Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3488

Evaluating the Efficacy of Traditional Machine Learning Models in Speaker Recognition: A Comparative Study Using the LibriSpeech Dataset

Gregorius Airlangga[1*]
[1]Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
[1]gregorius.airlangga@atmajaya.ac.id

open access

***Gregorius Airlangga**

Article History:
Submitted: 23-01-2024
Accepted: 24-01-2024
Published: 31-01-2024

Keywords:
Speech Recognition, Machine Learning, Naïve Bayes, Logistic Regression, Gradient Boosting

ABSTRACT

The efficacy of machine learning models in speaker recognition tasks is critical for advancements in security systems, biometric authentication, and personalized user interfaces. This study provides a comparative analysis of three prominent machine learning models: Naive Bayes, Logistic Regression, and Gradient Boosting, using the LibriSpeech test-clean dataset—a corpus of read English speech from audiobooks designed for training and evaluating speech recognition systems. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted as features from the audio samples to represent the power spectrum of the speakers' voices. The models were evaluated based on precision, recall, F1-score, and accuracy to determine their performance in correctly identifying speakers. Results indicate that Logistic Regression outperformed the other models, achieving nearly perfect scores across all metrics, suggesting its superior capability for linear classification in high-dimensional spaces. Naive Bayes also demonstrated high efficiency and robustness, despite the inherent assumption of feature independence, while Gradient Boosting showed slightly lower performance, potentially due to model complexity and overfitting. The study underscores the potential of simpler machine learning models to achieve high accuracy in speaker recognition tasks, particularly where computational resources are limited. However, limitations such as the controlled nature of the dataset and the focus on a single feature type were noted, with recommendations for future research to include more diverse environmental conditions and feature sets.

INTRODUCTION

The application of machine learning in audio processing, particularly in the field of speaker identification, has become a prominent area of research in the rapidly advancing domain of artificial intelligence (Dhakal, Damacharla, Javaid, & Devabhaktuni, 2019; Jahangir et al., 2021; Kabir, Mridha, Shin, Jahan, & Ohi, 2021). This research article aims to explore and evaluate the performance of several machine learning models, namely Naive Bayes, Logistic Regression, and Gradient Boosting Classifiers. These models are applied in the context of speaker recognition, with a focus on utilizing Mel-Frequency Cepstral Coefficients (MFCCs) extracted from audio samples. The primary objective is to critically assess these models' effectiveness in accurately identifying speakers from the LibriSpeech test-clean dataset, a widely recognized benchmark in the field. Speaker recognition, a subset of audio signal processing, has been a subject of extensive research over several decades (Bai & Zhang, 2021; Hanifa, Isa, & Mohamad, 2021; Kabir et al., 2021). The evolution of this field can be traced back to the early works focusing on simple pattern recognition techniques. Pioneering studies in this domain employed basic linear models, which laid the groundwork for more complex approaches. The introduction of Gaussian Mixture Models (GMM) marked a significant advancement, as exemplified in the research by (Barai, Chakraborty, Das, Basu, & Nasipuri, 2022; Kamiński & Dobrowolski, 2022; Sisman, Yamagishi, King, & Li, 2020), which provided a robust method for modeling voice characteristics.

As the field progressed, the integration of machine learning algorithms became prevalent. The use of Support Vector Machines (SVM) demonstrated notable improvements in speaker verification tasks. Concurrently, the importance of effective feature extraction was introduced MFCCs (Hanifa et al., 2021; Jahangir et al., 2020, 2021). These coefficients have since become standard in audio processing tasks due to their ability to succinctly capture the timbral aspects of sound. The advent of deep learning has ushered in a new era for speaker recognition. The research from (Awad, Elkaffas, & Fakhr, 2023; J. Wang, Wang, Wang, & Zhang, 2023; Zhao, Han, Ouyang, & Burke, 2023) is a groundbreaking work on Deep Neural Networks (DNN) opened possibilities for handling the intricacies of human speech with unprecedented accuracy. Further studies have explored various architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), each contributing to the field's advancement (Galván & Mooney, 2021; Goel, Goel, & Kumar, 2023; N. Singh & Sabrol, 2021). In today's digital age, the relevance and urgency of enhancing speaker recognition technology cannot be overstated. With the proliferation of voice-activated devices, secure voice authentication systems, and the increasing reliance on remote communication, the demand for accurate and

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3488

efficient speaker recognition systems is at an all-time high. This technology is pivotal not only in user convenience but also in critical applications such as security and forensic analysis (Khan, Laghari, Awan, & Jumani, 2021; Musile et al., 2021; Sadaf et al., 2023). The surge in virtual assistants and smart home devices further underscores the need for robust speaker identification systems.

The current state of the art in speaker recognition is dominated by deep learning methods. These approaches have shown remarkable proficiency in capturing the complex patterns inherent in audio data, significantly outperforming traditional machine learning methods in many scenarios. However, they come with their own set of challenges, including the need for substantial computational resources and large datasets for training. This research seeks to explore the potential of alternative machine learning models that may offer a more computationally efficient yet effective solution for speaker recognition tasks. This study is driven by the goal of investigating and comparing the effectiveness of Naive Bayes, Logistic Regression, and Gradient Boosting Classifiers in speaker recognition. The research aims to provide a comprehensive analysis of these models based on precision, recall, and F1-score metrics. By doing so, it seeks to offer insights into the applicability and suitability of these models in real-world scenarios, especially where resource constraints are a significant consideration.

Despite the extensive body of research in speaker recognition, there is a notable gap in the comparative analysis of more traditional machine learning models against the backdrop of advanced deep learning techniques (Saleem, Potgieter, & Arif, 2021; Serradilla, Zugasti, Rodriguez, & Zurutuza, 2022; Sharma & Mehra, 2020). This research is poised to fill this gap by offering a detailed evaluation of these traditional models, particularly in contexts where computational resources are limited or where the complexity of deep learning models is not justifiable. This study contributes to the field by presenting an empirical evaluation of Naive Bayes, Logistic Regression, and Gradient Boosting Classifiers in the context of speaker recognition. This area, while crucial, has not been thoroughly explored in existing literature, particularly in comparison with more advanced deep learning techniques. The findings from this study could provide valuable insights for researchers and practitioners in the field, aiding in the selection of appropriate machine learning models for specific audio processing applications. The remainder of the article is organized as follows: Section 2 details the literature review to guide our reader about the existing research and our contribution. Section 3 details the methodology, encompassing the dataset description, the process of feature extraction, and the intricacies of the machine learning models utilized. Section 4 presents a thorough analysis of the experimental results. Section 5 delves into a discussion of these results, exploring their implications and relevance in the broader context of speaker recognition. Finally, Section 6 concludes the article, summarizing the key findings, acknowledging the limitations of the study, and suggesting potential directions for future research in this domain.

LITERATURE REVIEW

The inception of speaker recognition can be traced back to the 1960s and 1970s, a period marked by foundational research. Early systems were primarily analog and hinged on simple pattern recognition techniques. For example, the work of (Clarke & others, 2022; Ezzameli & Mahersia, 2023; Oviatt & Cohen, 2022) emphasized the potential of voice as a biometric marker but was limited by the rudimentary technology of the era, which struggled with varying speech patterns and low processing capabilities. The 1990s saw a significant advancement with the introduction of Gaussian Mixture Models (GMM) by researchers like (Flynn, Giannetti, & Van Dijk, 2023). GMMs effectively modeled voice features, but they often faltered in complex acoustic environments or when dealing with highly variable speech (Chignoli, 2022). Around the same time, Support Vector Machines (SVM) emerged, as explored by (Alimi, Ouahada, & Abu-Mahfouz, 2020), offering improved classification capabilities. However, SVMs were often computationally intensive and struggled with large-scale data, limiting their practicality in more extensive systems.

The adoption of Mel-Frequency Cepstral Coefficients (MFCCs), as proposed by (Abimbola, Kostrzewa, & Kasprowski, 2023; Srinivasa Murthy, Koolagudi, & Jeshventh Raja, 2021; Yong, 2022), marked a significant improvement in feature extraction. While MFCCs were effective in capturing the timbral aspects of speech, they were not without limitations. Their performance could degrade in noisy environments, and they were sometimes insufficient in capturing the nuances of different speaking styles and accents. Deep learning, especially Deep Neural Networks (DNN), introduced by (Akhtarshenas, Vahedifar, Ayoobi, Maham, & Alizadeh, 2023; Alhaizaey, 2023; Y. Wang et al., 2023), has significantly impacted speaker recognition. These methods excel in handling complex patterns but require extensive computational resources and vast datasets for optimal performance, making them less accessible for smaller-scale or resource-constrained projects. Subsequent exploration of architectures like CNNs and RNNs pushed the boundaries further but also highlighted issues such as overfitting and the need for fine-tuning to specific tasks (Akay, Karaboga, & Akay, 2022; Archana & Jeevaraj, 2024; Menghani, 2023).

Despite advancements, the field faces significant challenges. One major issue is the balance between model complexity and computational efficiency. Deep learning models, while powerful, are not always feasible in resource-limited settings (Cavalcanti, Eriksson, & Barbosa, 2021; Merzoug, Mostefaoui, Kechout, & Tamraoui, 2020). Another critical challenge is the robustness of these systems in diverse conditions. Factors like background noise, the speaker's emotional state, and varied recording qualities significantly impact performance (Gheewalla, McClelland, & Furnham,

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3488

2021; A. Singh, Kaur, Kukreja, Kadyan, & Kumar, 2022). Furthermore, the susceptibility of speaker recognition systems to spoofing attacks raises serious security concerns. Ethical considerations are also increasingly at the forefront. Issues of privacy, consent, and potential biases in algorithmic design are crucial, especially as these technologies become more integrated into everyday life. The reviewed literature indicates that while significant strides have been made in speaker recognition, there are notable gaps and limitations in current methodologies (Toussaint & Ding, 2021; Wassink, Gansen, & Bartholomew, 2022). Many of the advanced techniques, particularly those involving deep learning, are resource-intensive and may not be practical for all applications. Furthermore, there is a need for models that maintain high accuracy and robustness in varied and realistic environmental conditions.

This research aims to address these gaps by investigating the performance of more computationally efficient machine learning models like Naive Bayes, Logistic Regression, and Gradient Boosting Classifiers. These models, while perhaps less explored in recent literature, may offer a viable alternative to deep learning approaches, particularly in scenarios where resources are limited or where large datasets are not available. By focusing on these traditional models, this study seeks to contribute to the field by providing insights into their applicability and effectiveness in realistic speaker recognition scenarios, thereby offering potential solutions to some of the highlighted limitations in the current state of the art.

METHOD

The dataset used in this study was the LibriSpeech test-clean subset, a collection of read English speech sourced from audiobooks. It is widely recognized in the speaker recognition field due to its diverse array of speakers and high-quality audio recordings. The test-clean subset was specifically chosen to ensure a robust evaluation of speaker recognition models, as it encompasses a variety of accents, intonations, and speech patterns. After that, we conduct a feature extraction, the process of feature extraction is critical in audio processing and speaker recognition. In this study, Mel-Frequency Cepstral Coefficients (MFCCs) were used to extract audio features. MFCCs represent the short-term power spectrum of a sound, computed as follows equation (1).

$$MFCC(i) = \sum_{k=0}^{N-1} \log\big(S(k)\big) \cdot \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{N}\right] \text{ for } \quad i = 1, \ldots, M \tag{1}$$

Where S(k) is the power spectrum, N is the number of filters in the mel-filter bank, and M is the number of cepstral coefficients. Then, the librosa library in Python was utilized for this purpose, calculating 40 MFCCs for each audio sample. The meaning of these coefficients was computed to form a single feature vector for each audio file, thus capturing the essential characteristics of the speaker's voice. After that, the model training and testing are executed. Three machine learning models were selected for this study: Naive Bayes, Logistic Regression, and Gradient Boosting Classifier, each implemented using scikit-learn's corresponding classes. The choice of these models was driven by their distinct properties and suitability for speaker recognition tasks. The dataset was split into training and testing sets using scikit-learn's train_test_split function, allocating 80% of the data for training and 20% for testing. To ensure the generalizability of the models, feature normalization was performed using the StandardScaler from scikit-learn. This normalization is crucial to avoid disproportionate influence of certain features on the model and is defined as follows equation (2).

$$x_{\text{normalized}} = \frac{x - \mu}{\sigma} \tag{2}$$

Where x is the original feature vector, μ is the mean, and σ is the standard deviation of the training features. The models were evaluated based on precision, recall, and F1-score, which are standard metrics in machine learning classification tasks. These metrics are defined as presented in the equation (3) – (5). Firstly, Precision is the ratio of correctly predicted positive observations to the total predicted positives, calculated in the equation (3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

Where TP is the number of true positives and FP is the number of false positives. Secondly, recall (Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class, calculated in the equation (4).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3488

Where TP is the number of true positives and FN is the number of false negatives. Lastly, F1-Score is he weighted average of precision and recall, providing a balance between the two metrics, calculated in the equation (5).

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

These metrics collectively provide a comprehensive assessment of the model's performance, accounting for both accuracy and robustness in predictions.

RESULT

In this section, the researcher will explain the results of the research obtained. As presented in table 1, The table presents a comparative analysis of three machine learning models—Naive Bayes, Logistic Regression, and Gradient Boosting—using four performance metrics: Precision, Recall, F1-Score, and Accuracy. Each of these metrics provides insights into different aspects of the models' performance in the context of speaker recognition tasks. The Naive Bayes model shows a precision of 0.97, recall of 0.96, F1-Score of 0.96, and an accuracy of 0.96. These results suggest that the Naive Bayes model is highly effective in speaker recognition, with a very high rate of correctly identifying speakers (precision) and successfully retrieving most of the actual positive cases (recall). The balanced F1-Score implies that the model does not significantly favor precision over recall or vice versa. Furthermore, the high accuracy indicates that the model correctly identifies both speakers and non-speakers with high reliability.

The success of the Naive Bayes model could be attributed to its probabilistic foundation, which, despite the assumption of feature independence (often violated in real-world scenarios), seems to perform exceptionally well with the MFCC features used in this study. This suggests that for the dataset and features at hand, the distributional assumptions made by Naive Bayes are sufficiently met, or its misestimations do not drastically impact performance. The Logistic Regression model scored the highest across all metrics with a precision of 0.99, recall of 0.99, F1-Score of 0.99, and accuracy of 0.99. These impressive results underscore the model's exceptional ability to classify speakers accurately. The near-perfect scores suggest that Logistic Regression has effectively captured the underlying patterns and relationships in the feature set for this task.

Table 1. The Comparative Results of Machine Learning Models

Methods	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.97	0.96	0.96	0.96
Logistic Regression	0.99	0.99	0.99	0.99
Gradient Boosting	0.92	0.91	0.91	0.91

The advantage of Logistic Regression is its capability to provide a linear decision boundary that can be calibrated with great precision. Given the high dimensionality of the feature space from the MFCCs, Logistic Regression seems to have found a linear relationship that very accurately separates the different classes (speakers). The results could indicate that the decision boundary is well-defined and that the features are linearly separable to a high degree. Gradient Boosting displayed a precision of 0.92, recall of 0.91, F1-Score of 0.91, and accuracy of 0.91. While still high, these scores are the lowest among the three models. Gradient Boosting is an ensemble learning method known for its robustness and ability to model complex non-linear relationships. The lower scores in comparison to the other models suggest that this complexity may not be necessary or adequately captured for the given speaker recognition task. It is possible that the iterative refinement process of Gradient Boosting did not converge to a solution that is as effective as the simpler models or that the model overfitted to the training data, reducing its generalizability to the test data.

Comparing the three models, Logistic Regression stands out as the superior model for this dataset and feature set, followed closely by Naive Bayes, with Gradient Boosting slightly trailing behind. The results could indicate that the feature space, defined by the MFCCs, is more amenable to linear classification methods, as evidenced by the performance of Logistic Regression and Naive Bayes. It is also noteworthy that the high performance across all models suggests that the MFCC features are robust indicators for speaker recognition. This aligns with the literature, which often cites MFCCs as a reliable feature for speech and speaker recognition tasks. The high precision and recall values across all models suggest that false positives and false negatives are minimal, which is crucial in applications such as security systems where misidentifications can have serious consequences. The results of this study have significant implications for the design of speaker recognition systems, especially in scenarios where computational efficiency is valued alongside predictive performance.

The findings indicate that simpler models like Naive Bayes and Logistic Regression not only provide competitive performance but also offer advantages in terms of computational cost and interpretability over more complex models like Gradient Boosting. This is particularly relevant for real-time applications or devices with limited processing power. While the results are promising, it is important to consider the broader applicability of the findings. The study used a clean, high-quality dataset, and further research is needed to assess model performance in less

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3488

controlled environments. Additionally, evaluating the models' resilience to adversarial attacks and their performance with different demographic groups would be critical to ensure robustness and fairness.

DISCUSSION

This study's findings contribute to the body of knowledge on speaker recognition by providing a comparative analysis of three distinct machine learning models using the LibriSpeech test-clean dataset. The discussion of these findings is multi-faceted, considering the performance of the models, the implications for real-world applications, and the broader context of machine learning in speaker recognition. The results presented in Table 1 show that the Logistic Regression model outperforms the Naive Bayes and Gradient Boosting models across all metrics. Such high precision and recall values are indicative of the model's ability to correctly classify most of the positive cases without a significant number of false positives. These findings are consistent with previous studies that have highlighted the effectiveness of Logistic Regression in binary classification tasks, particularly in fields where the decision boundary can be distinctly defined through linear relationships (James et al., 2013).

In contrast, while the Gradient Boosting model performed commendably, it did not achieve the same level of accuracy as the other models. This could be due to the complexity of the model and the possibility of overfitting despite the general robustness that Gradient Boosting is known for (Natekin & Knoll, 2013). It raises questions about the necessity and efficiency of using more complex models over simpler ones when dealing with high-dimensional data where linear separability is present. Naive Bayes, although based on the strong assumption of feature independence, showed remarkably high performance, which aligns with the findings of Zhang et al. (2004), who noted that Naive Bayes could work well in practice even when the independence assumption is violated. The result is a testament to the model's utility as a baseline in speaker recognition tasks and its potential for applications where computational simplicity is required.

The practical implications of these results are significant, especially considering the need for efficient and accurate speaker recognition systems in various applications, from mobile device authentication to security systems. The high accuracy of Logistic Regression suggests that it can be an excellent choice for real-world applications where the balance between performance and computational efficiency is crucial. Additionally, the results have implications for the deployment of speaker recognition systems in environments with limited computational resources. In such scenarios, simpler models like Naive Bayes and Logistic Regression may provide a more feasible solution without a substantial trade-off in performance.

The performance of the machine learning models in this study also contributes to the ongoing discourse on the suitability of different machine learning approaches to speaker recognition. It challenges the notion that more complex, non-linear models are always necessary for high-dimensional data problems. Instead, it suggests that there might be cases, such as the one explored in this study, where the data's inherent structure allows for the successful application of simpler linear models. While the study provides valuable insights, it is not without limitations. The LibriSpeech test-clean dataset, while diverse and high-quality, represents a controlled environment that may not fully capture the variability present in real-world audio samples. Factors such as background noise, varying microphone quality, and different acoustic environments were not considered and could affect the generalizability of the results.

Additionally, the study focused on a single feature set (MFCCs), and it is possible that the inclusion of other features or a combination of various feature types could yield different results. Future studies could explore the integration of additional acoustic and linguistic features to potentially improve the robustness and accuracy of the models. Future research should aim to validate these findings across more diverse and challenging datasets, including those with lower-quality audio samples and greater background noise. Further research could also explore the use of ensemble methods that combine the predictions of multiple models to enhance accuracy and reliability. Moreover, investigating the models' performance across different languages and dialects would be valuable, considering the global application of speaker recognition technology. It would also be beneficial to assess the fairness and bias of these models, ensuring that they perform equitably across speakers of different genders, ethnicities, and ages.

CONCLUSION

This research has provided a comparative analysis of three machine learning models—Naive Bayes, Logistic Regression, and Gradient Boosting—applied to the task of speaker recognition using the LibriSpeech test-clean dataset. The Logistic Regression model demonstrated superior performance, achieving near-perfect precision, recall, F1-score, and accuracy, closely followed by Naive Bayes. Gradient Boosting, while still showing high performance, lagged slightly behind the other two models in all metrics. The findings of this study underscore the efficacy of Logistic Regression for high-dimensional classification tasks where the decision boundary between classes is well-defined. The strong performance of the Naive Bayes model also highlights its value, especially in scenarios where computational simplicity and efficiency are paramount. In contrast, the slightly lower performance of the Gradient Boosting model suggests that increased model complexity does not always equate to better performance for every task and can lead to challenges such as overfitting.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3488

These results have important implications for the design and implementation of speaker recognition systems, particularly in resource-constrained environments. They suggest that simpler, more interpretable models can provide robust performance and should not be overlooked in favor of more complex models without due consideration. However, this study is not without limitations. The controlled environment of the LibriSpeech test-clean dataset may not capture the full range of variability encountered in real-world audio samples. Future research should, therefore, focus on testing these models in more diverse and challenging acoustic environments and on incorporating a broader range of feature types to enhance model robustness and accuracy.

Moreover, the broader social implications of deploying machine learning models for speaker recognition, such as issues of privacy, consent, and potential biases, warrant further exploration. Ensuring the equitable performance of these systems across different demographic groups is an important aspect that future research should aim to address. In conclusion, while more complex machine learning models continue to be developed, this research highlights the ongoing relevance and potential of traditional models for speaker recognition tasks. Future advancements in the field should consider not only the accuracy but also the computational efficiency, interpretability, and fairness of machine learning models, tailoring the choice of model to the specific demands and constraints of the application context.

REFERENCES

Abimbola, J., Kostrzewa, D., & Kasprowski, P. (2023). Optimization of MFCCs for Time Signature Detection Using Genetic Algorithm. *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, 459–462.

Akay, B., Karaboga, D., & Akay, R. (2022). A comprehensive survey on optimizing deep learning models by metaheuristics. *Artificial Intelligence Review*, 1–66.

Akhtarshenas, A., Vahedifar, M. A., Ayoobi, N., Maham, B., & Alizadeh, T. (2023). Federated Learning: A Cutting-Edge Survey of the Latest Advancements and Applications. *ArXiv Preprint ArXiv:2310.05269*.

Alhaizaey, Y. (2023). *Optimizing task allocation for edge compute micro-clusters*. University of Glasgow.

Alimi, O. A., Ouahada, K., & Abu-Mahfouz, A. M. (2020). A review of machine learning approaches to power system security and stability. *IEEE Access*, *8*, 113512–113531.

Archana, R., & Jeevaraj, P. S. E. (2024). Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, *57*(1), 11.

Awad, A. L., Elkaffas, S. M., & Fakhr, M. W. (2023). Stock Market Prediction Using Deep Reinforcement Learning. *Applied System Innovation*, *6*(6), 106.

Bai, Z., & Zhang, X.-L. (2021). Speaker recognition based on deep learning: An overview. *Neural Networks*, *140*, 65–99.

Barai, B., Chakraborty, T., Das, N., Basu, S., & Nasipuri, M. (2022). Closed-set speaker identification using VQ and GMM based models. *International Journal of Speech Technology*, *25*(1), 173–196.

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2021). Multiparametric analysis of speaking fundamental frequency in genetically related speakers using different speech materials: Some forensic implications. *Journal of Voice*.

Chignoli, G. (2022). *Speech components in phonetic characterisation of speakers: a study on complementarity and redundancy of conveyed information*. Sorbonne Nouvelle.

Clarke, C., & others. (2022). *Reviver Voce: The Voice, Technology, and Death*. Falmouth University.

Dhakal, P., Damacharla, P., Javaid, A. Y., & Devabhaktuni, V. (2019). A near real-time automatic speaker recognition architecture for voice-based user interface. *Machine Learning and Knowledge Extraction*, *1*(1), 504–520.

Ezzameli, K., & Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 101847.

Flynn, J. S., Giannetti, C., & Van Dijk, H. (2023). Anomaly Detection of DC Nut Runner Processes in Engine Assembly. *AI*, *4*(1), 234–254.

Galván, E., & Mooney, P. (2021). Neuroevolution in deep neural networks: Current trends and future challenges. *IEEE Transactions on Artificial Intelligence*, *2*(6), 476–493.

Gheewalla, F., McClelland, A., & Furnham, A. (2021). Effects of background noise and extraversion on reading comprehension performance. *Ergonomics*, *64*(5), 593–599.

Goel, A., Goel, A. K., & Kumar, A. (2023). The role of artificial neural network and machine learning in utilizing spatial information. *Spatial Information Research*, *31*(3), 275–285.

Hanifa, R. M., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers \& Electrical Engineering*, *90*, 107005.

Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., … Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. *IEEE Access*, *8*, 32187–32202.

Jahangir, R., Teh, Y. W., Nweke, H. F., Mujtaba, G., Al-Garadi, M. A., & Ali, I. (2021). Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications*, *171*, 114591.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 3, **Number** 2, November 2023
https://doi.org/10.47709/brilliance.v3i2.3488

Kabir, M. M., Mridha, M. F., Shin, J., Jahan, I., & Ohi, A. Q. (2021). A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *IEEE Access*, *9*, 79236–79263.

Kamiński, K. A., & Dobrowolski, A. P. (2022). Automatic Speaker Recognition System Based on Gaussian Mixture Models, Cepstral Analysis, and Genetic Selection of Distinctive Features. *Sensors*, *22*(23), 9370.

Khan, A. A., Laghari, A. A., Awan, S., & Jumani, A. K. (2021). Fourth industrial revolution application: network forensics cloud security issues. *Security Issues and Privacy Concerns in Industry 4.0 Applications*, 15–33.

Menghani, G. (2023). Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, *55*(12), 1–37.

Merzoug, M. A., Mostefaoui, A., Kechout, M. H., & Tamraoui, S. (2020). Deep learning for resource-limited devices. *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, 81–87.

Musile, G., Agard, Y., Wang, L., De Palo, E. F., McCord, B., & Tagliaro, F. (2021). based microfluidic devices: On-site tools for crime scene investigation. *TrAC Trends in Analytical Chemistry*, *143*, 116406.

Oviatt, S., & Cohen, P. R. (2022). *The paradigm shift to multimodality in contemporary computer interfaces*. Springer Nature.

Sadaf, M., Iqbal, Z., Javed, A. R., Saba, I., Krichen, M., Majeed, S., & Raza, A. (2023). Connected and Automated Vehicles: Infrastructure, Applications, Security, Critical Challenges, and Future Aspects. *Technologies*, *11*(5), 117.

Saleem, M. H., Potgieter, J., & Arif, K. M. (2021). Automation in agriculture by machine and deep learning techniques: A review of recent developments. *Precision Agriculture*, *22*, 2053–2091.

Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. *Applied Intelligence*, *52*(10), 10934–10964.

Sharma, S., & Mehra, R. (2020). Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—a comparative insight. *Journal of Digital Imaging*, *33*, 632–654.

Singh, A., Kaur, N., Kukreja, V., Kadyan, V., & Kumar, M. (2022). Computational intelligence in processing of speech acoustics: a survey. *Complex \& Intelligent Systems*, *8*(3), 2623–2661.

Singh, N., & Sabrol, H. (2021). Convolutional neural networks-an extensive arena of deep learning. A comprehensive study. *Archives of Computational Methods in Engineering*, *28*(7), 4755–4780.

Sisman, B., Yamagishi, J., King, S., & Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 132–157.

Srinivasa Murthy, Y. V, Koolagudi, S. G., & Jeshventh Raja, T. K. (2021). Singer identification for Indian singers using convolutional neural networks. *International Journal of Speech Technology*, *24*, 781–796.

Toussaint, W., & Ding, A. Y. (2021). Sveva fair: A framework for evaluating fairness in speaker verification. *ArXiv Preprint ArXiv:2107.12049*.

Wang, J., Wang, J., Wang, S., & Zhang, Y. (2023). Deep learning in pediatric neuroimaging. *Displays*, *80*, 102583.

Wang, Y., Zhang, T., Zhao, L., Hu, L., Wang, Z., Niu, Z., … others. (2023). RingMo-lite: A Remote Sensing Multi-task Lightweight Network with CNN-Transformer Hybrid Framework. *ArXiv Preprint ArXiv:2309.09003*.

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, *140*, 50–70.

Yong, A.-P. C. (2022). *The Mel-frequency cepstrum coefficient for music emotion recognition in machine learning*. Macquarie University.

Zhao, J., Han, X., Ouyang, M., & Burke, A. F. (2023). Specialized deep neural networks for battery health prognostics: Opportunities and challenges. *Journal of Energy Chemistry*.