Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Voice Recognition Based on Machine Learning Classification Algorithms: A Review

Hazheen Sarbast Mahmood [1], Adnan Mohsin Abdulazeez [2]
hazheen.sarbast@auas.edu.krd, adnan.mohsin@dpu.edu.krd
[1]Akre University for Applied Science- Technical College of Informatics- Akre- Department of Information Technology , Duhok, Kurdistan Region, Iraq.
[2]Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq.

Article Information

Submitted : 8 Jun 2024
Reviewed : 13 Jun 2024
Accepted : 30 Jun 2024

Keywords

Voice recognition, machine learning, classification algorithms, support vector machine, k-nearest neighbors, multilayer perceptrons, random forest, feature extraction, evaluation metrics, graphical user interface.

Abstract

One essential component of biometric identity is voice recognition technology, which uses speech pattern analysis to authenticate people. With an emphasis on machine learning classification techniques, this review article thoroughly examines the field of speech recognition. We examine the effectiveness of random forest (RF), multilayer perceptrons (MLP), k-nearest neighbours (KNN), and support vector machine (SVM) classifiers via painstaking analysis and empirical evaluation. Utilizing a collection of Sepedi speech audio files, our results demonstrate the remarkable accuracy of 99.86% that RF is capable of producing. Aside from visual aids for better understanding, assessment indicators like as accuracy, precision, recall, F-measure, and root mean square error (RMSE) clarify the effectiveness of the model. The research highlights how machine learning algorithms, especially reinforcement learning (RF), have the capacity to revolutionize speech recognition technology in a variety of contexts.

A. Introduction

One useful bio-feature identification technique is voice recognition. Its goal is to identify a person by listening to their recorded voice[1]. Numerous industries, including multi-factor authentication, banking, voice dialing, machine control, and safe access to highly classified regions, have used biometric recognition technology. A speech recognition system's objective is to transform the voice waveform into a parametric representation, which is then processed and utilised as the input data for a variety of recognition models and techniques
that are developed to meet the requirements of the system [2].

The science of artificial intelligence had a significant change in the 1980s, and voice recognition research has advanced significantly [3]. Feature extraction, model construction, model scoring, and data pre-processing were the steps involved in the procedure at that time. Several models and applications that have been created throughout time have their foundations in the earlier phases [4].

Voice recognition technology has been incredibly successful in recent years due to its increased affordability, dependability, and software's enrichment with machine learning techniques. These techniques have improved the software's efficiency in automated data analysis, pattern identification, and processing algorithms that improve the decision-making process automatically through learning [5]. In addition to recently developed cutting-edge feature extraction methods like Mel-Frequency Cepstral Coefficients (MFCC). The key benefit of the MFCC algorithm, which is a popular strategy for extracting speech features, is that it can create robust features even in the presence of noise in the signal and is effective at reducing errors [6, 7].

Speech-based systems may be used for a wide range of tasks, including control applications, speech recognition, voice verification, and much more. A smart home system that enables voice commands to control functions like reading the news and turning on and off lights and fans is suggested by Triyono et al.[8]. Alsaify et al. [9] and other researchers concentrated on the voice recognition issue. The primary goal was to assess the voice recognition capabilities of a general Gaussian Mixture Model (GMM) classification system. However, some academics use speech recognition software for health-related studies [10].

The phenomena of the human voice are largely reliant on the voice that creates it. Research demonstrates that no two people have precisely the same voice; in contrast to the signal qualities that allow for segment identification, the auditory features that distinguish the Voices are vague and hard to pin down [11]. The reviewed authors state that there are three primary sources of variation among Voices: (1) differences in speaking styles (including accents), (2) differences in vocal tract forms and vocal cords, and (3) differences in how Voices convey a message through the words or phrases they use. Since the third source—a voice's propensity to employ particular phrases, words, and syntactic structures—is hard to measure or control in an experiment, automatic voice recognition systems rely on the first two—low-level acoustic features of a speech signal—instead[12]. Signal processing's voice recognition is a hot issue with many of uses, particularly in security systems. Voice recognition plays a major role in voice-activated devices and systems [13]. Several uses of voice recognition technology include forensics,

remote computer access, client verification for financial transactions, and security control for sensitive data.

The primary goals of this review are to: To recognize the most recent developments in speech recognition technology and how they are used. To evaluate how well various feature extraction techniques, including MFCC, enhance speech recognition precision. To investigate the applications of speech recognition across a range of industries, such as smart home systems, healthcare, and security. To outline the primary obstacles to speech recognition technology and suggest possible lines of inquiry for further study.

The remainder of this review is structured as follows: section 2, methodology: This section describes the methodology used in the review, including information on the dataset that was utilised, strategies for detecting speech activity, and approaches for extracting features. The study's classifier models and performance assessment measures are also covered. section 3, summary of speech and voice recognition studies: A thorough synopsis of the research included in the literature review table is given in this section. It summarizes the most important discoveries, patterns, and restrictions found in many research studies on several facets of speech recognition technology and its uses.

Section 4, Discussion: The main conclusions, patterns, and revelations from the literature study are condensed and examined in this section. It summarizes the data from the preceding part and draws attention to the similarities, advantages, and drawbacks found in various investigations. Based on the studied literature, it also provides insights on the general status of speech recognition technology. Section 5: Conclusion: Based on the evaluated papers, this section offers prospective future research paths while also summarizing the review paper's general results and commenting on the advancements achieved in speech recognition technology. It offers a cogent conclusion to the study, highlighting the importance of speech recognition technology and the need for further advancements and research in this field.

- Background

I. Fundamental tasks of Voice Recognition

As seen in Fig. 1, voice recognition involves two primary tasks: voice identification and voice verification. The process of identifying if an unknown voice is indeed coming from a certain enrolled voice is known as voice verification. The voice must assert their identity, and the system verifies that claim. Voice verification is used in calling cards, online login, mobile phone fraud protection, and banking over the phone [14]. The task of voice identification is to pair an unfamiliar voice with one of the registered Voices. In this instance, the Voice gives a sample of their voice (without revealing their name), and the algorithm identifies which of the enrolled Voices the sample belongs to. Two possible applications of speech recognition technology are intelligent answering machines that can recognize individual callers and automatically tag recorded meetings for voice-dependent audio indexing.

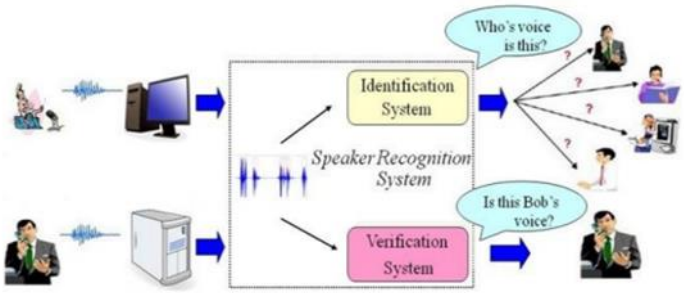

Figure 1. Voice Recognition fundamental tasks (Verification and Identification) [15]

II. Classification of Voice Recognition Systems

Another way to classify speech recognition systems is according to the constraints placed on the spoken text input. Some systems are text-dependent, while others are text-independent. Each Voice in the text-dependent instance uses a set word or phrase for testing and training purposes. Telephone-based services and access control are two examples of user-dependent services that primarily use text-dependent speech recognition technologies. In a text-independent setting, the training and testing text or phrase is not static [16]. The most versatile and often used text-independent speech recognition systems are those that deal with forensic analysis and surveillance operations, when voices could be considered uncooperative due to a lack of desire to be recognized. The recognition performance of text-dependent recognition is superior to that of text-independent recognition. Nonetheless, since text-independent recognition offers such versatility, developing text-independent recognition systems is becoming more and more popular.

III. Phases of Voice Recognition

As seen in Fig. 2, there are two separate steps in developing a speech recognition system: training and testing. A voice is captured during the training phase, and certain audio feature vectors are taken out to create a distinct model (voice-print) that may be used to identify the voice. Testing, often called the recognition phase, involves comparing the provided voice sample to the previously built model.

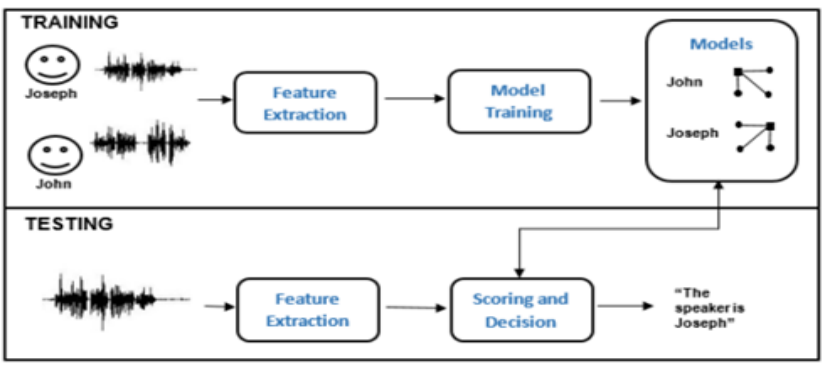

Figure 2. Voice Recognition Training and Testing[17]

B. Research Method

Fig. 3 depicts the methodical process of the suggested voice recognition system. In order to ascertain if an input speech signal contains speech or not, the voice activity detection technique is used [18]. Next, the audio feature vectors that were obtained are used to train four popular machine learning algorithms: support vector machine (SVM), k-nearest neighbours (k-NN), multilayer perceptrons (MLP), and random forest (RF) [19]. All of the models are trained and evaluated in the Waikato Environment for Knowledge Analysis (WEKA) using a 10-fold cross-validation in order to determine which classifier is the best. To consequently recognize the best-performing calculation with its ideal hyper-boundaries, Auto-WEKA was likewise utilized. RF was picked via Auto-WEKA as the ideal classifier model [20].

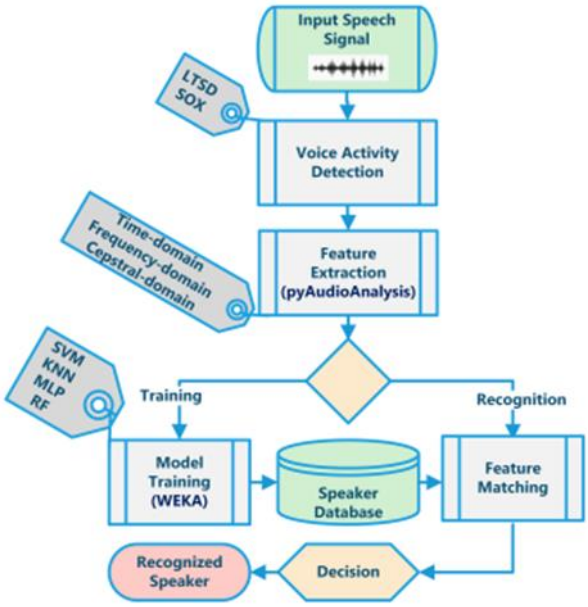

Figure 3. Detailed procedure for the planned speech recognition system [21]

i. Dataset

The Public Community for Human Language Innovation (NCHLT) drive of the Language Asset the executives Organization gave the dataset used in this examination [22]. The sound accounts of a few voices from Sepedi discourse make up the information. A dataset of 60 Voices, each with 150 sound examples remembering around three to five words for every sound document, was used to test the information. As displayed in Table 1, a sum of 7000 sound documents was used, coming about in a runtime of 385.7 minutes.

Table 1. Review of the 150-Speaker Audio Library [23]

Unit	Value
Quantity	617 MB
Time Spent	385.7 minutes
Occurrences	7000

ii. Detection of Voice Activities

In discourse handling, voice action recognition is a technique used to decide if discourse is available in sound or not [24]. To avoid bias in the training, this method isolates audible speech by filtering out background noise [25]. For this work, a noise reduction script from SOX1 was combined with the Long Term Spectral Divergence (LTSD) algorithm. An utterance is divided into overlapping frames by the LTSD algorithm, which then assigns a score to each frame based on the likelihood that vocal activity is present. Next, the likelihood is put together to extract every period that has voice activity.

iii. Feature Extraction

There are several distinguishing characteristics in the human voice that may be utilised to recognize Voices. One of the most crucial parts of voice recognition is feature extraction, which creates a vector that depicts the spoken signal. PyAudioAnalysis is an open-source, feature-extraction software written in Python [26]. PyAudioAnalysis carries out thirty-four short-term characteristics in total. Straight from the unprocessed audio samples, we derive the time-domain parameters Zero Crossing Rate (ZCR), Energy, and Entropy of Energy. The frequency domain attributes (Spectral Spread, Spectral Centroid, Spectral Flux, Spectral Entropy, Spectral Rolloff, Chroma Deviation, and Chroma Vector) are computed using the magnitude of the Discrete Fourier Transform (DFT) [27]. Last but not least, cepstral-domain features (MFCCs) are the result of inverting the logarithmic spectrum using the Inverse DFT. Figure 4 displays the temporal domain (ZCR), frequency domain (Spectral Centroid), and cepstral domain (MFCCs) features retrieved from a single spoken audio sample.

MFCCs are well-liked audio characteristics that are taken out of speech signals and used extensively in voice and speech recognition activities [28]. MFCCs are seen as representing the filter (vocal tract) in the source-filter model of speech. A psychoacoustically oriented filter bank is used to assist estimate MFCCs, which are then calculated using the discrete cosine transform and logarithmic compression. Applying the formula to the case where an M-channel filter bank generates Y (m), where m ranges from 1 to M, we may identify the MFCCs.

$$c_n = \sum_{m=1}^{M} \left[\log Y(m)\right] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right], \qquad (1)$$

In this case, n is the index of a cepstral coefficient.

C. Literature Review

[29] 2020. Using publicly accessible datasets, this research carried out a thorough review spanning both conventional and deep learning techniques in speech recognition. Even though the research was thorough, it lacked precise accuracy measurements, which makes it more difficult to evaluate the strategies' effectiveness directly.

[30] 2020. This research used machine learning and Internet of Things (IoT) approaches with a focus on vocal pathology monitoring systems. It provides a thorough analysis of the problems and applications related to these systems, but it doesn't say how accurate the models used are.

[31] 2022. Mel-Frequency Cepstral Coefficients (MFCC) and Multi-layer Perceptron models were used in this study, which yielded a 90.2% accuracy rate using Turkish voices. The study's capacity to analyze vocal signals is noteworthy, but its applicability is restricted to Turkish voices.
[32] 2021. This research used Convolutional Neural Networks (CNN) and MFCC with the Kaggle TensorFlow Speech Recognition Challenge dataset, and it achieved an accuracy of 88.21%. Although deep learning improves voice recognition, there is still room for improvement in terms of accuracy.
[33] 2020. In order to further our knowledge of deep learning applications in voice recognition, this research suggested a deep learning method for English speech recognition. It did not, however, address the suggested model's correctness.
[34] 2020. This study used Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), and Gaussian Bayes to classify respiratory illnesses with up to 98% accuracy. Although the great accuracy is excellent, huge feature sets would be needed for best results.
[35] 2022. This research used ResNet50 and Deep Neural Networks (DNN) to recognize gender using the Common Voice dataset, with a high accuracy of 98.57%. Its application is restricted to gender recognition, however.
[36] 2020. Using the TIMIT, RAVDESS, and BGC datasets, this work used a multi-layer architecture using K-Nearest Neighbours and Support Vector Machine to achieve gender detection accuracy of up to 96.8%. Like previous research, it is limited to gender identity.
[37] 2020. This study used a convolutional neural network to diagnose vocal pathology using the Saarbrucken voice database, with a 95.41% accuracy rate. While highlighting CNNs' efficacy, it also emphasizes the need for big datasets for training.
[38] 2020. This research achieved a 93.84% accuracy rate in Parkinson's disease diagnosis using Support Vector Machines. Although useful for diagnosing Parkinson's disease, its use is restricted to this particular condition.
[39] 2022. With a focus on speech recognition security, this study used convolutional neural networks to achieve accuracy as high as 96.87%. Although limited to security applications, the work is noteworthy for its trustworthy security analysis.
[40]2024. This extensive research investigated data augmentation methods for CNN-based voice categorization. Although it provides insightful information, accuracy measures are not specified.
[41] 2021. This research included a thorough overview of speech recognition technology; however, it lacked particular accuracy statistics for the techniques covered.
[42] 2021. In this study, deep learning neural networks and padding were used to classify Arabic speech. It does not state the accuracy attained, but it performs a good job of classifying Arabic alphabets.
[43] 2020. In order to illustrate how age and gender affect speech, this research used machine learning methods to analyze voice samples captured using cellphones. It does not, however, provide precise accuracy measures.

[44] 2021. This study looked at deep learning methods for speech recognition using a variety of sensors, including gyros, pictures, 3-axial magnetic sensors, EMG, EEG, EPG, EMA, PMA, and gyros. It organizes several technologies into taxonomy, however because of its wide scope, it lacks precise accuracy data.
[45] 2020. This research used signal processing techniques to conduct a survey on vocal disability identification approaches. Although thorough, it omits information about the precision of the techniques examined.
[46] 2022. This research achieved up to 77.49% accuracy in vocal pathology identification using deep neural networks and MFCC. Although useful for identifying disorders, accuracy varies according on the kind of illness.
[47] 2020. This study used discrete wavelet transform with MFCC, DWT, pitch, energy, and ZCR with an emphasis on emotion recognition. It stresses feature extraction approaches, however it doesn't give precise accuracy measurements.
[48] 2022. This thorough analysis used machine learning methods and recurrent neural networks (RNN) to study voice and vision systems. Although it offers a thorough summary, particular accuracy data is absent.
[49] 2020. With an emphasis on feature extraction, dimensionality reduction, and semantic comprehension, this research examined human-robot interaction systems. Despite being thorough, it lacks accuracy measurements.
[50] 2022. Using CNN for spoken language recognition with short recording intervals, this research achieved 100% accuracy in binary classification and 99.8% accuracy in multi-language classification. It is quite successful but only works with language recognition.
[51] 2020. This study discussed advancements in machine learning algorithms and flexible piezoelectric acoustic sensors for speech recognition. Though enlightening, it doesn't provide precise accuracy statistics.
[52] 2022. With a focus on Parkinson's disease, this research evaluated the effects of L-Dopa on voice characteristics and tracked the severity of the condition with good accuracy using SVM and machine learning techniques. It can only be used for Parkinson's disease.
[53]2020. By combining deep neural networks with MFCC and time-domain characteristics, this study enhanced sound detection. It doesn't state the level of precision attained.
[54] 2020. In order to identify emotions from speech, this research employed feature selection and Deep Convolutional Neural Networking (DCNN). Its results showed a range of accuracy across many datasets, including Emo-DB (95.10%), SAVEE (82.1%), IEMOCAP (83.8%), and RAVDESS (81.3%). Although it is quite precise, it can only identify certain emotions.
[55] 2020. This study investigated the detection of Alzheimer's illness using Logistic Regressions and spectrogram characteristics. It illustrates viability but leaves out accuracy measurements.
[56] 2020. This work employed SVM and Dynamic Time Warping to achieve 97% accuracy when using vocal commands to operate smart devices. It can only be evaluated using certain datasets.

[57] 2021. This review focused on CNN-based facial expression recognition (FER) applications. Although it offers a thorough analysis, it is devoid of precise accuracy data.
[58] 2020. Using machine learning techniques, this research was able to classify cardiac illnesses with 97.78% accuracy based just on heart sounds. Although very accurate, precise accuracy information for certain techniques is not given.

Table 2. Summary of Speech and Voice Recognition Studies

Reference	Year	Datasets Based	Model	Accuracy	Advantages	Limitations
[29]	2020	Publicly available datasets	Machine learning techniques	Not specified	This study conducted an elaborate survey covering both traditional and deep learning methods in voice recognition using publicly available datasets. Despite its comprehensive nature, the study did not provide specific accuracy metrics, which limits the direct assessment of the methods' performance.	Accuracy not provided
[30]	2020	Voice pathology surveillance systems	Internet of Things, Machine learning	Not specified	Focusing on voice pathology surveillance systems, this study integrated Internet of Things (IoT) and machine learning techniques. It offers an extensive review of the applications and challenges associated with these systems, although it does not specify the accuracy of the models employed.	Accuracy not provided
[31]	2022	Turkish Voices	Mel-Frequency Cepstral Coefficients, Multi-layer Perceptron	90.2%	This research used Turkish voices and applied Mel-Frequency Cepstral Coefficients (MFCC) and Multi-layer Perceptron models, achieving an accuracy of 90.2%. The study is notable for its success in processing voice signals but is limited to Turkish voices, restricting its generalizability.	Limited to Turkish Voices
[32]	2021	Kaggle TensorFlow Speech Recognition Challenge	Convolutional Neural Networks, MFCC	88.21%	Using the Kaggle TensorFlow Speech Recognition Challenge dataset, this study employed Convolutional Neural Networks (CNN) and MFCC, achieving an accuracy of 88.21%. While the use of deep learning enhanced speech recognition, the accuracy could still be improved.	Accuracy could be further improved
[33]	2020	English speech	Deep learning	Not specified	This research proposed a deep learning algorithm for English speech recognition, contributing to the understanding of deep learning applications in voice recognition. However, it did not specify the	Specific accuracy not provided

					accuracy of the proposed model.	
[34]	2020	Respiratory diseases	Support Vector Machines, k-Nearest Neighbor, Gaussian Bayes	Up to 98%	Targeting respiratory diseases, this research utilized Support Vector Machines (SVM), k-Nearest Neighbor (k-NN), and Gaussian Bayes, achieving up to 98% accuracy in disease classification. The high accuracy is impressive but may require large feature sets for optimal performance.	May require large feature sets
[35]	2022	Common Voice dataset	Deep Neural Networks, ResNet50	98.57%	Using the Common Voice dataset, this study applied Deep Neural Networks (DNN) and ResNet50, achieving a high accuracy of 98.57% in gender recognition. However, its scope is limited to gender recognition.	Limited to gender recognition
[36]	2020	TIMIT, RAVDESS, BGC datasets	Multi-layer architecture, K-Nearest Neighbors, Support Vector Machine	Up to 96.8%	This study employed a multi-layer architecture with K-Nearest Neighbors and Support Vector Machine on TIMIT, RAVDESS, and BGC datasets, achieving up to 96.8% accuracy in gender recognition. Similar to other studies, it is restricted to gender identification.	Limited to gender recognition
[37]	2020	Saarbrücken voice database	Convolutional Neural Network	95.41%	Using the Saarbrücken voice database, this research implemented a Convolutional Neural Network for voice pathology detection, achieving an accuracy of 95.41%. It highlights the effectiveness of CNNs but also notes the need for large datasets for training.	May require large datasets for training
[38]	2020	Not specified	Support Vector Machines	93.84%	This study used Support Vector Machines for the accurate diagnosis of Parkinson's disease, achieving an accuracy of 93.84%. While effective for Parkinson's diagnosis, its application is limited to this specific disease.	Limited to Parkinson's disease diagnosis
[39]	2022	Voice recognition security	Convolutional Neural Network	Up to 96.87%	Focusing on voice recognition security, this research utilized Convolutional Neural Networks, achieving up to 96.87% accuracy. The study is significant for its reliable security analysis but is confined to security applications.	Limited to voice recognition security
[40]	2024	Voice classification	CNN, Data augmentation	Not specified	This comprehensive study explored data augmentation techniques in voice classification using CNNs. While it offers valuable insights, it	Specific accuracy not provided

			techniques		does not specify accuracy metrics.	
[41]	2021	Voice recognition	Technology review	Not specified	This study provided a comprehensive review of voice recognition technologies, offering an extensive overview but lacking specific accuracy data for the methods discussed.	Specific accuracy not provided
[42]	2021	Arabic speech classification	Padding, Deep learning neural network	Not specified	Focusing on Arabic speech classification, this research employed padding and deep learning neural networks. It effectively classifies Arabic alphabets but does not specify the achieved accuracy.	Specific accuracy not provided
[43]	2020	Voice samples recorded through smartphones	Machine learning analysis	Not specified	This study analyzed voice samples recorded through smartphones, using machine learning techniques to demonstrate the effects of age and gender on voice. However, it does not provide specific accuracy metrics.	Specific accuracy not provided
[44]	2021	Various sensors (EMG, EEG, EPG, EMA, PMA, gyros, images, 3-axial magnetic sensors)	Deep learning techniques	Not specified	This review examined various sensors and deep learning techniques for voice recognition. It systematizes different technologies into a taxonomy but lacks specific accuracy data due to its broad focus.	Specific accuracy not provided. Focuses on a wide range of technologies, which may limit depth in individual areas.
[45]	2020	Voice disability detection	Signal processing techniques	Not specified	Conducting a survey on voice disability detection methods, this study used signal processing techniques. It is comprehensive but does not specify the accuracy of the methods reviewed.	Specific accuracy not provided
[46]	2022	Voice pathology detection	MFCC, Deep neural networks	Up to 77.49%	This study used MFCC and deep neural networks for voice pathology detection, achieving up to 77.49% accuracy. While effective for disorder detection, accuracy varies based on the type of disease.	Accuracy may vary based on the type of disease
[47]	2020	Common human emotions (e.g., anger, joy, sadness)	MFCC, DWT, pitch, energy, ZCR with discrete wavelet	Not specified	This study emphasized feature extraction algorithms for improved emotion recognition using a combination of MFCC, DWT, and other features. However, specific accuracy metrics were not provided.	Specific accuracy not provided

			transform			
[48]	2022	Speech and vision systems	Recurrent Neural Network (RNN), Machine Learning algorithms	Not specified	This comprehensive review focused on machine learning architectures and applications in speech and vision systems. While it provides a detailed overview, specific accuracy data for the models discussed is not provided.	Specific accuracy not provided
[49]	2020	Human-Robot Interaction systems	Feature extraction, Dimensionality reduction, Semantic understanding	Not specified	Investigating voice-based perception in Human-Robot Interaction systems, this study explored feature extraction, dimensionality reduction, and semantic understanding. However, it does not specify the accuracy of the methods discussed.	Specific accuracy not provided
[50]	2022	Spoken language identification	Convolutional Neural Network (CNN)	100% (binary classification), 99.8% (multi-language classification)	Focusing on language identification, this research achieved high accuracies of 100% for binary classification and 99.8% for multi-language classification using CNNs. However, its applicability is limited to language identification tasks.	Limited to language identification
[51]	2020	Speech processing	Flexible piezoelectric acoustic sensors, Machine learning algorithms	Not specified	Reviewing advancements in speech recognition, this study utilized advanced sensors and machine learning algorithms. However, it does not provide specific accuracy metrics for the methods discussed.	Specific accuracy not provided
[52]	2022	Parkinson's disease	Machine Learning algorithms, Support Vector Machine	High accuracy	Investigating Parkinson's disease, this research achieved high accuracy in tracking disease severity and evaluating the effect of L-Dopa on voice parameters. While significant for disease management, its scope is limited to Parkinson's disease.	Limited to Parkinson's disease
[53]	2020	Voice identification	Feature fusion, Deep Neural Network	Not specified	This study focused on voice identification, employing feature fusion and deep neural networks. It improved voice identification by integrating MFCC and time-domain features but did not provide specific accuracy metrics.	Specific accuracy not provided
[53]	2020	Emotion identification with speech	Selection of features, Deep Convolutional	Emo-DB: 95.10%, SAVEE: 82.1%, IEMOCAP: 83.8%,	Investigating emotion identification, this research achieved high accuracies using DCNN with feature selection on various datasets. However, its application is limited to emotion	Limited to emotion recognition

			Neural Networking (DCNN)	RAVDESS: 81.3%	recognition tasks.	
[55]	2020	Alzheimer's disease identification	Spectrogram features, Logistic Regression CV	Not specified	Assessing Alzheimer's disease identification, this study explored the feasibility of using speech data and machine learning. However, it does not provide specific accuracy metrics for the methods discussed.	Specific accuracy not provided
[56]	2020	Smart healthcare system	Support Vector Machine, Dynamic Time Warping	97%	Focusing on smart healthcare systems, this research effectively controlled devices through speech commands, achieving 97% accuracy. However, its evaluation was limited to specific datasets.	Limited evaluation on diverse datasets
[57]	2021	Facial expression recognition	Convolutional Neural Network	Not specified	Reviewing facial expression recognition, this study discussed applications and CNN algorithms but did not provide specific accuracy metrics.	Specific accuracy not provided
[58]	2020	Cardiac disease classification	Machine Learning algorithms	97.78%	Investigating cardiac disease classification, this study achieved high accuracy of 97.78% from heart sounds. However, it does not provide specific accuracy metrics for the methods discussed.	Specific accuracy not provided

D. Result Comparison And Discussion

The literature review table encapsulates a broad spectrum of studies exploring various facets of voice recognition technologies. Key findings highlight the efficacy of deep learning models, such as Convolutional Neural Networks (CNN) and Multi-layer Perceptron (MLP), in achieving high accuracy rates, particularly in specific tasks like gender recognition, voice pathology detection, and spoken language identification. The use of MFCCs as a feature extraction method emerges as a common trend across numerous studies, underscoring its importance in enhancing model performance. However, a notable limitation is the specificity of datasets and application domains, which often restricts the generalizability of the findings. While some studies report impressive accuracy metrics, such as 98.57% in gender recognition and 100% in binary language classification, these results are frequently limited to particular languages, diseases, or controlled environments. Additionally, the need for large datasets and computational resources for training sophisticated models like CNNs is a recurring challenge. Overall, the literature suggests a strong trend towards integrating machine learning and deep learning techniques in voice recognition, albeit with varying degrees of success and applicability across different use cases.

E. Conclusion

This review paper has synthesized the current state of voice recognition technologies, highlighting significant advancements and ongoing challenges within the field. The integration of machine learning and deep learning methods has markedly improved the accuracy and reliability of voice recognition systems, as evidenced by numerous studies achieving high performance metrics. The widespread adoption of MFCCs for feature extraction further demonstrates their critical role in processing speech signals effectively. Despite these advancements, the field still grapples with issues of dataset specificity and the need for substantial computational resources. Future research should focus on developing more generalized models that can perform well across diverse datasets and application domains. Additionally, efforts to optimize computational efficiency and reduce the dependency on large datasets will be crucial. Exploring innovative approaches, such as data augmentation techniques and hybrid models that combine traditional and deep learning methods, may also yield significant improvements. By addressing these challenges, the field of voice recognition can continue to advance, paving the way for more versatile and accessible speech processing applications.

F. References

[1] A. T. Ali, H. S. Abdullah, and M. N. Fadhil, "Voice recognition system using machine learning techniques," *Materials Today: Proceedings,* pp. 1-7, 2021.

[2] F. Cordella, A. Paffi, and A. Pallotti, "Classification-based screening of Parkinson's disease patients through voice signal," in *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2021, pp. 1-6.

[3] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella, and S. V. Alluri, "Real-time acoustic based depression detection using machine learning techniques," in *2020 International conference on emerging trends in information technology and engineering (ic-ETITE)*, 2020, pp. 1-6.

[4] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 32, pp. 1034-1047, 2021.

[5] A. J. Moshayedi, A. S. Roy, A. Kolahdooz, and Y. Shuxin, "Deep learning application pros and cons over algorithm deep learning application pros and cons over algorithm," *EAI Endorsed Transactions on AI and Robotics,* vol. 1, 2022.

[6] W. Wan, W. Sun, Q. Zeng, L. Pan, and J. Xu, "Progress in artificial intelligence applications based on the combination of self-driven sensors and deep learning," in *2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2024, pp. 279-284.

[7] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access,* vol. 8, pp. 221640-221653, 2020.

[8] L. Triyono, T. Yudantoro, S. Sukamto, and I. Hestinigsih, "VeRO: Smart home assistant for blind with voice recognition," in *IOP Conference Series: Materials Science and Engineering*, 2021, p. 012016.

[9] A. Baha'A, H. S. A. Arja, B. Y. Maayah, and M. M. Al-Taweel, "A dataset for voice-based human identity recognition," *Data in Brief,* vol. 42, p. 108070, 2022.

[10] H. A. Kholidy, A. Berrouachedi, E. Benkhelifa, and R. Jaziri, "Enhancing Security in 5G Networks: A Hybrid Machine Learning Approach for Attack Classification," in *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 2023, pp. 1-8.

[11] R. Amin, M. A. Al Ghamdi, S. H. Almotiri, and M. Alruily, "Healthcare techniques through deep learning: issues, challenges and opportunities," *IEEE Access,* vol. 9, pp. 98523-98541, 2021.

[12] C. Yu, M. Kang, Y. Chen, J. Wu, and X. Zhao, "Acoustic modeling based on deep learning for low-resource speech recognition: An overview," *IEEE Access,* vol. 8, pp. 163829-163843, 2020.

[13] J. Andreu-Perez, H. Pérez-Espinosa, E. Timonet, M. Kiani, M. I. Girón-Pérez, A. B. Benitez-Trinidad*, et al.*, "A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels," *IEEE Transactions on Services Computing,* vol. 15, pp. 1220-1232, 2021.

[14] M. Akay, Y. Du, C. L. Sershen, M. Wu, T. Y. Chen, S. Assassi*, et al.*, "Deep learning classification of systemic sclerosis skin using the MobileNetV2 model," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 2, pp. 104-110, 2021.

[15] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE access,* vol. 8, pp. 107562-107582, 2020.

[16] O. K. Toffa and M. Mignotte, "Environmental sound classification using local binary pattern and audio features collaboration," *IEEE Transactions on Multimedia,* vol. 23, pp. 3978-3985, 2020.

[17] O. Asmae, R. Abdelhadi, C. Bouchaib, S. Sara, and K. Tajeddine, "Parkinson's disease identification using KNN and ANN Algorithms based on Voice Disorder," in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, 2020, pp. 1-6.

[18] Y. Huang, J. Jing, and Z. Wang, "Fabric defect segmentation method based on deep learning," *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-15, 2021.

[19] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," *IEEE Journal of Biomedical and Health Informatics,* vol. 25, pp. 2162-2171, 2020.

[20] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE transactions on biomedical circuits and systems,* vol. 14, pp. 535-544, 2020.

[21] L. Lu, J. Mao, W. Wang, G. Ding, and Z. Zhang, "A study of personal recognition method based on EMG signal," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 14, pp. 681-691, 2020.

[22] U. Sehar, S. Kanwal, K. Dashtipur, U. Mir, U. Abbasi, and F. Khan, "Urdu sentiment analysis via multimodal data mining based on deep learning algorithms," *IEEE Access,* vol. 9, pp. 153072-153082, 2021.

[23] J. Liu, Z. Zhu, Y. Zhou, N. Wang, G. Dai, Q. Liu*, et al.*, "4.5 BioAIP: A reconfigurable biomedical AI processor with adaptive learning for versatile intelligent health monitoring," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, 2021, pp. 62-64.

[24] W. E. Villegas-Ch, J. García-Ortiz, and S. Sánchez-Viteri, "Identification of emotions from facial gestures in a teaching environment with the use of machine learning techniques," *IEEE Access,* vol. 11, pp. 38010-38022, 2023.

[25] T. Toivonen, I. Jormanainen, J. Kahila, M. Tedre, T. Valtonen, and H. Vartiainen, "Co-designing machine learning apps in K–12 with primary school children," in *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)*, 2020, pp. 308-310.

[26] S. Srivastav, K. Guleria, and S. Sharma, "Predictive Machine Learning Approaches for Cervical Cancer Detection: An Analytical Comparison," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2023, pp. 951-956.

[27] L. Guo, Z. Lu, and L. Yao, "Human-machine interaction sensing technology based on hand gesture recognition: A review," *IEEE Transactions on Human-Machine Systems,* vol. 51, pp. 300-309, 2021.

[28] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access,* vol. 9, pp. 79236-79263, 2021.

[29] N. H. Tandel, H. B. Prajapati, and V. K. Dabhi, "Voice recognition and voice comparison using machine learning techniques: A survey," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 459-465.

[30] F. T. Al-Dhief, N. M. a. A. Latiff, N. N. N. A. Malik, N. S. Salim, M. M. Baki, M. A. A. Albadr*, et al.*, "A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms," *IEEE Access,* vol. 8, pp. 64514-64533, 2020.

[31] U. Ayvaz, H. Gürüler, F. Khan, N. Ahmed, T. Whangbo, and A. A. Bobomirzaevich, "Automatic Speaker Recognition Using Mel-Frequency Cepstral Coefficients Through Machine Learning," *Computers, Materials & Continua,* vol. 71, 2022.

[32] A. Mahmood and U. Köse, "Speech recognition based on convolutional neural networks and MFCC algorithm," *Advances in Artificial Intelligence Research,* vol. 1, pp. 6-12, 2021.

[33] Z. Song, "English speech recognition based on deep learning with multiple features," *Computing,* vol. 102, pp. 663-682, 2020.

[34] M. Aykanat, Ö. Kılıç, B. Kurt, and S. B. Saryal, "Lung disease classification using machine learning algorithms," *International Journal of Applied Mathematics Electronics and Computers,* vol. 8, pp. 125-132, 2020.

[35] A. A. Alnuaim, M. Zakariah, C. Shashidhar, W. A. Hatamleh, H. Tarazi, P. K. Shukla*, et al.*, "Speaker gender recognition based on deep neural networks

and ResNet50," *Wireless Communications and Mobile Computing,* vol. 2022, p. 4444388, 2022.

[36] M. A. Uddin, M. S. Hossain, R. K. Pathan, and M. Biswas, "Gender recognition from human voice using multi-layer architecture," in *2020 International conference on innovations in intelligent systems and applications (INISTA)*, 2020, pp. 1-7.

[37] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. Khanapi Abd Ghani, M. S. Maashi, B. Garcia-Zapirain*, et al.*, "Voice pathology detection and classification using convolutional neural network model," *Applied Sciences,* vol. 10, p. 3723, 2020.

[38] Z. K. Senturk, "Early diagnosis of Parkinson's disease using machine learning algorithms," *Medical hypotheses,* vol. 138, p. 109603, 2020.

[39] W. Ibrahim, H. Candra, and H. Isyanto, "Voice recognition security reliability analysis using deep learning convolutional neural network algorithm," *Journal of Electrical Technology UMY,* vol. 6, pp. 1-11, 2022.

[40] H. Bakır, A. N. Çayır, and T. S. Navruz, "A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification," *Multimedia Tools and Applications,* vol. 83, pp. 17601-17628, 2024.

[41] R. M. Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Computers & Electrical Engineering,* vol. 90, p. 107005, 2021.

[42] A. Asroni, K. R. Ku-Mahamud, C. Damarjati, and H. B. Slamat, "Arabic speech classification method based on padding and deep learning neural network," *Baghdad Science Journal,* vol. 18, pp. 0925-0925, 2021.

[43] F. Asci, G. Costantini, P. Di Leo, A. Zampogna, G. Ruoppolo, A. Berardelli*, et al.*, "Machine-learning analysis of voice samples recorded through smartphones: the combined effect of ageing and gender," *Sensors,* vol. 20, p. 5022, 2020.

[44] W. Lee, J. J. Seong, B. Ozlu, B. S. Shim, A. Marakhimov, and S. Lee, "Biosignal sensors and deep learning-based speech recognition: A review," *Sensors,* vol. 21, p. 1399, 2021.

[45] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access,* vol. 8, pp. 66749-66776, 2020.

[46] M. Zakariah, Y. Ajmi Alotaibi, Y. Guo, K. Tran-Trung, and M. M. Elahi, "[Retracted] An Analytical Study of Speech Pathology Detection Based on MFCC and Deep Neural Networks," *Computational and Mathematical Methods in Medicine,* vol. 2022, p. 7814952, 2022.

[47] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology,* vol. 23, pp. 45-55, 2020.

[48] S. P. Yadav, S. Zaidi, A. Mishra, and V. Yadav, "Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN)," *Archives of Computational Methods in Engineering,* vol. 29, pp. 1753-1770, 2022.

[49] A. A. Badr and A. K. Abdul-Hassan, "A review on voice-based interface for human-robot interaction," *Iraqi Journal for Electrical and Electronic Engineering,* vol. 16, pp. 1-12, 2020.

[50] F. M. Rammo and M. N. Al-Hamdani, "Detecting the speaker language using CNN deep learning algorithm," *Iraqi Journal for Computer Science and Mathematics,* vol. 3, pp. 43-52, 2022.

[51] Y. H. Jung, S. K. Hong, H. S. Wang, J. H. Han, T. X. Pham, H. Park*, et al.*, "Flexible piezoelectric acoustic sensors and machine learning for speech processing," *Advanced Materials,* vol. 32, p. 1904020, 2020.

[52] A. Suppa, G. Costantini, F. Asci, P. Di Leo, M. S. Al-Wardat, G. Di Lazzaro*, et al.*, "Voice in Parkinson's disease: a machine learning study," *Frontiers in neurology,* vol. 13, p. 831428, 2022.

[53] R. Jahangir, Y. W. Teh, N. A. Memon, G. Mujtaba, M. Zareei, U. Ishtiaq, *et al.*, "Text-independent speaker identification through feature fusion and deep neural network," *IEEE Access,* vol. 8, pp. 32187-32202, 2020.

[54] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors,* vol. 20, p. 6008, 2020.

[55] L. Liu, S. Zhao, H. Chen, and A. Wang, "A new machine learning method for identifying Alzheimer's disease," *Simulation Modelling Practice and Theory,* vol. 99, p. 102023, 2020.

[56] A. Ismail, S. Abdlerazek, and I. M. El-Henawy, "Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping," *Sustainability,* vol. 12, p. 2403, 2020.

[57] S. M. S. Abdullah and A. M. Abdulazeez, "Facial expression recognition based on deep learning convolution neural network: A review," *Journal of Soft Computing and Data Mining,* vol. 2, pp. 53-65, 2021.

[58] A. Yadav, A. Singh, M. K. Dutta, and C. M. Travieso, "Machine learning-based classification of cardiac diseases from PCG recorded heart sounds," *Neural Computing and Applications,* vol. 32, pp. 17843-17856, 2020.