DOI: 10.24996/ijs.2025.66.5.23

ISSN: 0067-2904

# A Hybrid ANN-LSTM Speaker Identification Using Advanced Feature Extraction Techniques

**Maha Adnan shanshool*, Husam Ali Abdalmohsen**

*computer science department, college of science, University of Baghdad, Iraq*

Received: 16/12/2023 Accepted: 30/6/2024 Published: 30/5/2025

**Abstract**

Over the past decades, speaker identification has gained the attention of many researchers and security companies because of its many applications in identifying individuals. Therefore, through this work, a speaker identification system has been designed and implemented. The system undergoes a preprocessing phase that involves the removal of silence, the removal of outliers, the quantization of features, and the extraction of linear predictive coding (LPC) and Mel-frequency cepstral coefficients (MFCC) features. Additionally, the system performs a mean and standard deviation analysis on all features. The third phase involved applying deep learning techniques such as convolutional neural networks (CNN), artificial neural networks (ANN), long-short-term memory (LSTM), and random forests (RF). The proposed work's novel idea is a hybrid architecture, generated from ANN and LSTM. The proposed hybrid speaker identification system exhibits exceptional processing efficiency and achieves a remarkable accuracy rate of 94.63% and 99.2%, respectively. This study makes a substantial contribution to the advancement of speech recognition technologies by highlighting the adaptability and practical value of the hybrid ANN-LSTM model, especially in situations where speed is of the essence. This work was applied to a large dataset that was combined from three different sources: TIMIT, Prominent Leaders, Fluent Speech Command, and the GSCC dataset, which are all comprised of audio files only.

**Keywords:** Speaker Identification, Mel-frequency Cepstral coefficients (MFCC), linear predictive coding (LPC), artificial neural network (ANN), long short term memory (LSTM), Ghadeer speech crowed corpus (GSCC)

# الطريقة الهجينة والميزات المتقدمة لتحديد هوية المتحدث *ANN-LSTM*

**مها عدنان شنشول*,حسام علي عبد المحسن**

قسم الحاسوب, العلوم , بغداد, بغداد, العراق

**الخلاصة**

لقد حظي تحديد هوية المتحدث باهتمام العديد من الباحثين وشركات الأمن خلال العقود الماضية بسبب تطبيقاته العديدة في تحديد هوية الأفراد. ولذلك ومن خلال هذا العمل تم تصميم وتنفيذ نظام التعرف على المتحدثين. يمر النظام بمرحلة المعالجة المسبقة والتي تتمثل في إزالة الصمت وإزالة القيم المتطرفة والتكميم، ومرحلة استخراج الميزات، واستخلاص ميزات التشفير التنبئي الخطي (LPC) ومعاملات ميل التردد الرأسي (MFCC)، وإجراء تحليل المتوسط والانحراف المعياري على جميع الميزات . أما المرحلة الثالثة فكانت تطبيق

*Email: maha.jabr2101m@sc.uobaghdad.edu..iq

تقنيات التعلم العميق مثل الشبكة العصبية التلافيفية (CNN) ، والشبكة العصبية الاصطناعية (ANN) ، والذاكرة طويلة المدى (LSTM) ، والغابة العشوائية .(RF) الفكرة الجديدة للعمل المقترح هي الهندسة المعمارية الهجينة، الناتجة عن (ANN) و .LSTMيُظهر نظام تحديد السماعات الهجين المقترح كفاءة معالجة استثنائية ويحقق معدل دقة ملحوظًا يبلغ 94.63% و99.2%. تقدم هذه الدراسة مساهمة كبيرة في تطوير تقنيات التعرف على الكلام من خلال تسليط الضوء على القدرة على التكيف والقيمة العملية لنموذج ANN–LSTM الهجين، خاصة في المواقف التي تكون فيها السرعة أمرًا جوهريًا. تم تطبيق هذا العمل على مجموعة بيانات كبيرة تم دمجها من ثلاثة مصادر مختلفةTIMIT :، والقادة البارزون، ومجموعة بيانات Fluent Speech Commandومجموعة بياناتGSCC ، والتي تتكون جميعها من ملفات صوتية فقط.

## 1. Introduction

Voice communication is utilized in the modern world to convey ideas and emotions. The majority of commerce is conducted via voice communication, which helps to build trust. Audio is produced by humans through their mouths, throats, and vocal cords, but the structure of the vocal cord modulates the fundamental frequency, which affects the voice frequency, which determines the voice of the human being . Every human voice is distinct from one another by its frequency[1]. Therefore, extracting features related to the frequency domain has been significant in voice identification, such as MFCCs. The MFCC features and the vector quantization approach were used, followed by handling a large volume of data.

The field of speaker identification is divided into two approaches: identification and verification. The goal of speaker identification is to match the incoming voice sample with the available voice samples, whereas the goal of speaker verification is to identify the claimant [2].

The speech of the human being is strongly related through its frequency to many life feelings, such as pain, psychological state, and emotion. Therefore, we created an automatic disease prediction system (ADPS) that uses only the patient's voice to estimate how much pain the patient is feeling, requiring no additional language processing [3]. The psychological state of the human being and its impact on voice were one of the many challenges that were faced. For speaker verification and identification, voice and speaker prints are significant behavioral aspects. To accurately identify speakers, a variety of techniques, algorithms, frameworks, and datasets are employed, depending on a multitude of variables [4]. We used a GMM-CNN classifier to figure out who was speaking and what emotions they were showing, along with a CASA pre-processing module to get rid of noise, since noise is what makes it hard to figure out who was speaking in noisy places [5].

In computer vision and related domains, deep learning, particularly using convolutional neural networks (CNNs), has led to significant advancements in recent years. This development is attributed to the move away from creating features and then separate subsystems to learning characteristics and recognition systems from scratch using almost raw data. But for speaker clustering, custom processing chains like MFCC features and GMM-based models are still frequently used [6].

## 2. Related Work

Through this section, we will mention state-of-the-art research that deployed the MFCC features in general and trained their systems using deep or machine learning from 2018 until now.

In 2018, [7] proposed two contributions. The first was a completely automated pipeline using computer vision methods. Open-source material was used to construct the dataset, collecting movies from YouTube. An active speaker verification system was proposed using a two-stream synchronization Convolutional Neural Network (CNN) and speaker identification using CNN-based facial recognition. The second contribution uses the dataset to apply and compare various cutting-edge speaker identification systems to set baseline performance. The highest performance for both identification and verification is obtained by a CNN-based design, with an accuracy of 92%.

In 2019, [8] introduced the potential application of neural networks to speech identification. The concept of an artificial neuron as an object utilized in speech identification was defined, and typical approaches to speech recognition were specifically taken into consideration. The use of a neural network for speech recognition was explored, and methods for carrying out this task were provided. With a score of 92.1%, accuracy utilizing neural networks with little training data and a high i-vector dimension outperforms the competition.

In 2019, [1], a novel combination of ANN and support vector machine (SVM) classifiers using MFCC, LPC, and zero crossing rate (ZCR) as feature extraction algorithms was proposed. There are 640 voices overall that are used as input for 20 speakers. There are thirty-two distinct terms used by each speaker. The best accuracy of 93.1 was reached with the ANN method. In 2021, [9] proposes comparison studies of voice recognition systems that use machine learning techniques. The MFCC feature and the differential of accuracy were deployed, and 7 classifiers were used for classification. The study showed that using the RF machine learning classifier increased the accuracy, reaching 97.9%, and was superior to other classifiers.

In 2022, [2] proposes speaker recognition based on machine learning algorithms. Support Vector Machine (SVM) and Random Forest (RF) models are used with statistical features and Mel-Frequency Cepstral Coefficients (MFCC). A new voice dataset was collected for training and evaluating speaker recognition models. The performed experiments showed that SVM achieved 94% identification accuracy, and the feature selections are RFE, MRMR, and CHI-2.

In 2022, [9] proposed a deep learning technique to develop an automated speaker detection system. On both the open and closed sets of the TIMIT and LibriSpeech datasets, CNN and LSTM algorithms have been used.

In this work, the proposed system consisted of three steps: pre-processing the audio, feature extraction using Mel frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC), and finding the mean and standard deviation for each audio. The new approach uses hybrid ANN-LSTM to obtain the best accuracy rate with low time consumption.

The current state-of-the-art limits its capabilities to a specific language or accent. A lot of research on speaker identification makes use of certain datasets that may not be very diverse in terms of languages spoken or accents. To achieve generality, we used three types of datasets covering three different accents and merged them to form one general dataset, which was treated as a single test bed. In order to achieve generality, we had to capture multiple characteristics of the voice signal. Robust speaker identification frequently requires a mix of features. In our research, we used different types of features and combined them to form a single feature vector.

## 3. Classification methods

Some of the following effective machine-learning algorithms will be utilized in our suggested system, while others will be employed for comparison.

### 3.1 ANN (Artificial Neural Network)

Artificial neural networks, or ANNs, are a fundamental idea in machine learning and artificial intelligence. They consist of an input layer, one or more hidden layers, and an output layer that is made up of interconnected nodes. During training, the network learns to adjust the weights assigned to each connection between nodes[10]. Artificial neural networks (ANNs) use activation functions to give the model nonlinearity and enable it to understand complex relationships in the data. Typically, training entails using algorithms such as gradient descent for forward propagation, which involves making predictions, and back propagation, which involves changing weights based on errors, as mentioned in figure 1.

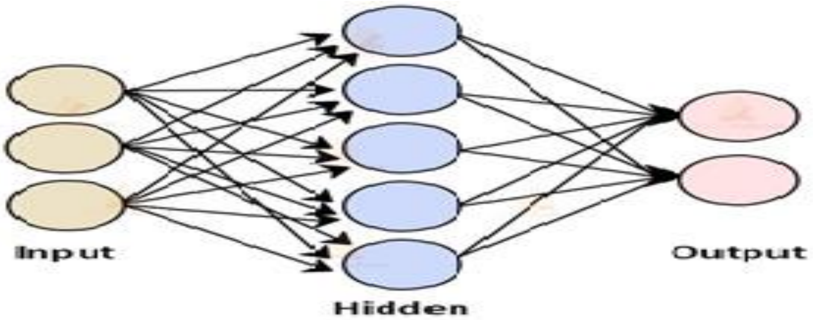

**Figure1:** Architecture of Ann [1]

### 3.2 Memory for Long Short Term (LSTM)

One typical problem with classic RNN models is that they become less effective at deriving context from time steps of states that are much farther in the past as the time step increases. Long-term reliance is the name given to these phenomena. The recurring nature of a standard RNN, as well as the network's deep layers, contribute to the frequent occurrence of exploding and vanishing gradient problems[11, 12]. Furthermore, memory cells with multiple gates are placed in a hidden layer to introduce the LSTM model[13]. In order to solve this issue, the block of a hidden layer with an LSTM unit is shown in Figure 2.

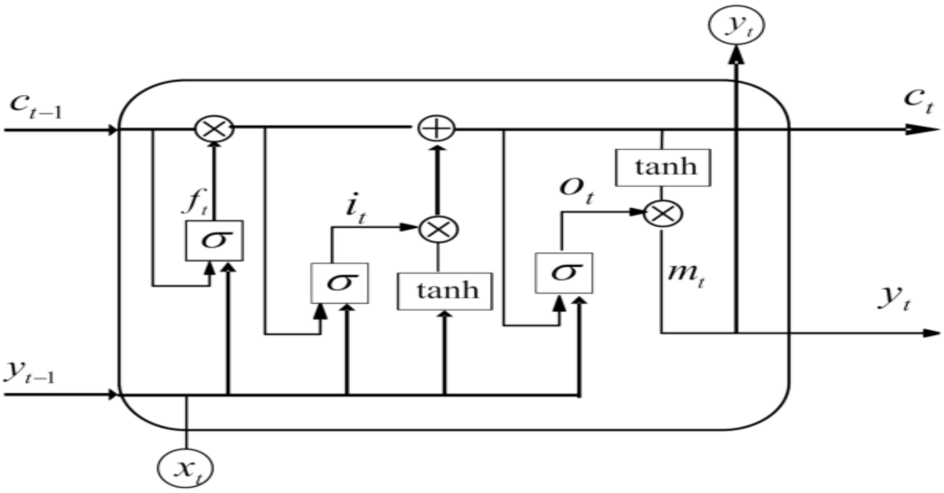

**Figure 2 :**Long short-term memory (LSTM) architecture [14]

## 4. Proposed System Architecture

In the proposed system, several stages will be used to identify the speaker’s voice, which includes performing pre-processing on the audio data, such as deleting silence from the beginning and end of the signal, as well as deleting outliers and quantizing the data. After pre-processing, the features are extracted, namely mfcc, LPC mean, and STD. Then the classification phase begins, obtaining the results as mentioned in figure 3.

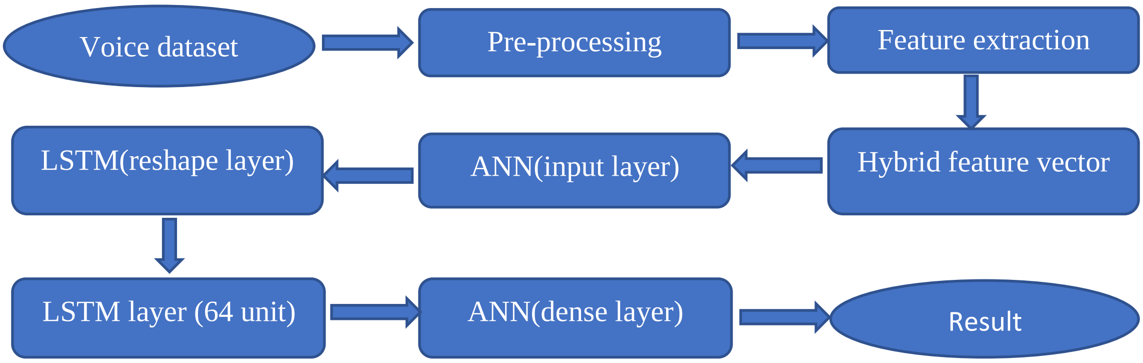

**Figure 3**: The block diagram of the ANN+LSTM

### 4.1Datasets

The dataset generated in this work was generated by merging three datasets, which are TIMIT, prominent, and fluent speech, by taking 728 speakers from all three datasets. 10 sentences for each speaker were taken randomly. The length of the audio clips ranges from 1 to 4 seconds. The language used is English with different accents. The final number of samples is 7280 at a sample rate of 16 kHz. We divided the data set into 80% for training, 582.4 samples entering the training phase, and the remainder for testing. Merged datasets aim to obtain a large number of audio files.

### 4.1.1TIMIT

The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers from eight major American English dialects, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic, and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance [6].

### 4.1.2Prominent Dataset

The voice dataset (Prominent Leader's Talks) was utilized, which included audio snippets of five well-known leaders both domestically and internationally. A certain amount of noise is included in it, which allows the crowd's cheers to be audible. The Kaggle website provided the dataset [15, 16].

### 4.1.3Fluent Speech Command

The dataset contains 97 speakers saying 248 different phrases. The 248 utterances map to 31 unique intents that are divided into three slots: action, object, and location. The goal in preparing this dataset was to provide a benchmark for end-to-end spoken language understanding models [5].

Three types of data sets were used and combined by taking the number of speakers, and 10 sentences per speaker were randomly taken. Audio clips range in length from 1 to 4 seconds.

The language used was English, and the final number of samples was 728, while the average frequency was 16,000.

As shown in Table 1, the datasets used in this paper were downloaded from https://www.Kaggle.com/datasets. And https://www.data.mendeley.com/datasets.

### 4.1.4 GSCC (Ghadeer-Speech-crowed-corpus)

GSCC is the first date set to contain two language solo recordings and a recorded recording. Furthermore, it includes additional details about each speaker, such as their height, weight, age, and gender. At Baghdad University, the data collection process involved recording the voices of various student groups from all departments within the GSCC. Containing 210 speakers, 105 female and 105 male, this data set was divided into two groups: first, each speaker spoke 9 individually defined sentences in Arabic and English. In the second group, a mixture of speakers created a sense of crowding in the speech. The crowding began with two speakers mixing, then three, then four, and finally five speakers speaking at the same time[17].

**Table 1:** The datasets were used for merging and GSCC dataset

| Input file data | name of dataset | file format | file size |
|---|---|---|---|
| Voice recognition | prominent leaders speech | wave | 16khz |
| | TIMIT | wave | 16khz |
| | Fluent | wave | 16khz |
| | GSCC | wave | 44khz |

### 4.2Preprocessing

Numerous processes were applied during this stage. To begin, there is the remove silence step. We remove silence from the start and end of the audio signal. The second process is to eliminate outliers from all signals. These preprocess operations are applied to improve the quality of audio files and Apply quantization to bit numbers as mentioned in figure 4.

```
Algorithm (1): The preprocessing Algorithm
Input: (audio files paths, Silence removal threshold, Outlier removal threshold, and Number of quantization bits)
Output: Remove silence and outliers from audio files
begin
Step 1: Load audio file
Step 2: Remove silence from the audio using the remove silence function.
Step 3: Remove outliers from the trimmed audio using the remove outliers function. Quantize the audio using the quantize audio function.
Step 4: Plots of the original, altered, and eliminated outlier audio signals are displayed.
End
```

The provided input outlines a multi-step audio signal processing method that includes processes like silence removal, outlier removal, and audio quantization. This pipeline is critical for tasks like audio denoising or preprocessing before further analysis. In the initial step, the audio files are loaded, creating the framework for the subsequent procedures. As stated in Step 2, in order to increase signal clarity, all pauses and background noise must be eliminated. The third step is to remove outliers, which could be anomalies or aberrations in the audio data. Another step in this process is quantization, which involves reducing the precision of the audio data for storage or compression purposes. Step 4 enables a final analysis by displaying the original audio alongside the modified and outlier-removed versions. We compare the signal's changes throughout the process. All things considered, this pipeline demonstrates a systematic way to enhance audio quality and prepare it for further analysis or use.

### 4.2Feature extraction

At this stage, important features are extracted. Feature extraction is the process of taking specific mathematically determined data and extracting it from the source signal [13]. In this study, we used 40 MFCC coefficients [14][18] and 12 LPC coefficients .in addition to the use of statistical characteristics essential for simulating the audio signal's spectrum properties. Furthermore, the retrieved audio contains statistical qualities such as the mean and standard deviation. This produces a hybrid feature vector that forms a comprehensive representation of the audio signal by encapsulating both statistical and spectral information, as shown in Figure 5.

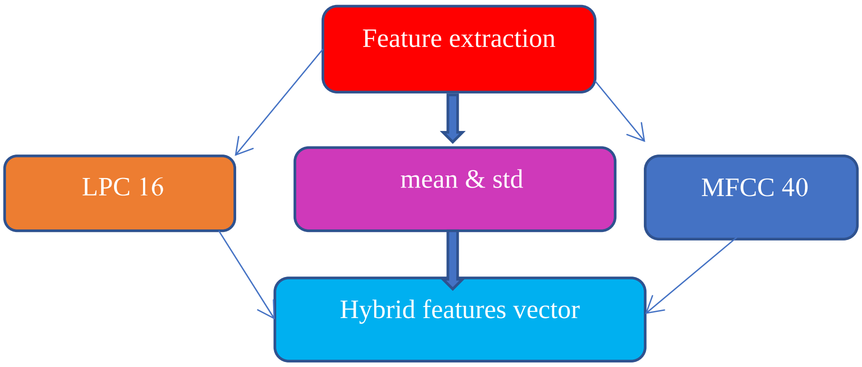

**Figure 5**: Block diagram features the extraction process

### 4.2.1 MFCC (Mel Frequency Cepstral Coefficients)

MFCCs are a popular technique for extracting features from speech and audio signal processing. By expressing audio signals in a manner better suited to machine learning, in this research, we used the 'librosa.feature.mfcc' [19] function to extract 40 mfcc from an audio signal.

### 4.2.2 LPC(Linear Predictive Coding)

LPC is a technique for simulating the spectral envelope of an audio signal that is often used for speech synthesis and analysis[20]. The audio signal's linear predictive model is represented by LPC coefficients. This work used the function 'compute_lpc' to calculate the LPC coefficients using Scipy's method '1filter'.

### 4.2.3 Feature Combination

Statistical metrics (standard deviation and mean) are computed for both MFCC and LPC features once they have been extracted. To create a single feature vector, the mean and standard deviation of the MFCC and LPC coefficients are concatenated. The machine learning model uses the generated feature vector as its input data. All things considered, the algorithm produces a feature vector that captures the spectral and statistical properties of audio signals by merging the feature combinations for the standard deviation and mean of MFCC and the standard deviation and mean of LPC. Using this feature vector, deep learning and machine learning models are trained to categorize audio signals into the many groups that the labels represent.

```
Algorithm(2): a combination of feature extraction steps
Extract mfcc
Input: path audio file , sr(sample rate , n_mfcc=40, hop_length=512, n_fft(Fourier transfer)=1024
Output The features vector that combine the standard deviation and mean of MFCC and standard deviation and mean of LPC.
Begin
Step1: Load audio file (audio) with librosa.
Step2: Calculate hop length based on overlap.
Step3: Apply the Hamming window to the audio using apply hamming window
step4: Extract MFCC features from audio using extract mfcc function
Step5: Calculate the mean of MFCC features (mfcc_mean)
Step6: Calculate autocorrelation sequence (auto_corr) for the audio
Step7: Window the autocorrelation sequence (windowed auto_corr)
Step8: Calculate LPC coefficients using Levinson-Durbin algorithm
Step9: Calculate mean and  standard deviation of the audio
Step10: Combine audio features (MFCC mean and LPC coefficients) with statistics Store the hybrid features in the variable hybrid_features
End
```

Step 1: Use the librosa library to load the audio file.
Step 2: The hop duration controls how long it takes for an audio frame to follow one another. We found that the audio has a sample rate of 16000 Hz and an overlap percentage of 0.5.

Step 3: The Hamming window is one type of window function that is used to create audio frames before spectral analysis. It reduces spectral leakage and improves spectrum analysis precision.
Step 4: Mel-frequency cepstral coefficients (MFCCs) are features that are widely used in speech recognition and other audio processing applications. They capture the spectral characteristics of the audio signal. In order to improve accuracy, we came up with 40 coefficients for MFCC following trials 13 and 20.

In this study, we present a new approach, a hybrid method called ANN-LSTM, to obtain better accuracy in low time on merged datasets and applied systems on sets of learning algorithms. This work proposes a combination of MFCC, LPC, mean, and standard deviation as a feature extraction technique with a new hybrid ANN-LSTM. The overall architecture of ANN-LSTM networks is made up of many layers that are merged. This layer consists of an input layer, a layer to reshape the input for the LSTM layer, and an LSTM layer from the Keras library. This layer represents the type of RNN (recurrent neural network) that is good at learning long-term dependencies. The layer consists of 64 neurons and is fully connected, with softmax activation occurring in the dense layer. It generates the network's final output. When compiling models, use the Adam Optimizer and the sparse categorical cross-entropy loss function.

```
Algorithm(3): Hybrid ANN+LSTM
Split the dataset into 80 % training sets and 20 % testing sets.
Begin
Input :( ANN input) starts with an input layer for the features and save weights of ANN.
Step1.Reshaping for LSTM.
Step2.LSTM layer is added with 64 units and returns sequences=False, meaning it returns the last output of the sequence.
```

**Step3**.A dense output layer with softmax activation is added for classification.
**Step4**.The model is compiled with the Adam optimizer and sparse categorical cross-entropy loss.
**Step5**.Model Training The model is trained on the training data.
**Step6**.Evaluation use evaluation metrics (accuracy) are computed.
**End.**

In the beginning, the dataset is divided into separate training and testing sets, with 80% of the data being used for training and 20% being held aside to assess the model's performance on untested data. Subsequently, the model architecture starts with an input layer that is made to fit the dataset's features and save weights. After that, data reshaping is used to make the data fit the LSTM layer's input specifications, which is essential for efficiently processing sequential data. Next, the 64-unit LSTM layer is added to the model architecture, with the requirement that it return just the final output in the sequence. After that, a dense output layer is added, which makes use of softmax activation to help in classification. The compilation stage sets up the training procedure after the model architecture is constructed. The sparse categorical cross-entropy loss function is chosen because it works well for multi-class classification tasks with sparse labels, whereas the Adam optimizer is used because it is good at optimizing complex models. After compilation, the model is trained using the training dataset, going through 300 iterations to improve its performance and fine-tune its parameters.

### 4.3 Experiment results and Comparison with state-of-the-art

When applied to the GSCC dataset, the suggested technique demonstrated strong performance, with an accuracy rate of 94.63% and a high accuracy of 99.2%. It combines an artificial neural network (ANN) with an LSTM model. This model uses a large set of features, such as statistical measurements like mean and standard deviation, as well as Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC). Three separate datasets are combined to provide a more varied and rich training environment, which improves the model's generalization capabilities across various data distributions. Essential components of the audio signals are captured by the integration of MFCC and LPC characteristics, and the ANN-LSTM architecture effectively processes the sequential data to learn intricate temporal patterns. The difficulties of audio signal analysis are well addressed by this hybrid approach, which offers a highly accurate and dependable system that shows promise for a range of applications. Table 3 provides a summary of the accuracy rates and methodology applied in multiple investigations conducted over a variety of years. 2019 demonstrated the ANN's ability to recognize complex patterns. The ability of a convolutional neural network (CNN) to automatically learn hierarchical features was demonstrated in 2021 with an accuracy of 90.82%. This capability is particularly helpful for applications that use images. The following year, 2022, saw the investigation of a combined method using CNN and Long Short-Term Memory (LSTM), which yielded accuracy rates of 79.05% and 66.83%, respectively. The intention behind positioning both systems next to each other was to benefit from each other's advantages in handling intricate patterns and sequential dependencies. Additionally, in that same year, a Random Forest (RF) model achieved an accuracy of 83%, demonstrating the versatility and effectiveness of ensemble learning strategies. These results highlight the fluidity of machine learning techniques and the significance of choosing models that are customized to the particulars of the given task.

**Table 2:** Experiment results and comparison study

| Ref | Year | Dataset | Feature extraction | Method | Accuracy |
|---|---|---|---|---|---|
| Proposed method ANN-LSTM | 2024 | Three datasets merged<br>, GSCC | Mfcc,lpc,mean and std | ANN-LSTM | **94.63%**<br>**99.2 %** |
| [1] | 2019 | 20 speaker (640 voice) | MFCC, LPC, ZC | ANN | **93 %** |
| [7] | 2021 | Voxceleb | GRU-CNN | CNN | **90.82%** |
| [4] | 2022 | Timit, librispeech | MFCC | CNN<br>LSTM | **79.05 %**<br>**66.83 %** |
| [2] | 2022 | 150 speakers with a total of 3,000 data samples and about six hours of speech. | MFCC | RF | **83 %** |

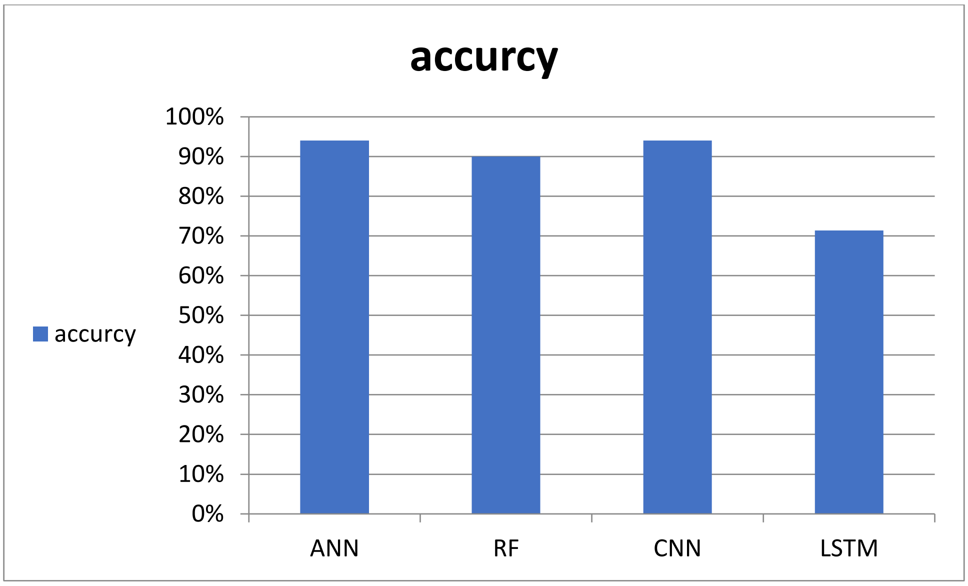

Figure 6: Comparison methods for the proposed system

## 5 .Discussion

The performance comparison of several models applied to diverse datasets is shown in Figure 6. Interestingly, the Artificial Neural Network (ANN) model showed an astounding 94% accuracy when applied to the combined dataset. In comparison, the Random Forest (RF) model's accuracy was just slightly lower at 90%. The Convolutional Neural Network (CNN) achieved outcomes that were exactly the same as the Annular Neural Network (ANN) at 94%. The Long Short-Term Memory (LSTM) model did worse than the others, with an accuracy of 71.4%. The inherent benefits and drawbacks of different model architectures could account for this discrepancy. LSTMs excel at handling sequential data, while ANNs and CNNs excel at spotting intricate patterns in structured and image data, respectively. A unique approach for the hybridization of ANN and LSTM is proposed in light of these findings. The best elements of both architectures will be carefully combined so that the model can more effectively recognize patterns in sequential as well as organized data. The hybrid ANN-LSTM mode achieves an astonishing 94.63% accuracy and 99.2% accuracy, matching the performance of the individual models. The enhanced accuracy of the hybrid model can be attributed to the complimentary nature of ANN and LSTM. While the ANN component excels at spotting intricate patterns in the combined dataset, the LSTM component helps the model better understand sequential dependencies in the data. The result is a more precise and dependable forecasting model. The hybrid model selection becomes particularly significant when the dataset contains elements of both sequential and structured features. By fusing the benefits of ANNs and LSTMs, the hybrid model reduces the shortcomings of the individual architectures and creates a prediction model that is more sophisticated and precise. The experimental results indicate that the hybrid ANN-LSTM model performs more precisely than the models used alone. This outcome highlights how important it is to adjust the model architecture to match the specific dataset attributes. Future research endeavors could concentrate on refining the hybrid model and exploring its applicability in many domains.

## 6 .Conclusions

To clear things up, our study demonstrates the efficacy of a hybrid speaker recognition system that combines artificial neural networks (ANN) and long short-term memory (LSTM) networks. This method produces impressive accuracy rates when it comes to speaker identification from audio recordings. Our method significantly improves speech recognition technologies by utilizing features discovered through Mel-frequency cepstral coefficients (MFCC) and linear predictive coding (LPC), as well as mean and standard deviation analysis. The suggested hybrid architecture is very well suited for real-time identification scenarios because it not only improves accuracy but also provides outstanding processing efficiency. Our model has a remarkable accuracy rate of 94.63% and 99.2%, indicating that it has great potential for real-world use. Furthermore, the model's resilience and flexibility are demonstrated by the utilization of a varied dataset, which includes audio files from TIMIT, Prominent Leaders, Fluent Speech Command, and GSCC. This work makes a substantial contribution to the field of speaker identification, demonstrating the effectiveness of hybrid deep learning architectures and opening the door for further developments in speech recognition technology.

## References

.

**[1].** N. Chauhan, , T. Isshiki, and D. Li, "Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database," *IEEE 4th International Conference On Computer and Communication Systems (ICCCS),* 2019.

**[2].** B.A. Alsaify, , H. S. Abu Arja, B .Y. Maayah, M.M. Al-Taweel, R. Alazrai, and M.I. Daoud, "Voice-Based Human Identification using Machine Learning", *13th International Conference on Information and Communication Systems(ICICS),* pp. 205-208, 2022.

**[3]**. H.A,Abdulmohsin, "*Automatic health speech prediction system using support vector machine*". *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*. 2022. Springer.

**[4].** O, Mamyrbayev,.et al., "Voice Identification Using Classification Algorithms". 2020.

**[5]**. A.B,Nassif,et al., *"CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions".* **103**: p. 107141.2021.

**[6].** Y,Lukic, et al. "Speaker identification and clustering using convolutional neural networks". IEEE 26th international workshop on machine learning for signal processing (MLSP). 2016.

**[7].** .A., J.S, Nagrani, Chung, and A.J.a.p.a. Zisserman. "Voxceleb: a large-scale speaker identification dataset". 2017.

**[8].** O.Z ,Mamyrbayev, et al." *Voice verification using I-vectors and neural networks with limited training data".* p. 36-43.2019.

**[9].** N. N. Prachi, F.M. Nahiyan, Md. Habibullah and R. Khan, "Deep Learning Based Speaker Recognition System with CNN and LSTM Techniques," *Interdisciplinary Research in Technology and Management (IRTM)*, p.6, Nov. 20, 2022. DOI: 10.1109/IRTM54583.2022.9791766.

**[10]**. R., M.J.I.J.o.I, Dastres. *"Artificial neural network systems"*.vol. **21**(2): p. 13-25.2021.

**[11]**. Z.C., J, Lipton, Berkowitz, and C.J.a.p.a. Elkan, *"A critical review of recurrent neural networks for sequence learning"*. 2015.

**[12]**. H.S.,Abdullah, N.H. Ali, and N.A.J.I.J.o.S. Abdullah,*" Evaluating the Performance and Behavior of CNN, LSTM, and GRU for Classification and Prediction Tasks"*. p. 1741-1751. 2024.

**[13].** M.,Mansoor, and M. Al Tamimi, *"PLAGIARISM DETECTION SYSTEM IN SCIENTIFIC PUBLICATION USING LSTM NETWORKS"*. Pages 17-24.*vol.14.2022.*

**[14]**. Ye, F. and J. Yang, "A deep neural network model for speaker identification". Applied Sciences, 11(8):**p**. 3603.2021.

**[15].** A.T. Ali, H.S. Abdullah and M. N. Fadhil, "Speaker Recognition System Based on Mel Frequency Cepstral Coefficient and Four Features," *Iraqi journal of computers, communications, control and systems engineering*, vol. 21, no. 4, pp. 82-89, Dec. 2021. DOI: 10.33103/uot.ijccce.21.4.8.

[**16].** A.A. ,Tahseen, Abdullah, H. S., and Fadhil, M. N.," *Voice recognition system using machine learning techniques"*. Materials Today: Proceedings, 2021.

**[17]**. G.Q., Ali, H.A.J.F.P. Abdulmohsin, and Applications, "*Speaker Identification in Crowd Speech Audio using Convolutional Neural Networks"*. **16**(2). 2024.

**[18]** . A,Mahmood,. and K.J.A.i.A.I.R. Utku.**"** *Speech recognition based on convolutional neural networks and MFCC algorithm"*. 1(1): p. 6-12.2021.

**[19].** B. McFee, C. Raffel, D. Liang, D. P.W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, " librosa: Audio and music signal analysis in python,*" Proceedings of the14th Python in Science conference*, vol. 8, pp. 18-25, 2015.

**[20]**. M.M, Kabir, et al.,*" A survey of speaker recognition: Fundamental theories, recognition methods and opportunities"*. **9**: p. 79236-79263.2021.