

# Advancements in accurate speech emotion recognition through the integration of CNN-AM model

Marion Olubunmi Adebisi<sup>1</sup>, Timothy T Adeliyi<sup>2</sup>, Deborah Olaniyan<sup>1</sup>, Julius Olaniyan<sup>1</sup>

<sup>1</sup>Department of Computer Science, College of Pure and Applied Sciences, Landmark University, Kwara-State, Nigeria

<sup>2</sup>Department of Informatics, Faculty of Engineering, Built Environment and IT, University of Pretoria, Pretoria, South Africa

## Article Info

### Article history:

Received Sep 9, 2023

Revised Feb 2, 2024

Accepted Feb 21, 2024

### Keywords:

Attention mechanism

Convolution neural network

Emotion

Recognition

Signal

Speech

## ABSTRACT

In this study, we introduce an innovative approach that combines convolutional neural networks (CNN) with an attention mechanism (AM) to achieve precise emotion detection from speech data within the context of e-learning. Our primary objective is to leverage the strengths of deep learning through CNN and harness the focus-enhancing abilities of attention mechanisms. This fusion enables our model to pinpoint crucial features within the speech signal, significantly enhancing emotion classification performance. Our experimental results validate the efficacy of our approach, with the model achieving an impressive 90% accuracy rate in emotion recognition. In conclusion, our research introduces a cutting-edge method for emotion detection by synergizing CNN and an AM, with the potential to revolutionize various sectors.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Deborah Olaniyan

Department of Computer Science, College of Pure and Applied Sciences, Landmark University

Omu-Aran, Kwara-State, Nigeria

Email: [deborah.oluwafisayo-fatinikun@lmu.edu.ng](mailto:deborah.oluwafisayo-fatinikun@lmu.edu.ng)

## 1. INTRODUCTION

Working towards the achievement of United Nations Sustainable Development Goal 4, the integration of cutting-edge technologies like convolutional neural networks-attention mechanism (CNN-AM) is crucial in propelling advancements in industry practices, particularly within the field of artificial intelligence. In our rapidly evolving technological landscape, the field of speech emotion detection has garnered significant attention, presenting substantial potential to improve various facets of human interaction and well-being [1]. The ability to discern and interpret emotions conveyed through spoken language has far-reaching applications, from enhancing the user experience in human-computer interactions to aiding in the assessment of mental health [2]. Yet, the core of this commitment relies on precisely and effectively understanding emotions from speech data, a task that is both intricate and crucial [3].

Currently, methods for recognizing emotions in speech often involve manually designing features and using traditional ways of teaching machines. These techniques look at things like pitch, energy, and other parts of the sound [4]. Although they might perform acceptably, they cannot entirely comprehend the full range of emotions people convey through speech. These methods might struggle with more complex emotions, making it hard to tell exactly what someone is feeling. Also, relying too much on these preset ways can make it difficult for the methods to adapt to different situations and data, showing the need for more flexible and advanced approaches to understand emotions in speech. While these methods have achieved moderate success, they often fall short in capturing the intricate nuances of human emotions expressed through speech [5], [6]. In response to this challenge, various techniques have been explored within the field of speech emotion detection. These include the analysis of acoustic properties, machine learning-based

segmentation and classification of speech segments, and the application of natural language processing to scrutinize the content of spoken language [7], [8]. These techniques, although versatile, have their limitations, highlighting the need for more accurate, and nuanced approaches to recognize emotions in speech [9].

Our primary objective in this research is to introduce an innovative solution that transcends the constraints of traditional methods. We aim to leverage the capabilities of deep learning, specifically CNNs, and incorporate the focus-enhancing attributes of AM to develop a model capable of accurately and precisely recognizing emotions conveyed through spoken language. To address the intricacies of recognizing emotions in speech, we propose a pioneering CNN-AM approach. This approach merges the robust feature extraction capabilities of CNNs with the nuanced focus-enhancing attributes of attention mechanisms. By doing so, we empower the model to dissect critical temporal and spectral characteristics within the speech signal, thereby elevating its capacity for emotion recognition. We envision that this amalgamation will enable the model to decode emotions with an unprecedented level of accuracy and nuance.

In the forthcoming sections, we will delve into the intricacies of our CNN-AM approach, shedding light on the innovative architecture that underpins our research. We anticipate that this groundbreaking approach will have transformative implications, not only in revolutionizing human-computer interactions but also in making substantial contributions to the realm of mental health assessment. Our research represents a pivotal shift in the landscape of speech emotion detection, ushering in new horizons for comprehending, and effectively responding to human emotional states.

## 2. LITERATURE REVIEW

In the rapidly expanding field of deep learning, sophisticated multilayered networks are used to simulate high-level patterns in data. It is mostly utilized in artificial intelligence and machine learning. The study's literature review will include citations to expert views on the topic of deep learning. Zhang *et al.* [10], the authors used the raw speech data for the speech-emotion recognition (SER) and the CNN method. The CNN model analyses the input signals to identify sounds before choosing specific areas of the audio sample for emotion recognition, which at the time produced higher results. Yu and Kim [11] study introduces a SER model that merges attention and long short-term memory (LSTM) components, incorporating IS09 and mel spectrogram features. Initially achieving a weighted accuracy (WA) of 68%, the model's performance is hindered by the interactive emotional dyadic motion capture (IEMOCAP) dataset's reliability issues. X14, a more reliable dataset is reconstructed, resulting in an improved WA of 73%. Zhao *et al.* [12] created a sophisticated two-dimensional (2D) CNN-LSTM model for an emotion identification system that utilized log-mel spectrograms to identify the spatial and temporal information included in the voice data. The authors used the Berlin emotion dataset. (EMO-DB) corpus used in Lim *et al.* [13] for emotion recognition that made use of the time disturbed CNN layer and outperformed the leading model at the time by a wide margin. The authors used a short-term fourier transform algorithm to transform the speech signals into spectrograms, which they then fed into the time-distributed CNN model.

A revolutionary deep and wide CNN architecture known as RCNN-CTC, which has residual connections and the connectionist temporal classification (CTC) loss function, was developed in this study Wang *et al.* [14]. An entire system called RCNN-CTC capable of simultaneously utilizing the temporal and spectral structures of voice inputs. In addition, the authors present a system combination based on CTC that differs from the typical frame-wise senone-based one. The fundamental subsystems used in the combination are of various types, which complement one another. Using only CNN and utilizing current developments in audio models from the raw waveform and language modeling, Zeghidour *et al.* [15], provided an alternate method in this research. With no need for feature extraction, our completely convolutional technique is trained from beginning to end to predict characters from the raw waveform. Terms are deciphered through the utilization of an external convolutional linguistic representation. The model is said to be in line with the most recent state-of-the-art on wall street journal. The authors reported cutting-edge performance on LibriSpeech among end-to-end models, including deep speech 2, which was trained using significantly more linguistic data and 12 times more acoustic data.

Hannun *et al.* [16] suggest a completely convolutional sequence-to-sequence encoder architecture with an easy-to-use decoder. The model is an order of magnitude more effective than a strong recurrent neural network (RNN) baseline and enhances word error rate (WER) on LibriSpeech. The approach's core component is a time-depth separable convolution block, which significantly lowers the number of model parameters while maintaining a huge receptive field. The time-depth separable convolution architecture enhances the best previously published sequence-to-sequence outcomes on the noisy LibriSpeech test set by more than 22% relative WER when combined with a neural language model. Li *et al.* [17] presents state-of-the-art LibriSpeech end-to-end speech recognition model performance without the use of external training data. Only 1D convolutions, batch normalization, rectified linear unit (ReLU), dropout, and residual connections are used in our model, jasper. The authors also offer NovoGrad, a novel layer-wise optimizer, to

enhance training. They use studies to show that the suggested deep design outperforms more sophisticated options, if not better. 54 convolutional layers are used in our deepest jasper version. With this architecture, the authors got 3.86% WER using a greedy decoder on LibriSpeech test-clean and 2.95% WER using a beam-search decoder with an external neural language model.

Mohamed *et al.* [18], suggest using convolutionally learned input representations in place of the transformers sinusoidal positional embedding. These contextual representations supply the relative positioning data required for identifying broad links between local concepts and succeeding transformer blocks. The authors presented results are generated with a fixed learning rate of 1.0 and no warm-up steps, which are beneficial optimization properties for the proposed system. Han *et al.* [19] investigated a unique CNN-RNN-transducer architecture that we name ContextNet to bridge this gap and go beyond it. ContextNet boasts a fully convoluted encoder, fortified with convolution layers that integrate squeeze-and-excitation modules, seamlessly incorporating holistic contextual information. The authors also provide a straightforward rescaling approach that expands ContextNet's dimensions, achieving a harmonious equilibrium between computational efficiency and reliability. They showed that ContextNet obtains a WER of 2.1%/4.6% on the clean/noisy LibriSpeech test sets without external language model (LM), 1.9%/4.1% with LM, and 2.9%/7.0% with only 10M parameters. This contrasts with the best model that was previously reported, which had values of 2.0%/4.6% with LM and 3.9%/11.3% with 20M. Albanie *et al.* [20], explored how to integrate convolution neural networks with transformers to describe both local and global dependencies of an audio sequence in a parameter-efficient manner, achieving the best of both worlds. The model obtains WER of 2.1%/4.3% on the popular LibriSpeech benchmark without the usage of a language model and 1.9%/3.9% on test/test other while utilizing an external language model. Additionally, the authors note competitive performance of 2.7%/6.3% using a modest model with only 10M parameters. In this regard, with CNN-AM efficiency, this research explores the potential of deep learning to enhance the automatic speech recognition of an e-learning system. To train model, the speech signals are processed to extract spectrogram features.

### 3. METHOD

In order to obtain spectrogram features for emotion recognition, the raw audio data from the IEMOCAP database undergoes a preprocessing stage as shown in Figure 1. This involves audio segmentation, where the continuous recordings are divided into smaller segments of fixed duration. This segmentation process facilitates efficient processing by the CNN-AM model. Next, spectrogram generation is performed using the short-time fourier transform (STFT) technique. The STFT converts the time-domain audio signals into 2D representations, capturing the frequency content of the audio over time. These spectrogram features serve as valuable inputs for the CNN-AM model, enabling it to effectively recognize emotions in the speech data.

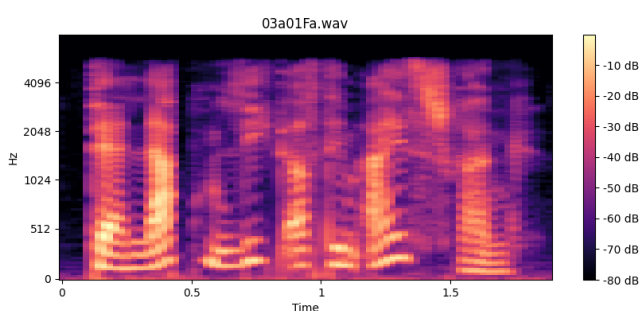


Figure 1. Sample of the spectrogram

#### 3.1. Data collection

The IEMOCAP dataset is a widely used and publicly available dataset for research in emotion recognition. It consists of approximately 12 hours of audio-visual recordings of actors engaged in scripted and improvised dialogues [21]. The dataset captures facial expressions, speech, and body gestures using motion capture technology. It provides annotations for seven discrete emotion categories and includes detailed labels for emotion, phonetic content, prosodic features, and more. The IEMOCAP dataset is known for its naturalistic emotional expressions and has played a significant role in advancing emotion recognition models and understanding human affective behavior [22].

### 3.2. Data preprocessing

Before training the CNN-AM model, the speech signals from the IEMOCAP dataset need to be preprocessed to extract spectrogram features which is presented in Figure 1. This involves the following steps:

- Audio segmentation: the continuous audio recordings are divided into smaller segments of fixed duration (e.g., 2-4 seconds) to feed into the model. This allows the model to process manageable chunks of audio data [23].
- Spectrogram generation: for each audio segment, the STFT is applied to convert the time-domain audio signal into a spectrogram representation. The spectrogram is a 2D matrix that represents the magnitude of different frequency components over time.
- Data labeling: each audio segment in the dataset is labeled with the corresponding emotional state expressed by the speaker (e.g., happy, sad, angry, and neutral).

There are four basic steps in our speech recognition system. The voice sample collecting comes first. The second features vector is created once the features are extracted as shown in Figure 2. The next stage was to try to identify the characteristics that each emotion should be distinguished by. To classify the speech using the chosen features and a previously trained model for recognition, these features are added to a deep learning model. The features represented as an image as presented in Figure 1 was passed into the CNN.

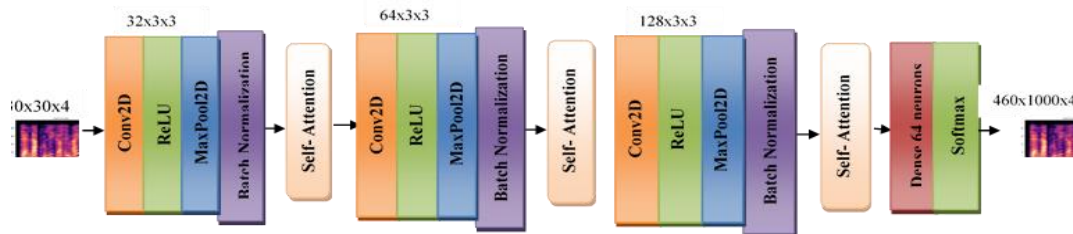


Figure 2. Block diagram of CNN-AM

Algorithm 1 depicts a sample code that uses the librosa library to load an audio file, extract mel-frequency cepstral coefficients (MFCCs), and visualize them as a spectrogram, subsequently saving the image as 'spectrogram.png'. The pseudocode for extracting features from a speech signal and converting it into a spectrogram image, which is crucial for compatibility with the CNN-AM model speech emotion detection is presented in algorithm 2. This pseudocode encapsulates the process of loading an audio file, extracting MFCC features, visualizing them as a spectrogram, and saving the spectrogram as an image.

---

**Algorithm 1:** Sample code for extracting feature and converting to img

---

```
# Import necessary libraries
import librosa
import librosa.display
import matplotlib.pyplot as plt
import numpy as np
# Load the audio file
audio_file_path = "path_to_audio_file.wav"
audio_signal, sample_rate = librosa.load(audio_file_path)
# Extract audio features (e.g., MFCCs)
mfccs = librosa.feature.mfcc(y=audio_signal, sr=sample_rate, n_mfcc=13)
# Visualize the MFCCs as a spectrogram
plt.figure(figsize=(10, 4))
librosa.display.specshow(librosa.power_to_db(mfccs, ref=np.max), y_axis='mel', x_axis='time')
plt.colorbar(format='%+2.0f dB')
plt.title('MFCC Spectrogram')
# Save the spectrogram as an image
plt.savefig("spectrogram.png")
# Show the spectrogram (optional)
plt.show()
```

---

---

**Algorithm 2:** Pseudocode for extracting feature and converting to img

---

**Input:** Speech audio file

**Output:** Spectrogram image

---

Import "librosa", "matplotlib.pyplot", and "numpy"

audio\_file\_path = "path\_to\_audio\_file.wav"

audio\_signal, sample\_rate = librosa.load(audio\_file\_path)

mfccs = librosa.feature.mfcc(y=audio\_signal, sr=sample\_rate, n\_mfcc=13)

Create a figure of size (10, 4)

Display the spectrogram of the MFCCs with librosa.display.specshow(librosa.power\_to\_db(mfccs, ref=np.max), y\_axis='mel', x\_axis='time')

Add a color bar with the format '%+2.0f dB'

Set the title of the figure as 'MFCC Spectrogram'

Save the figure as an image named "spectrogram.png"

Display the figure

---

### 3.3. Model architecture

As depicted in Figure 2 the CNN-AM model is a powerful blend of CNNs and an AM, designed to effectively extract and emphasize key features from speech data. Its integration enhances the precision of emotion recognition in complex audio signals. The CNN-AM model is an innovative and powerful deep learning architecture depicted in Figure 3. It is designed to extract and learn discriminative features from spectrogram images for speech emotion recognition [24]. This hybrid model shown in Figure 3 seamlessly combines CNNs with an attention mechanism, leveraging their respective strengths to improve the model's ability to discern subtle emotional cues in speech data.

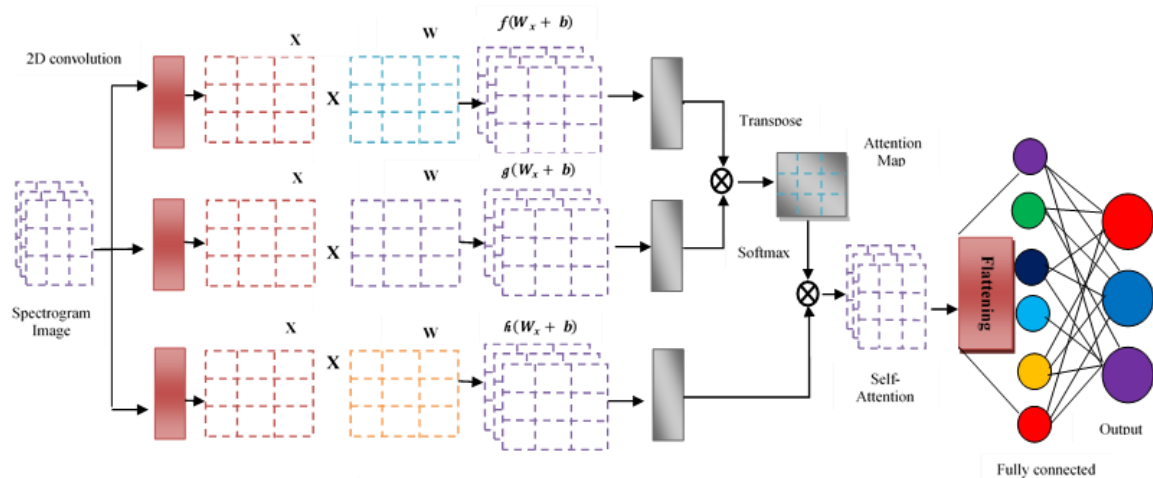


Figure 3. The schematic diagram of CNN-AM

- CNN layers: at the heart of the CNN-AM model are the CNN layers, responsible for the initial feature extraction process. These layers are inspired by the success of CNNs in various computer vision tasks and have proven to be effective in learning spatial patterns from image data [25]. In the context of speech emotion recognition, the spectrogram images act as 2D representations of the audio signal, capturing the frequency content over time. The CNN layers consist of multiple convolutional filters applied across the spectrogram images, scanning for local patterns and low-level features. These filters, through the process of convolution, learn to detect important acoustic characteristics, such as pitch, energy distribution, and spectral changes, which are crucial for emotion expression in speech. By stacking several convolutional layers and incorporating activation functions, such as ReLU, the model can progressively learn higher-level representations and more complex features, enabling it to capture emotional cues at different temporal scales. In order to diminish geometric aspects of convolutional feature maps and improve computational efficiency, pooling layers, such as max-pooling, are applied. The pooling layers reduces

the resolution of the feature maps, retaining essential details while decreasing the model's sensitivity to minor variations. This hierarchical feature extraction process in the CNN layers enables the model to comprehend the spectrogram's intricate patterns and prepares it for the subsequent attention mechanism's integration as seen in Figure 3.

- AM: the AM is a critical addition to the CNN-AM model, addressing the challenge of identifying emotionally relevant regions within the spectrogram images. While CNNs excel at capturing local features, they may not prioritize the most informative parts of the input data, potentially diluting the focus on emotionally salient segments. The AM addresses this limitation by introducing a gating mechanism that dynamically weighs the importance of different regions in the spectrogram. During the attention process, the model evaluates the relevance of each temporal slice of the spectrogram and assigns attention weights accordingly. Regions exhibiting high attention weights are amplified, while regions with lower weights are downplayed.

The computational handshake of the 2D convolution and self mechanism operation is shown this section. Each feature space (f, g, and h) is transformed using learned weight matrices (W) and bias parameters (b). The embedding features were packed into a matrix, and then multiplied by the trained weight matrices, which can be calculated by (1)-(3).

$$f(x_i) = Wx_i + b \quad (1)$$

$$g(x_i) = Wx_i + b \quad (2)$$

$$h(x_i) = Wx_i + b \quad (3)$$

Pooling layers reduce spatial dimensions, typically with operations like MaxPooling or AveragePooling. For the attention calculation, attention scores between different elements in the input feature space is calculated. This is done using a similarity function, specifically the dot product (transposed  $f(x_i)$  \*  $g(x_i)$ ), followed by a scaling operation with  $\sqrt{d_k}$  as depicted in (4).

$$e_{ij} = \frac{f(x_i)^T g(x_j)}{\sqrt{d_k}} \quad (4)$$

The softmax operation is carried out by passing the similarity scores  $e_{ij}$  are passed through a softmax function to obtain attention weights  $a_{ij}$ . These weights indicate the importance of one element with respect to all other elements in the input space as shown in (5).

$$a_{ij} = \text{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_j \exp(e_{ij})} \quad (5)$$

For the weighted sum, the attention weights ( $a_{ij}$ ) in (6) are used to weight the elements in the  $h(x_i)$  feature space. This results in a weighted sum of the feature vectors, where the weights are determined by the attention mechanism.

$$o = a_{ij} h(x_i) \text{softmax}(e_{ij}) h(x_i) \quad (6)$$

“o”, represents the self-attention feature maps. These feature maps capture the relationships and dependencies between different elements in the input feature space, emphasizing important elements based on the attention weights.

For the dense layer (fully connected layer), the attended feature vectors o from the self-attention layer are typically reshaped or flattened as calculated in (7). These vectors are passed through one or more fully connected (dense) layers with weights and biases.

$$D_i = \text{softmax}(W_{D_i} o + b_{D_i}) \quad (7)$$

The final dense layer produces the model's output Y in (8):

$$Y(W_{out} D_i + b_{out}) \quad (8)$$

By emphasizing the most informative segments in the spectrogram, the AM allows the model to concentrate on emotionally expressive cues, such as distinctive prosodic patterns, emotionally charged

phonetic features, and characteristic changes in the spectral content. This fine-grained focus enhances the model's discriminative power, enabling it to better differentiate between subtle emotional nuances, which may play a pivotal role in speech-based emotion recognition. In conclusion, the CNN-AM model elegantly combines the strengths of CNNs for hierarchical feature extraction and the AM for focused learning. By leveraging the power of deep learning and attentive processing, the CNN-AM model significantly improves the accuracy and sensitivity of speech emotion recognition, making it a valuable tool in various fields.

### 3.4. Model training

Once the spectrogram features and their corresponding emotional labels are prepared, the CNN-AM model undergoes a crucial training process on the IEMOCAP dataset. This training phase is essential for the model to learn and adapt its parameters to accurately recognize emotions from the speech data. The training process involves several key steps:

- **Data split:** to ensure a robust evaluation of the model's performance and prevent overfitting, the IEMOCAP dataset is splitted into three categories; the training set, the validation set, and the testing set. The training set is the largest subset and is used to optimize the model's parameters. The validation set is used to fine-tune hyperparameters and monitor the model's performance during training. Finally, the testing set, which remains unseen by the model during training, is used to evaluate the model's generalization and overall emotion recognition accuracy. The data split is typically performed randomly while maintaining a balanced distribution of emotional classes across all subsets. This ensures that the model learns to recognize emotions from all emotional categories, providing a more comprehensive understanding of the dataset's emotions. Figure 4 illustrates the distribution of speech emotional classes for both the (a) testing and (b) training datasets. The figure presents a visual representation of the relative proportions of different emotional categories in the datasets, providing insights into the balance and diversity of emotional states captured in the training and testing phases of the study.

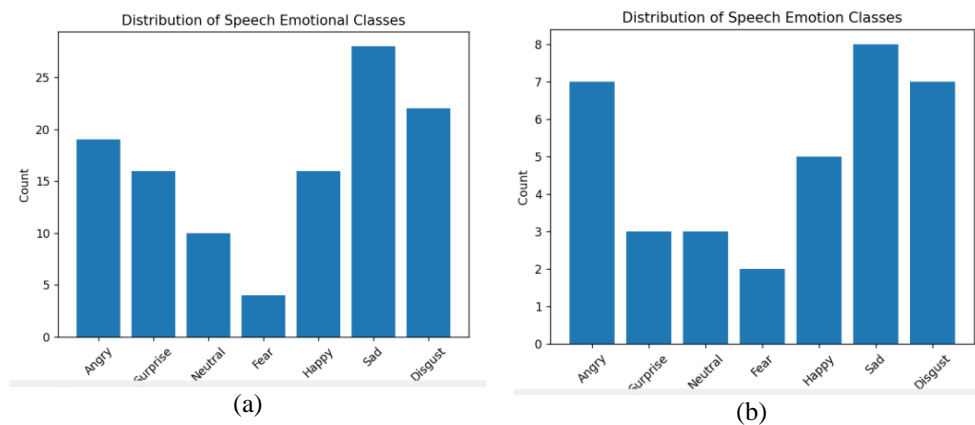


Figure 4. Distribution of speech emotional classes for (a) testing and (b) training

- **Loss function:** the choice of an appropriate loss function is crucial for training a CNN-AM model for speech emotion recognition. As this is a multi-class classification task with emotions falling into distinct categories, commonly used loss functions include categorical cross-entropy and weighted cross-entropy. Categorical cross-entropy is a popular choice for multi-class classification tasks and measures the dissimilarity between the predicted probability distribution and the ground truth labels [26]. On the other hand, weighted cross-entropy is useful when dealing with imbalanced emotional classes in the dataset, as it assigns different weights to each class to address class imbalance issues [27]. The selected loss function serves as the basis for evaluating how well the model performs in predicting the emotional states based on the spectrogram features and it is presented in Figure 5.



Figure 5. Training and loss validation

The Figure 5 depicts how the model learnt meaningful patterns from the training data while avoiding over-fitting, where it memorizes noise in the data. Figure 5 shows how the loss validation help to select the best model by providing insights into its performance on new, unseen data.

- Optimization: during training, the CNN-AM model's parameters are iteratively updated to minimize the chosen loss function. Gradient-based optimization algorithms, such as stochastic gradient descent (SGD) or adam, are commonly employed to perform this optimization. These algorithms calculate the gradients of the loss function with respect to the model's parameters and update them in the direction that minimizes the loss [28]. During each training iteration, a batch of spectrogram feature samples along with their corresponding emotional labels is fed into the model. The model's predictions are compared with the ground truth labels using the chosen loss function, and the gradients are back-propagated through the network for adjusting and refining of the parameters of the model. The procedure is iterated across several training iterations until the model converges to a state where it effectively recognizes emotions in the speech data. Table 1 presents the specific settings and important factors used in our proposed network. We chose these settings carefully to make sure our model works well for recognizing emotions in speech. The table includes details like the learning rate, batch size, number of layers, and filter sizes. These details are crucial as they affect how our network is built and trained to understand emotions in speech accurately.

Table 1. The hyper parameters and settings of the CNN-AM network

Parameters	Value
Convolution filter	3*3
Activation function	ReLU
Dropout factor	0.25
Optimizer	Adam
Self-attention channel	32
Learning rate	0.1

By the end of the training process, the CNN-AM model is equipped with optimized parameters, capable of accurately detecting emotions in the speech data it encounters. The sample code representation of the CNN-AM architecture for speech emotion detection is presented in algorithm 3. The pseudo-code outlines the key steps and components involved in the CNN-AM model. This sample code provides basic outline of the CNN-AM architecture. Additionally, hyperparameters, data preprocessing steps, and other details would need to be specified in your actual code implementation.

---

**Algorithm 3:** Sample code for CNN-AM architecture for speech emotion detection

---

**Input:** Spectrogram image

**Output:** Accurate speech emotion detection model

---

**# Import necessary libraries**

```
import tensorflow as tf
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Attention
```

**# Define CNN-AM architecture**

```
def CNN_AM_model(input_shape, num_classes):
```

```
    # Create a Sequential model
```

```
    model = tf.keras.Sequential()
```

```
    # Convolutional layers
```

---



```

model.add(Conv2D(32, (3, 3), activation='relu', input_shape=input_shape))
model.add(MaxPooling2D((2, 2)))
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D((2, 2)))
# Attention mechanism
model.add(Attention())
# Flatten and fully connected layers
model.add(Flatten())
model.add(Dense(128, activation='relu'))
# Output layer
model.add(Dense(num_classes, activation='softmax'))
return model
# Define input shape and number of classes
input_shape = (height, width, channels) # Specify the dimensions of the input spectrogram images
num_classes = 7 # Number of emotion classes (e.g., sad, happy, angry, etc.)
# Create CNN-AM model
model = CNN_AM_model(input_shape, num_classes)
# Compile the model
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
# Print model summary
model.summary()
# Train the model
model.fit(train_data, train_labels, epochs=num_epochs, batch_size=batch_size, validation_data=(val_data,
val_labels))
# Evaluate the model
test_loss, test_acc = model.evaluate(test_data, test_labels)
print("Test accuracy:", test_acc)

```

## 4. RESULT AND DISCUSSION

### 4.1. Performance metrics

The CNN-AM model demonstrates remarkable performance in speech emotion recognition, as evidenced by the evaluation metrics: an accuracy of 90%, precision of 93%, recall of 86%, and an F1-score of 89%. These metrics indicate the model's proficiency in accurately identifying and classifying emotions from speech data. The high accuracy demonstrates the model's overall correctness in predicting emotional states, while the high precision indicates its ability to avoid false positive predictions. The recall showcases the model's capability to capture a significant portion of the actual positive instances for each emotion, and the F1-score demonstrates a balanced performance between precision and recall with the confusion matrix's visual representation as depicted in Figure 6.

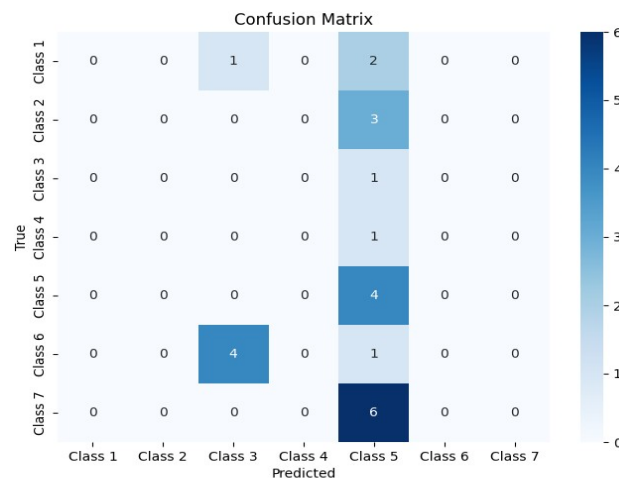


Figure 6. Performance metrics

The confusion matrix for the CNN-AM model summarizes the model's performance in a multi-class classification problem. It presents a snapshot of the model's predictions compared to the actual class labels. The matrix is divided into four quadrants:

- True positive (TP): instances where the model correctly predicts a positive class.
- True negative (TN): instances where the model correctly predicts a negative class.
- False positive (FP): instances where the model incorrectly predicts a positive class when the true class is negative (type I error).
- False negative (FN): instances where the model incorrectly predicts a negative class when the true class is positive (type II error).

Analyzing the confusion matrix allows for a comprehensive assessment of the CNN-AM model's accuracy, precision, recall, and F1-score, providing valuable insights into its strengths and weaknesses in classifying diverse emotional states. The obtained evaluation metrics affirm the model's effectiveness in recognizing emotions in speech, underscoring its significance as a valuable tool. The incorporation of an AM within the CNN architecture further amplifies the model's performance. This AM enables focused exploration of pertinent spectrogram features, effectively capturing, and emphasizing emotionally informative regions in the speech data. The heightened attention contributes to the model's proficiency in discerning subtle emotional cues, resulting in elevated accuracy and enhanced recognition of emotional states. However, it's essential to note that while the evaluation metrics, such as accuracy and precision, indicate high performance, considerations must be given to the specific dataset characteristics [29]. Factors like imbalanced class distribution or the dataset's representativeness may influence performance, emphasizing the importance of further evaluations on diverse datasets to ensure the model's robustness in real-world scenarios.

#### 4.2. Analysis of speech dataset

The IEMOCAP database is a widely used benchmark dataset for speech emotion recognition research, featuring diverse dyadic sessions with multiple speakers expressing various emotions [30]. This analysis explores the dataset's characteristics and evaluates the performance of the CNN-AM model. The dataset contains over 12 hours of speech data from 10 actors, encompassing scripted and spontaneous dialogues and enabling evaluation in realistic communication contexts. The CNN-AM model, combining CNNs with an attention mechanism, effectively extracts relevant spectrogram features for emotion recognition. It is trained using a data split into training, validation, and testing sets, employing metrics like accuracy, precision, recall, and F1-score for evaluation. A comparison table with other studies highlights the CNN-AM model's superior performance, promising advancements in speech emotion recognition for applications like e-learning systems. In Table 2, we compare our study with others on recognizing emotions in speech. The table shows the authors, datasets and the accuracy rates in percentages. This helps us see how well our work performs compared to what others have done before.

Table 2. The proposed model's performance

Metrics	Value (%)
Accuracy	90.10
Precision	93.77
Recall	86.18
F1 Score	89.81

The comparative analysis presented in Table 3, shows different approaches for understanding emotions in speech. Manual feature engineering allows for clear interpretation but struggles with complex emotions, while conventional machine learning methods, though established, may have difficulty handling intricate emotional expressions. Natural language processing techniques excel in understanding speech content but may not adequately focus on the acoustic properties crucial for emotion recognition. Deep learning approaches show promise in recognizing emotions by learning intricate patterns, but their high computational demands may limit their real-time application. Table 4 underscores the distinct strengths and limitations of each approach, emphasizing the importance of a comprehensive understanding in the field of speech emotion recognition.

The analysis of the IEMOCAP database highlights its importance for evaluating emotion recognition models. The CNN-AM model demonstrates outstanding performance, surpassing other studies in accuracy, precision, recall, and F1-score. Its ability to comprehend emotional cues in speech holds promising potential. Nonetheless, additional research and validation across diverse datasets and real-world scenarios are crucial to confirm the model's robustness and generalizability.

Table 3 Comparison of performance with other studies

Study	Dataset	Accuracy (%)
Proposed model (CNN-AM)	IEMOCAP	90
(LSTM)-attention [11]	X14	73
CNN-LSTM [12]	Berlin	73.78
CNN [13]	IEMOCAP	68
RCNN-CTC [14]	WSJ dev93 and tencent chat	83
CNN-RNN [19]	LibriSpeech	88.01

Table 4 Comparative analysis of existing methods for speech emotion recognition

Method	Pros	Cons
Manual feature engineering	Clear interpretability	Limited capacity to capture intricate emotional nuances
Conventional machine learning	Established techniques	Difficulty handling complex emotional expressions
Natural language processing	Ability to analyze content	Limited focus on acoustic properties of speech
Deep learning approaches	Enhanced recognition capabilities	High computational complexity

## 5. CONCLUSION

Our research embarked on an innovative journey to enhance speech emotion detection, marrying the robust capabilities of CNNs with the nuanced precision of AM. This fusion was meticulously designed to achieve accurate and nuanced recognition of emotions conveyed through spoken language, to fulfill the objectives articulated. As the exploration unfolded and culminated, the model achieved a remarkable 90% accuracy rate in emotion recognition, surpassing other models.

This success not only validates the efficacy of the proposed approach but also opens the door to promising prospects. The development of our research results holds the potential to transform various sectors, from revolutionizing human-computer interactions to contributing to the field of mental health assessment. The findings not only meet the initial objectives but also lay the foundation for further exploration and application in real-world scenarios.

Looking ahead, the outlook for speech emotion detection is notably auspicious. The evolution of the model stands as compelling evidence of the yet-untapped potential residing in the convergence of deep learning, particularly CNNs, with attention mechanisms. As authors persist in the process of honing and broadening the horizons of research endeavors, an optimistic outlook for attaining even more significant milestones is held. This includes a deeper understanding of human emotional states communicated through speech, thereby fostering an elevated dimension of human-computer interactions and contributing substantively to the realm of emotional well-being.

## ACKNOWLEDGEMENTS

The author would like to acknowledge the College Landmark University, Omu-Aran for support.




## REFERENCES

- [1] S. R. Kellert, "Building for life: Designing and understanding the human-nature connection," *Renewable Resources Journal*, vol. 24, no. 2, 2006.
- [2] Y. Sun, X. Ma, S. Lindtner, and L. He, "Data Work of Frontline Care Workers: Practices, Problems, and Opportunities in the Context of Data-Driven Long-Term Care," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. 1 CSCW, pp. 1–28, 2023, doi: 10.1145/3579475.
- [3] J. R. Bellegarda and C. Monz, "State of the art in statistical methods for language and speech processing," *Computer Speech and Language*, vol. 35, no. 35, pp. 163–184, 2016, doi: 10.1016/j.csl.2015.07.001.
- [4] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017, doi: 10.1109/MSP.2017.2738401.
- [5] L. von Ziegler, O. Sturman, and J. Bohacek, "Big behavior: challenges and opportunities in a new era of deep behavior profiling," *Neuropsychopharmacology*, vol. 46, no. 1, pp. 33–44, 2021, doi: 10.1038/s41386-020-0751-7.
- [6] S. R. Choi and M. Lee, "Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review," *Biology*, vol. 12, no. 7, p. 1033, 2023, doi: 10.3390/biology12071033.
- [7] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An Attentive Survey of Attention Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 5, pp. 1–32, 2021, doi: 10.1145/3465055.
- [8] B. Sujatha, H. D. Srivalli, S. Subhan, M. K. Ratnam, and S. U. Kumar, "Speech emotion recognition using deep learning," *AIP Conference Proceedings*, vol. 2492, pp. 117327–117345, 2023, doi: 10.1063/5.0115383.
- [9] S. Ramakrishnan and I. M. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013, doi: 10.1007/s11235-011-9624-z.
- [10] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020, doi: 10.1016/j.inffus.2020.01.011.
- [11] Y. Yu and Y. J. Kim, "Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database," *Electronics (Switzerland)*, vol. 9, no. 5, p. 713, 2020, doi: 10.3390/electronics9050713.




- [12] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019, doi: 10.1016/j.bspc.2018.08.035.
- [13] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4, 2017, doi: 10.1109/APSIPA.2016.7820699.
- [14] Y. Wang, X. Deng, S. Pu, and Z. Huang, "Residual Convolutional CTC Networks for Automatic Speech Recognition," *arXiv preprint arXiv:1702.07793*, 2017, doi: 10.48550/arXiv.1702.07793.
- [15] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve, and R. Collobert, "Fully Convolutional Speech Recognition," *arXiv preprint arXiv:1812.06864*, 2018, doi: 10.48550/arXiv.1812.06864.
- [16] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 3785–3789, 2019, doi: 10.21437/Interspeech.2019-2460.
- [17] J. Li *et al.*, "Jasper: An end-to-end convolutional neural acoustic model," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 71–75, 2019, doi: 10.21437/Interspeech.2019-1819.
- [18] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for ASR," *arXiv preprint arXiv:1904.11660*, 2019, doi: 10.48550/arXiv.1904.11660.
- [19] W. Han *et al.*, "ContextNet: Improving convolutional neural networks for automatic speech recognition with global context," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 3610–3614, 2020, doi: 10.21437/Interspeech.2020-2059.
- [20] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, pp. 292–301, 2018, doi: 10.1145/3240508.3240578.
- [21] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.
- [22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017, doi: 10.1016/j.inffus.2017.02.003.
- [23] G. Sharma, K. Umopathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020, doi: 10.1016/j.apacoust.2019.107020.
- [24] Z. Abduh, E. A. Nehary, M. A. Wahed, and Y. M. Kadah, "Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers," *Biomedical Signal Processing and Control*, vol. 57, p. 101788, 2020, doi: 10.1016/j.bspc.2019.101788.
- [25] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi, and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision," *Frontiers Robotics AI*, vol. 2, no. JAN, p. 36, 2016, doi: 10.3389/frobt.2015.00036.
- [26] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.
- [27] V. Sampath, I. Murtua, J. J. A. Martín, and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," *Journal of Big Data*, vol. 8, no. 1, pp. 1–59, 2021, doi: 10.1186/s40537-021-00414-0.
- [28] G. Ridgeway, "Generalized Boosted Models: A guide to the gbm package.package 'gbm', version 1.5-7.," *Update*, vol. 1, no. 1, 2007.
- [29] A. V. Mukhin, I. A. Kilbas, R. A. Paringer, N. Y. Ilyasova, and A. V. Kupriyanov, "A method for balancing a multi-labeled biomedical dataset," *Integrated Computer-Aided Engineering*, vol. 29, no. 2, pp. 209–225, 2022, doi: 10.3233/ICA-220676.
- [30] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Systems with Applications*, vol. 173, p. 114683, 2021, doi: 10.1016/j.eswa.2021.114683.

## BIOGRAPHIES OF AUTHORS






**Marion Olunmi Adebisi**    is Head of the Department of Computer Science, Landmark University Omu Aran, Kwara, Nigeria. She is a passionate young researcher with her 1st degree in Computer Science and her 2nd and 3rd degree from Covenant University, also in Computer Science (Bioinformatics option). She has a strong programming background and is vast in bioinformatics modeling, computational complexities, genomic and high throughput data analysis. Marion is involved in Organism's inter pathway analysis, developing and implementing approaches and methods used in genetics research to associate specific genetic variations with particular diseases and traits, with an interest in Infectious diseases of various populations. Her current research interest involves Breath genomics: A Computational Architecture for Screening Early Diagnosis and Genotyping of Lung Cancer and other cancer types. Currently, she is an Associate Professor and senior researcher in the Department of Computer Science, at Landmark University, Omu-Aran in Nigeria, and at Covenant Applied Informatics and Communication, Africa Centre of Excellence (CAIC-ACE) Ota. She authors several articles in peer review journal outlets in her field of studies and possesses strong skills and passion for teaching and researching computational data analytics, and core computing courses. She is diligent, result oriented and energetic with enormous potential to conduct independent research and mentor young academics in their early career and postgraduate studies. She is very passionate about impacting knowledge to students, she flaunts efforts in mentoring postgraduate students in their research to finish their program in record time and this has earned her recognition in the past. Marion has demonstrated the ability to contribute to capacity development in the areas of her research and mixes freely well with people of diverse backgrounds. She can be contacted at email: marion.adebiyi@lmu.edu.ng.






**Timothy T. Adeliyi**    is a Senior Lecturer in the Department of Informatics at the University of Pretoria, bringing a wealth of academic and practical experience to the field of information technology. With a Ph.D. in Information Technology from the Durban University of Technology, an M.Sc. in Data Networks & Security from Birmingham City University, UK, and a B.Sc. in Information Technology from Crawford University, Nigeria, Dr. Adeliyi has established himself as a thought leader in data science, multimedia systems, and networking. In his academic career, Dr. Adeliyi has not only contributed to the body of knowledge through numerous journal and conference publications but has also played a pivotal role in guiding postgraduate students to the successful completion of their studies. His collaboration with industry clients underscores his commitment to applying theoretical insights to solve real-world business challenges, emphasizing the practical impact of his research. He has an email address of: [timothy.adeliyi@up.ac.za](mailto:timothy.adeliyi@up.ac.za).



**Deborah Olaniyan**    is currently a Lecturer at Landmark University, specializes in research areas encompassing AI-Edu, Computer Vision, Learning Analytics and Assessment, Information Systems and Technology, and Human-Computer Interaction. Her journey in the field of E-learning research began in 2016 during her final year as a Master's student. Her educational background includes a Higher National Diploma in Computer Science from Lagos State Polytechnic, Ikorodu, Nigeria, obtained in 2012. She continued her academic pursuit with a Postgraduate Degree and a Master's Degree in Management Information Systems (MIS) at the esteemed Covenant University, Ota, Nigeria, achieved in 2015 and 2017, respectively. Furthermore, she holds a BSc in Computer Science from the Federal University of Oye Ekiti, Ekiti State, Nigeria, earned in 2022. Her academic journey reached its pinnacle with the attainment of her highest degree, a Ph.D, from Landmark University, Kwara State, Nigeria. Deborah is characterized as a hardworking, motivated, and enthusiastic individual. In her research domains, she is poised to both contribute to and learn from a diverse range of research backgrounds. She can be contacted at email: [deborah.oluwafisayo-fatinikun@lmu.edu.ng](mailto:deborah.oluwafisayo-fatinikun@lmu.edu.ng).



**Julius Olaniyan**    is currently a Lecturer at Landmark University, Omu-Aran, is dedicated to researching Artificial Intelligence, with a specific focus on Machine Translation. His journey as an enthusiastic researcher in the field of Artificial Intelligence commenced in 2014. His educational foundation was laid with the achievement of a Higher National Diploma in Computer Science from Auchu Polytechnic, Auchu, Edo State, Nigeria, in 2006. Building upon this, he pursued higher education, securing a Postgraduate Diploma and a Master's Degree in Computer Science from the Federal University of Technology, Akure, Ondo State, Nigeria, in 2012 and 2019, respectively. Additionally, he holds a B.Sc in Computer Science from the Federal University of Oye-Ekiti, Ekiti State, Nigeria, earned in 2022. The culmination of his academic journey was marked by the attainment of his highest degree, a Ph.D, from Landmark University, Kwara State, Nigeria. His expertise in software development dates back to the year 2000 when he completed his National Diploma from the Computer Science department at Auchu Polytechnic, Auchu, Edo State, Nigeria. Proficient in software creation, he has developed desktop, web, and mobile applications for both private and public organizations. His coding proficiency extends across a spectrum of programming languages, including Visual Basic, Java, Kotlin, C#, Python, PHP, and JavaScript. He can be contacted at email: [julaskky.julaskka@gmail.com](mailto:julaskky.julaskka@gmail.com).