed *et al*. (2024) is the integration of cnns with bilstm networks without attention mechanism (as depicted in Figure 2.2 & Table 2.2). The model is designed to reduce complexity and computational overhead, thereby making the model more suitable for real-time applications and deployment on resource-constrained devices. The exclusion of the attention mechanism also simplified the training process and reduced the risk of overfitting, particularly on smaller datasets.

Figure 2.2: Proposed Model Architecture

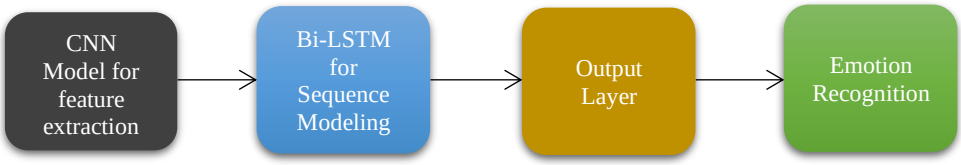


Table 2.2: Network Architecture of the Model

S/N	Layer Type	Layer Name	No. Of Neurons	Activation Function
1	Convolutional Layer	2D	32 x 3	Relu
2	Convolutional Layer	2D	64 x 3	Relu
3	Pooling Layer	Maxpooling2D	-	-
4	Dropout	Dropout	07	-
5	Flattening Layer	Flatten	-	-
6	Fully Connected Layer	Dense	128	Relu
7	Dropout	Dropout	0.5	-
8	Reshaping Layer	Reshape	128 x 1	-
9	Bi – Directional LSTM Layer	Bidirectional LSTM	64	-
10	Bi – Directional LSTM Layer	Bidirectional LSTM	64	-
11	Output Layer	Dense	-	Softmax

CNN Model For Feature Extraction

The CNN plays a pivotal role in this study by extracting emotion-relevant features from mel-spectrograms and MFCCs 2D representations of speech signals. These features capture both spectral and temporal variations that correlate with different emotions. For example, happiness may manifest as a higher pitch and energy, while sadness is typically associated with a lower pitch and slower rhythm (Ristea *et al*., 2022). CNNs excel at detecting such localized patterns from structured data like spectrograms, making them an ideal choice for automated feature extraction in SER (Ristea & Ionescu, 2021). In this study, mel-spectrograms serve as input to the CNN model. The spectrograms translate the voice signal to time-frequency domain, where the frequency is on the vertical axis while time is on the horizontal axis. The energy is reflected by the spectrogram pixel, providing a rich input data that records tones, rhythm, and prosody; essential indicators for identifying emotions. (Khalil *et al*., 2018). CNN scans the spectrogram using convolutional filters to find significant patterns, like format shifts, pitch variations or energy changes, which are the signs of a particular emotion. The CNN structure guarantees that low-level characteristics are recorded in the network's early layers while higher-level patterns appear at a deeper level. (Dixit & Satapathy, 2023).

The CNN model used in this study consists of two convolutional layers, followed by ReLU activation, batch normalization, and max-pooling layers. Each layer learns features from the spectrogram input, ensuring the model detects patterns that contribute to emotional expression.

The convolution layer applies a filter (or kernel) over small regions of the inputted spectrogram, producing a feature map that represents the important local features as defined in equation 7:

$$Y\ (i,j)\ = \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} \mathrm{I}\ (\mathrm{i}+\mathrm{p},\mathrm{j}+\mathrm{q})\ .\ K(p,q) \qquad (7)$$

where: I is the inputted spectrogram, K is the filter of size 3×3, (i,j) are the coordinate positive on the output feature map.

Equation 7 describes how the 3×3 filter slides over the inputted spectrogram, performing element-wise multiplication. The filter is designed to detect specific patterns like energy bursts across frequency bands or pitch changes, where high activation values indicate the presence of the pattern (Dixit & Satapathy, 2023). In this study, the first convolutional layer captures the basic frequency peaks and energy shifts within the short time frames; on the other hand, the deeper layers detect and capture temporal patterns such as the rise and fall in pitch over a longer period. After the convolutional operations, the output is passed through a Rectified Linear Unit (ReLU) activation function. ReLU allows the CNN to learn more complex relationships between input features (Njoku *et al*., 2021). The ReLU function is defined as: (equation 8)

$$ReLU(x)\ =\ max(0,x) \qquad (8)$$

The function in equation 8 ensures that only positive activations are passed to the next layer, allowing the network to focus on important patterns ignoring irrelevant noise.

The next step is batch normalization. This is applied to standardize the extracted feature map by ensuring zero mean and unit variance across the mini-batch. The normalization helps to stabilize the training process and prevents overfitting. This process ensures that the model generalizes well on unseen data (Dixit & Satapathy, 2023).

The max-pooling follows the activation and normalization steps to reduce its spatial dimensions. The max-pooling layer selects the maximum value within a region of 2×2 within the feature map as shown in equation 9:

$$P(i,j) = max\{Y(i+p,j+q)|p,q \in [0,1]\} \qquad (9)$$

Max-pooling achieve two objectives: (1) Dimensionality reduction: It reduces the size of the feature map, making the model more computationally efficient. (2) Salient feature retention: It preserves the most important activations, ensuring that critical patterns (such as prominent pitch or energy shifts) are not lost. The combination of convolution, activation, and pooling ensures that the CNN can capture localized emotional cues across both time and frequency (Dixit & Satapathy, 2023).

After max-pooling, the reduced feature maps are flattened into a 1D vector to prepare them for the (dense) layers. Flattening converts the multidimensional feature maps into a single vector: The flattening operation converts the 2D feature map Y_3 into a 1D vector z (equation 10):

$$z\ = \mathrm{Flatten}\ (Y_3) \qquad (10)$$

This vector z serves as the input to the dense layers, which map the extracted features to the corresponding emotion classes. The dense layer assigns weights to each feature, learning how different patterns in the audio spectrogram correlate with specific emotions (Rasheed *et al*., 2024).

BILSTM For Sequence Modeling and Classification

While CNNs extract local acoustic features from individual frames of the input spectrogram, these features alone are insufficient for accurate emotion recognition. Emotional meaning is conveyed over time, requiring a model that can learn and retain sequential patterns (Li *et al*., 2020). The BiLSTM captures both forward and backward dependencies, making it ideal for tasks where context evolves across time like speech emotion recognition. For instance, emotional tone might change across a sentence ("*To shikenan, nagode*"), and BiLSTMs can interpret both past and future cues to determine the overall emotion accurately.

The BiLSTM network builds on the feature representations extracted by the CNN, treating these features as a sequence of input vectors. Each vector corresponds to a temporal frame in the speech signal, and the task of the BiLSTM is to model how these frames evolve and predict the underlying emotion at the sequence level. Unlike standard LSTMs, which only process sequences in one direction (past to present), BiLSTM networks process sequences in both directions from start to end and from end to start. This bidirectional processing ensures that the network captures context from both preceding and succeeding frames, improving classification accuracy. The output from the BiLSTM layers is fed into fully connected layers, before the final classification of the emotional states. The layer employed a softmax function with Relu activation that translates the integrated features into a probability distribution over the possible emotions, which provided the basis for the model's predictions. A Dense Unit of size 128 and Relu activation are connected to the degree of freedom to synchronize the data and adapt to the model.

Model Training

The model was trained using the backpropagation algorithm, with the Adam optimizer selected for its efficiency in handling sparse gradients and its adaptive learning rate properties. During training, the model learned to map the preprocessed audio data to the correct emotional labels using a loss function that measured prediction accuracy. The process was iterated for 60 epochs with a large number of training examples from the self-captured Hausa audio dataset as well as the RAVDESS datasets to ensure that the model generalize well on new, unseen data. The training set was split into the ratio of 60:20:20 for training, validation, and testing respectively.

The training parameters:

Hyperparameter tuning is a critical step in optimizing the performance of deep learning models. In this study, the hyperparameters was selected based on a combination of

empirical testing and established best practices from prior research in Speech Emotion Recognition (SER) Hussain *et al*., (2021). The learning rate was set to 0.001, as it provides a balance between convergence speed and model stability. A dropout rate of 0.3 was introduced to mitigate overfitting, ensuring that the model generalizes well to unseen data (Sandugash & Anargul, 2023). A batch size of 32 was selected based on empirical studies and practical considerations. Prior research, including that of Tarunika *et al*., (2018) has shown that batch sizes in the range of 32-64 strike an optimal balance between computational efficiency and model performance. A smaller batch size (e.g., 16) could lead to slower convergence due to high variance in gradient updates, whereas a larger batch size (e.g., 128) could require excessive memory and potentially lead to less effective generalization. The batch size of 32 ensures a steady learning process, smooth convergence, and reduced risk of overfitting.

Evaluation Metrics

The evaluation metrics of accuracy, precision, recall, and F1-score are selected to evaluate the performance of the proposed model. The selected metrics offer a thorough evaluation to accurately categorize emotions, especially when there are class disparities. (Singh *et al*., 2021). The definitions and equations used to calculate the metrics for performance evaluation are explained accordingly.

Accuracy is the ratio of correctly predicted classes to the total number of classes (Elbanna *et al*., 2021) (see equation 10).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

where;

TP are the True Positives, **TN** are the True Negatives, **FP** are the False Positives, **FN** are the False Negatives

Precision (Positive Predictive Value) is the precision that measures the proportion of correctly predicted positive instances among all instances predicted as positive (equation 11).

$$Precision = \frac{TP}{TP + FP} \qquad (11)$$

High precision yields fewer false positives, a situation that is desirable in applications.

Recall (Sensitivity or True Positive Rate) which is known as sensitivity, measures the proportion of the actual positive instances that were correctly classified. (equation 12).

$$Recall = \frac{TP}{TP + FN} \qquad (12)$$

High recall means that the model is effective in recognising the true positive classes.

F1-Score is the harmonic mean of precision and recall. It offers a balanced measure when there is an unbalanced class distribution. It considers both precision and recall, especially when one metric is lower than the other (Jeong *et al*., 2023) (equation 13).

$$F1 - score = \frac{Precision \times Recall}{Precision + Recall} \qquad (13)$$

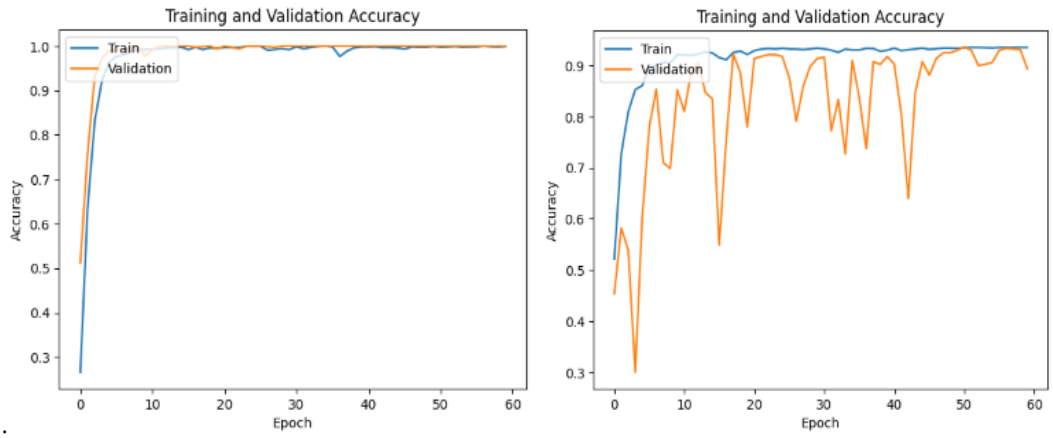
The F1-score is useful when false positives and false negatives are critical since it balances both aspects of classification performance. A high F1-score means strong performance in precision and recall (Zhou & Beigi, 2020).

3. RESULTS AND DISCUSSION

Local Hausa and RAVDESS emotional datasets were implemented and used in the training of the suggested speech emotion recognition model. The model equipped with Tensoflow, Sklearn, Numpy, Matpotlib, searbon, and librosa libraries was implemented in the Jupyter Notebook Python Anaconda 3.0 environment.

Experimental Results with the Self-Captured Hausa Dataset

The proposed model was trained on the self-captured local Hausa dataset achieving an overall accuracy of 100%. The dataset which included 4,000 audio samples was divided into four emotional classes of joy, anger, sad, and surprise. To maximise the performance, the model was trained using 60 epochs. The training duration lasted for about 28 minutes, and 37 seconds. Similarly, the technique in (Rasheed *et al*., 2024) which included an attention mechanism was evaluated using the self-captured Hausa dataset. The model was also trained using 60 epochs to compare the effectiveness against the proposed model that removed the attention mechanism. The training duration of (Rasheed *et al*., 2024) lasted for about 45 minutes, 53 seconds achieving an accuracy value of 90%. While the attention mechanism in the benchmark model added sophistication and allowed for more focused learning, it increased the risk of overfitting, especially when working with locally sourced datasets like that of the Hausa speakers. This might mean that the simpler architecture proposed, offered a more balanced approach, providing strong performance with a reduced risk of overfitting. The training and validation accuracy curve shown in Figure 3.1 (a) provides evidence that the proposed model is robust. The alignment of the two accuracy curves means that the model did not overfit. When compared with the Rasheed *et al*. (2024) model Figure 3.1 (b), results show that the training and validation accuracies of the model increased rapidly, indicating that the model quickly learns the key features necessary for the recognition. However, as training progresses, a gap begins to form between the training and validation curves. The training accuracy continues to rise, while the validation accuracy starts to plateau and even slightly declines. This pattern suggests that the benchmark model is overfitting to the training data, becoming too tailored to the specific patterns within the training set, which may not generalize well to unseen data when compared with the proposed model.



(a) (b)

Figure 3.1: Training and Validation Accuracies of Proposed Model (a) vs that of Benchmark (Rasheed *et al*., 2024) Model (b) on Self-captured Hausa Dataset

Figure 3.1 underscores the effectiveness of the hybrid CNN-BLSTM model architecture without attention mechanism in recognizing emotions from speech data, particularly when using a culturally specific dataset. The high accuracy achieved demonstrates the potential of leveraging culturally relevant datasets to improve the performance and applicability of emotion recognition systems in diverse linguistic contexts.

Classification Report of the Models With Hausa Datasets

The classification report as illustrated in Figure 3.2, shows that the Proposed model Figure 3.2 (a) achieved exceptionally high precision, recall, and F1-scores across all four emotion classes—happy, angry, sad, and surprise. In all cases, the precision, recall, and F1-score recorded 1.00, when compared with the benchmark model that recorded lower accuracy as seen in Figure 3.2 (b), Similarly, the model was able to correctly classify each emotion without any record of misclassifications. This is also reflected in the F1-scores, which further confirm the model's accuracy and reliability.

Classification Report:

	precision	recall	f1-score	support
surprise	1.00	1.00	1.00	200
sad	1.00	1.00	1.00	200
happy	1.00	1.00	1.00	200
angry	1.00	1.00	1.00	200
accuracy			1.00	800
macro avg	1.00	1.00	1.00	800
weighted avg	1.00	1.00	1.00	800

(a)

Classification Report:

	precision	recall	f1-score	support
surprise	0.88	1.00	0.94	200
sad	0.86	0.90	0.88	200
happy	0.87	0.70	0.78	200
angry	0.98	1.00	0.99	200
accuracy			0.90	800
macro avg	0.90	0.90	0.90	800
weighted avg	0.90	0.90	0.90	800

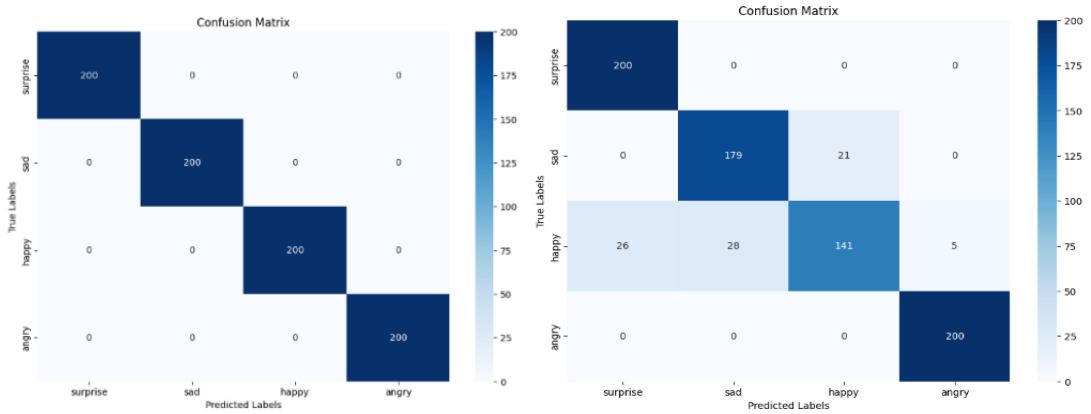
(b)

Figure 3.2: Classification Reports of the Proposed Model (a) vs that of Benchmark (Rasheed *et al*., 2024) Model (b) on the Self-captured Hausa Datasets

Figure 3.2 results underscore the effectiveness of the proposed model in categorizing emotions from the Hausa speech emotion data. The precision, recall, and F1-scores across all emotion classes show that the simplified model did not compromise the model's strength.

Confusion Matrix Report on The Self-Captured Hausa Datasets

The confusion matrix provides a visualization of the performance of the proposed model. As shown in Figure 3.3 (a), the confusion matrix for the proposed model revealed that all predictions fall perfectly on the diagonal, indicating that the model correctly classified every instance of each emotion without any misclassifications. There are no off-diagonal elements in the matrix, meaning there were no instances where the model confused one emotion with another. However, benchmark (Rasheed *et al*., 2024) model results Figure 3.3 (b) shows that there are some off-diagonal elements where "happy" instances was misclassified as "surprise", "sad" and vice versa. For example, the model incorrectly classified (21) instances of "sad" class as "happy", and misclassified (59) instances of "happy" emotions as "sad" (28), "surprise" (26) and "angry" (5) categories respectively. These misclassifications suggest that the model struggled to distinguish between these emotions, possibly due to the complexities introduced by the attention mechanism in the model's structure.



(a) (b)

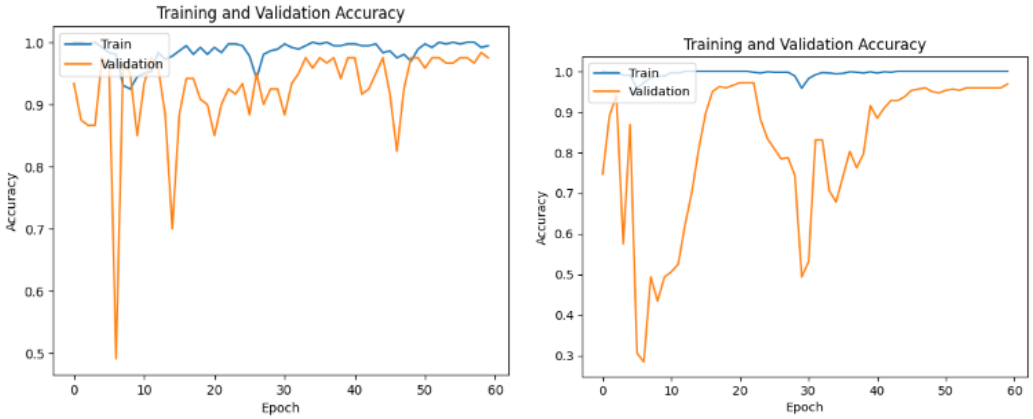
Figure 3.3: Confusion Matrix of Proposed Model (a) vs Confusion Matrix of Benchmark (Rasheed *et al*., 2024) Model (b) on Self-captured Hausa Dataset

Figure 3.3 results indicate that the proposed model is not only accurate but also consistent in its predictions across all emotion classes unlike that of the benchmark (Rasheed *et al*., 2024) model that misclassified some classes. This level of performance is particularly impressive given the model's simplified architecture, which omits the attention mechanism found in benchmark (Rasheed *et al*., 2024) model. The results also demonstrated that the proposed model's design integrating CNNs and BiLSTM without attention mechanism can provide a robust framework for emotion recognition.

Experimental Results of the Models on RAVDESS Datasets

The training and validation process of the proposed model Figure 3.4 (a) was also carried out using the RAVDESS dataset. This dataset provided a robust test of the model's ability to generalize to different linguistic and cultural contexts beyond the

Hausa dataset. The training duration stood at 18 minutes, and 23 seconds.



(a) (b)

Figure 3.4: Training and Validation Accuracies of Proposed Model (a) vs that of Benchmark (Rasheed *et al*., 2024) Model (b) on Self-captured Hausa Dataset

Figure 3.4 shows that the proposed model achieved a final accuracy of 96% while Rasheed *et al*., (2024) Model reached 94%, indicating strong generalization to the RAVDESS data. The training duration for the benchmark model lasted 31 minutes, and 26 seconds. The accuracy curves of both models demonstrated high performance, there are noticeable differences in their behavior during training, particularly in the validation curves. The Rasheed *et al*., (2024) model's validation curve (Figure 3.4 (b)) exhibits a slight divergence from the training curve, with the validation accuracy plateauing early in the process and continue to plateau as the training progresses. This suggests that the model is overfitting, meaning it is learning patterns that are too specific to the training data and less effective on unseen validation data. Overfitting often may occur when model is too complex or rely heavily on mechanisms like attention, which introduced unnecessary parameters (Sandugash & Anargul, 2023). The proposed model's validation curve on the other hand, followed the training curve with slight deviation, indicating that the model generalized well without significant overfitting.

Classification Reports of the Models on Ravdess Datasets

The classification report for the proposed model trained on the RAVDESS dataset was also provided. Figure 3.5 (a) shows that the proposed model achieved an overall accuracy of 96%. The report indicated that the proposed model achieved high precision, recall, and F1-scores, with most metrics at or close to 100%, indicating that the model was highly effective in identifying emotions. A slight drop was noted in the "happy" class, where the F1-score was 0.87, and scored the highest Precision, Recall, and F1 Score of 1.0 in the “sad”, “calm”, “angry”, and “disgust” classes respectively. On the other hand, Rasheed *et al*., (2024) model performed slightly lower, with 94% accuracy, and precision, recall, and F1-scores ranging from 0.85 to 1.00. The most significant drop was again observed in the "disgust" class, which had the lowest F1-score (0.86), demonstrating that the proposed model outperformed the benchmark model.

Classification Report:

	precision	recall	f1-score	support
angry	1.00	1.00	1.00	40
calm	1.00	1.00	1.00	40
disgust	1.00	1.00	1.00	40
fear	1.00	0.97	0.99	40
happy	1.00	0.78	0.87	40
neutral	0.85	1.00	0.92	40
sad	0.97	0.95	0.96	40
surprise	0.89	0.97	0.93	40
accuracy			0.96	320
macro avg	0.96	0.96	0.96	320
weighted avg	0.96	0.96	0.96	320

(a)

Classification Report:

	precision	recall	f1-score	support
angry	1.00	1.00	1.00	40
calm	0.95	0.93	0.94	40
disgust	0.89	0.82	0.86	40
fear	0.85	1.00	0.92	40
happy	0.95	0.97	0.96	40
neutral	0.95	0.90	0.92	40
sad	0.97	0.95	0.96	40
surprise	1.00	0.97	0.99	40
accuracy			0.94	320
macro avg	0.95	0.94	0.94	320
weighted avg	0.95	0.94	0.94	320

(b)

Figure 3.5: Classification Report of the Proposed Model (a) vs the benchmark (Rasheed *et al*., 2024) Model (b) on RAVDESS Dataset

Confusion Matrix of The Models on RAVDESS Datasets

Figure 3.6 displays the confusion matrix for the proposed model when tested on the RAVDESS dataset. The confusion matrix illustrates how well the model performed in classifying the eight (8) emotion classes. The diagonal elements of the matrix represent the instances where the model correctly classified the emotions. The matrix revealed that the model accurately predicted 4 emotions; “angry”, “calm”, and “neural” and also made a few misclassifications, particularly between the "fear", “happy”, "sad", and “surprise” classes. A small number of "fear" (1) instance were incorrectly classified as "sad", (5) instances of “happy” class were incorrectly predicted as “surprise” and (4) instances as “neutral” class, and (2) instances of “sad” category was misclassified as “neutral” emotion, indicating that the model modestly struggles to differentiate between these emotions.

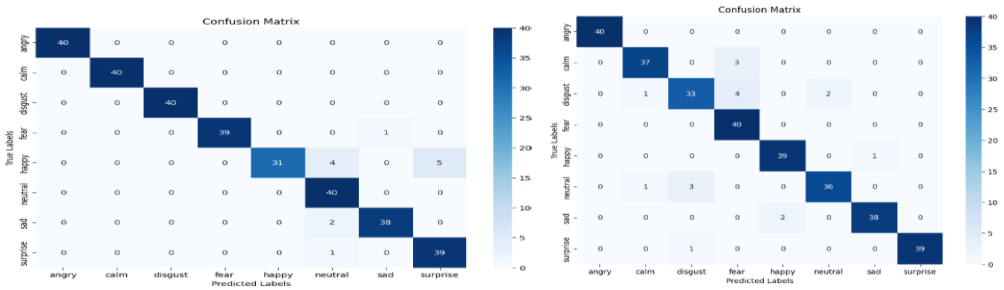


Figure 3.6: Confusion Matrix Report of the Proposed Model (a) vs benchmark (Rasheed *et al*., 2024) Model (b) on RAVDESS Dataset

The misclassifications reported in Figure 3.6 suggest that while the model is highly accurate overall, it finds certain emotions modestly challenging to distinguish, possibly due to similarities in their expression within the RAVDESS dataset; however, performed a little better when compared with the benchmark (Rasheed *et al*., 2024) model. This suggests that enhancing the model's ability to accurately classify more nuanced emotional expressions, particularly those that are easily confused, could improve its performance even further, making it more reliable in diverse applications.

Analysis of Results

The results from the experiments carried out on both the self-captured Hausa dataset and the RAVDESS dataset showed the effectiveness of the prepared model. This proposed model achieved an accuracy of 100% on the self-captured Hausa dataset and 96% on the RAVDESS dataset, with results suggesting that the model is highly capable of capturing and classifying emotional cues in speech data across different cultural contexts even without the attention mechanism.

Comparative Analysis of Models' Performance with Hausa and RAVDESS Datasets

The proposed model consistently outperformed the benchmark Rasheed *et al*., (2024) technique. The proposed model recorded a perfect accuracy of 100% with the locally sourced Hausa dataset, while the benchmark model recorded 90% accuracy. The proposed model maintained a high accuracy of 96% for the RAVDESS dataset, compared to 94% achieved by the benchmark model.

Evident in the Hausa dataset, are the differences highlighting the efficiency and robustness of the proposed model's simple architecture which did not only achieve higher performance but avoided overfitting as well. While its validation accuracy plateaued and slightly declined, the benchmark model with the attention mechanism shared signs of overfitting as its training accuracy continued to increase. This differences shows that the benchmark model became too focused to the training data, thereby losing its ability to generalise effectively to new, unseen data. In addition, the training time and computational resources required for both models further underscore the efficiency of the proposed approach. On the Hausa dataset, the proposed model had a shorter training time of approximately 28 minutes and 37 seconds and 18 minutes and 23 seconds on the RAVDESS data contrary to the Rasheed *et al*., (2024) model which took about 45 minutes and 53 seconds and 31 minutes, 26 seconds on both datasets, showing that the attention mechanism added a computational burden. This indicates that the proposed model achieved better performance while maintaining computational efficiency which in turn, makes it more suitable for applications with limited resources.

The comparative analysis across both datasets shows that the suggested CNN-BiLSTM model with the attention mechanism not only achieved higher accuracy but did so with better efficiency and generalisability. This relates with the wider trend in deep learning towards optimising model efficiency without losing performance (Harár *et al*., 2016). These discoveries suggest that, when properly designed, simpler architectures can effectively record complex/complicated patterns in speech emotion recognition, outperforming more complicated models that are prone to overfitting.

One potential concern is overfitting, particularly observed in Training and Validation Curve of the Proposed Model with RAVDESS dataset. While techniques such as dropout regularization and early stopping were employed to mitigate this risk, future work should explore further generalization techniques, such as data augmentation or adversarial training, to ensure robustness against unseen data. Additionally, expanding the dataset to include a more diverse range of dialects and speakers from different cultural backgrounds would enhance the model's applicability.

Comparative Analysis with Existing Literature

To evaluate the effectiveness of the proposed model, its performance is also compared with recent studies on Speech Emotion Recognition.

Rasheed *et al*., (2024) utilized a CNN-BiLSTM model with attention mechanisms and achieved an accuracy of 94% on the RAVDESS dataset. However, their model had increased computational overhead due to attention layers.

Hussain *et al*., (2021) employed a hybrid CNN-LSTM approach and obtained 89.5% accuracy. Their study highlighted the impact of Bag of Acoustic Words (BoAW) in improving feature representation on the RAVDESS dataset.

Zhou & Beigi (2020) proposed a Time-Delay Neural Network (TDNN) for SER, leveraging transfer learning techniques, and achieved a competitive accuracy of 92.3% on the IEMOCAP dataset.

Table 4.2: Comparative Analysis with Existing Literature

Study	Model Used	Dataset	Accuracy
Rasheed *et al*., (2024)	CNN-BiLSTM with Attention	RAVDESS	94%
Hussain *et al*., (2021)	Hybrid CNN-RNN	RAVDESS	89.5%
Zhou & Beigi (2020)	TDNN with Transfer Learning	IEMOCAP	92.3%
Proposed Model	CNN-BiLSTM (No Attention)	Hausa Emotional Speech Dataset	100%
Proposed Model	RAVDESS	RAVDESS	96%

The Proposed model, which integrates CNN with BiLSTM without attention mechanisms, achieved a comparable accuracy of 100% on Hausa Emotional Speech Datasets and 96% on RAVDESS Datasets, demonstrating a balance between computational efficiency and accuracy. Unlike models relying on attention mechanisms, the proposed approach reduces complexity, making it more suitable for real-time applications, particularly on resource-constrained devices. Additionally, the inclusion of the Hausa Emotional Speech Dataset enhances the cultural adaptability of SER models, addressing linguistic biases present in prior works.

The comparative analysis suggests that our model offers an optimal trade-off between accuracy and computational efficiency while expanding the applicability of SER systems to underrepresented languages.

CONCLUSION

In conclusion, an improved model comprising of CNN BiLSTM without attention mechanism was proposed for speech emotion recognition task. The technique demonstrated excellent recognition accuracies of 100% and 96% on the locally sourced Hausa dataset and the RAVDESS dataset respectively. The proposed model, which extracted spatial and temporal information from audio data without the attention mechanism helped reduce time complexity. The model was made simpler compared to the base model. The results show that the proposed method worked very well in cross-cultural settings. It also demonstrated how crucial it is to use culturally inclusive speech datasets to increase the generalisability and resilience of emotion identification systems. The proposed improved technique provides an answer for practical uses, in settings with limited resources.

Future Work

Building on the findings of this study, several areas of future research can further enhance the effectiveness and applicability of SER models for underrepresented languages.

Expansion To Other Languages

One key direction for future research is expanding this approach to other African and underrepresented languages. Developing emotional speech datasets for languages such as Yoruba, Igbo, Kanuri, Bura, and other African Languages would increase linguistic diversity and improve SER models' ability to generalize across different cultural contexts. Cross-linguistic adaptation techniques, including transfer learning and zero-shot learning, could be explored to make SER models more robust across multiple languages.

Enhancing Emotion Categories

Future research should incorporate additional emotional states such as fear, disgust, neutrality, and contempt to provide a more comprehensive emotional spectrum. This expansion would enhance model accuracy and ensure the recognition of more nuanced emotional expressions in speech.

Exploring Advanced Model Architectures

To further improve the performance of SER systems, future studies could explore advanced deep learning architectures such as transformer-based models and self-supervised learning approaches. Transformer models, such as Wav2Vec and SpeechT5, have demonstrated strong capabilities in speech processing and could offer improved feature extraction and classification for emotion recognition tasks. Additionally, hybrid models combining CNN, BiLSTM, and attention mechanisms could be evaluated to determine their trade-offs in accuracy and computational efficiency.

Increasing Dataset Diversity

Future research should focus on collecting a more diverse dataset with greater representation of dialectal variations and socio-demographic diversity. Including speakers from different age groups, educational backgrounds, and geographic regions would enhance the generalizability of SER models. Furthermore, expanding the dataset size beyond 4,000 samples would strengthen model robustness and reduce the risk of overfitting.

Real-World Deployment and Applications

To ensure practical applications of SER models, future work should explore deployment on real-world systems, such as virtual assistants, call centers, and mental health monitoring platforms. Evaluating the model's effectiveness in natural, noisy environments will help assess its viability in real-world interactions. Additionally, developing lightweight, energy-efficient models for mobile and embedded systems would further enhance the accessibility of SER technologies.

Ethical Statement

The Ethical Committee of the Nigerian Defence Academy approved the current experiment.

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Mujtaba K. Tahir and Martins E. Irhebhude

Acquisition, analysis, or interpretation of data: Mujtaba K. Tahir, Adeola O. Kolawole, and Martins E. Irhebhude.

Drafting of the manuscript: Mujtaba K. Tahir and Adeola O. Kolawole.

REFERENCES

Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensor Journal*, *4*(21), 1249. https://doi.org/https://doi.org/10.3390/s21041249

Ahmed, S. A., Mahmood, B. N., Mahmood, D. J., & Namq, M. M. (2024). ENHANCING KURDISH SIGN LANGUAGE RECOGNITION THROUGH RANDOM FOREST CLASSIFIER AND NOISE REDUCTION VIA SINGULAR VALUE DECOMPOSITION (SVD). *Science Journal of University of Zakho*, *12*(2), 257-267. https://doi.org/10.25271/sjuoz.2024.12.2.1263

Al-Tamimi, J. (2022). *JalalAl-Tamimi/Praat-Silence-Detection: Praat-Silence detection*. In Zenodo.

Ali, S. H., Mohammed, A. I., Mustafa, S. M. A., & Salih, S. O. (2025). WEB VULNERABILITIES DETECTION USING A HYBRID MODEL OF CNN, GRU AND ATTENTION MECHANISM. *Science Journal of University of Zakho*, *13*(1), 58-64. https://doi.org/10.25271/sjuoz.2025.13.1.1404

Barazida, N. (2021). 40 Open-Source Audio Datasets for ML. In: Towards Data Science.

Calzone, O. (2022). An Intuitive Explanation of LSTM. *An Intuitive Explanation of LSTM*.

Dixit, C., & Satapathy, S. M. (2023). Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Systems with Applications*, *240*, 122579. https://doi.org/10.1016/j.eswa.2023.122579

Duttaa, D., Halderb, S., & Gayen, T. (2023). Intelligent Part of Speech tagger for Hindi. *Procedia Computer Science*, *218*(3), 604-611.

Elbanna, G., Scheidwasser-Clow, N., Kegler, M., Beckmann, P., Hajal, K. E., & Cernak, M. (2021). BYOL-S: Learning Self-supervised Speech Representations by Bootstrapping. *Proceedings of Machine Learning*

Research, *86*(6), 365-375. https://doi.org/arXiv:2206.12038v4 [cs.SD]

Goncalves, L., Leem, S.-G., Lin, W.-C., Sisman, B., & Busso, C. (2023). VERSATILE AUDIO-VISUAL LEARNING FOR HANDLING SINGLE AND MULTI MODALITIES IN EMOTION REGRESSION AND CLASSIFICATION TASKS. *Prime AI*, *18*(2), 5587-5598.

Gong, Y., Chung, Y.-A., & Glass, J. (2021). Audio Spectrogram Transformer. *MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA*, *23*(3), 524-577. https://doi.org/arXiv:2104.01778v3 [cs.SD]

Harár, P., Burget, R., & Dutta, M. K. (2016). Emotion recognition using MFCC and RBF network. *International Journal of Engineering and Technology*, *8*(1), 209-315. https://doi.org/10.21817/ijet/2016/v8i1/160801309

Hasan, Z. F. (2022). An Improved Facial Expression Recognition Method Using Combined Hog and Gabor Features. *Science Journal of University of Zakho*, *10*(2), 54-59. https://doi.org/10.25271/sjuoz.2022.10.2.897

Hassan, D. A. (2025). DEEP NEURAL NETWORK-BASED APPROACH FOR COMPUTING SINGULAR VALUES OF MATRICES. *Science Journal of University of Zakho*, *13*(1), 1-6. https://doi.org/10.25271/sjuoz.2025.13.1.1345

Hassan, M. M., & Ahmed, D. (2023). BAYESIAN DEEP LEARNING APPLIED TO LSTM MODELS FOR PREDICTING COVID-19 CONFIRMED CASES IN IRAQ. *Science Journal of University of Zakho*, *11*(2), 170-178. https://doi.org/10.25271/sjuoz.2023.11.2.1037

Hossain, M. S., & Muhammad, G. (2018). Deep learning approach for emotion recognition from audio-visual data. *Personal and Ubiquitous Computing*, *22*(1), 3-14. https://doi.org/https://doi.org/10.1007/s00779-017-1072-7

Ibrahim, U. A. (2021). Hausa Speech Dataset. *Mendley Data*, *23*(1). https://doi.org/10.17632/z38hsttxb.1

Irhebhude, M. E., Kolawole, A., & Goshit, N. (2023). Perspective on Dark-Skinned Emotion Recognition Using Deep-Learned and Handcrafted Feature Techniques. *Emotion Recognition Using Deep-Learned and Handcrafted Feature Techniques*, *2*(5), 25-50. https://doi.org/10.5772/intechopen.109739

Irhebhude, M. E., Kolawole, A. O., & Zubair, W. M. (2024). Sign Language Recognition Using Residual Network Architectures for Alphabet And Diagraph Classification. *Journal of Computing and Social Informatics*, *4*(1), 11-25. https://doi.org/10.33736/jcsi.7986.2025

Jeong, E., Kim, G., & Kang, S. (2023). Multimodal Prompt Learning in Emotion Recognition Using Context and Audio Information. *Mathematics Journal*, *11*(2), 2908-2923. https://doi.org/https://doi.org/10.3390/math11132908

Kim, Y., Lee, H., Provost, & Mower, E. (2015). Deep learning approaches for emotion recognition from speech and non-speech. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, *25*(2), 433-440. https://doi.org/https://doi.org/10.1145/2818346.2820779

Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, *244*, 108580. https://doi.org/10.1016/j.knosys.2022.108580

Njoku, J. N., Caliwag, A. C., Lim, W., Kim, S., Kim, S., Hwang, H.-J., & Jeong, J.-W. (2021). Deep Learning Based Data Fusion Methods for Multimodal Emotion Recognition. *The Journal of Korean Institute of Communications and Information Sciences*, *47*(1), 79-122.

Rasheed, B. H., Yuvaraj, D., Alnuaimi, S. S., & Priya, S. S. (2024). Automatic Speech Emotion Recognition Using Hybrid Deep Learning Techniques. *International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING*, *12*(15), 87-96.

SaiDhruv, Y. H., k, P., & Vardhan, M. V. (2023). Speech Emotion Recognition Using LSTM Model. *International Conference on Recent Trends in Data Science and its Applications*, *5*(4), 692-697. https://doi.org/rp-9788770040723.133

Saunders, M. (2023). *Hausa*. Oxford University Press, London.

Schonevelda, L., Othmanib, A., & Abdelkawyb, H. (2020). Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition: A Review of Recent Progress. *Signal Processing Magazine*, *37*(5), 141-152. https://doi.org/10.1109/MSP.2020.3006019

Selvaraj, M., Bhuvana, R., & Padmaja, S. (2016). Recognising emotions using deep learning. *International Journal of Engineering and Technology*, *8*(1). https://doi.org/10.21817/ijet/2016/v8i1/160801309

Singh, P., Srivastava, R., Rana, K. P. S., & Kumar, V. (2021). A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, *229*, 107316. https://doi.org/10.1016/j.knosys.2021.107316

Sundar, B. S., Rohith, V., Suman, B., & Chary, K. N. (2022). Emotion Detection in Text Using Machine Learning and Deep Learning Techniques. *International Journal for Research in Applied Science & Engineering Technology*, *10*(6), 2276-2282. https://doi.org/https://doi.org/10.22214/ijraset.2022.44293

Tarunika, K., Pradeeba, R. B., & Aruna, P. (2018). Accuracy of speech emotion recognition through deep neural network and k-nearest. *International Journal of Engineering Research in Computer Science and Engineering,*, *5*(2), 2320-2394.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). Automatic recognition of emotion in speech and facial expressions: A multimodal approach. *IEEE Transactions on Affective Computing*, *9*(4), 578-584.

Vaaras, E., Ahlqvist-Björkroth, S., Drossos, K., Lehtonen, L., & Räsänen, O. (2023). Development of a speech emotion recognizer for large-scale child-centered audio recordings from a hospital environment. *Speech Communication*, *148*, 9-22. https://doi.org/10.1016/j.specom.2023.02.001

Zaman, K., Sah, M., Direkoglu, C., & Unoki, M. (2023). A Survey of Audio Classification Using Deep Learning. *IEEE ACCESS*, *10*(7), 290-350. https://doi.org/10.1109/ACCESS.2023.3318015

Zhou, S., & Beigi, H. (2020). A Transfer Learning Method for Speech Emotion Recognition from Automatic Speech Recognition. *Electrical Engineering and Systems Science*, *5*(7), 2356-2363. https://doi.org/arXiv:2008.02863v2