

Analysis Of Feature Importance And Signal To Noise Ratio Impact On Gradient Boost Model-Based Multilingual Speaker Recognition In Noisy Environment

Agbaji Adai Samuel¹

Department of Computer Engineering
University of Uyo,
Uyo, Akwa Ibom State, Nigeria
agajisam@yahoo.com

Silas Abraham Friday²

Department of Electrical/ Electronic Engineering
Adekeke University, Ede Osun State, Nigeria
silas.abraham@adelekeuniversity.edu.ng

Emmanuel Ubom³

Department of Electrical and Electronic Engineering,
Akwa Ibom State University, Ikot Akpaden,
Mkpat Enin, Akwa Ibom State.

¹emmanuelubom@aksu.edu.ng, ²ubongukommi@aksu.edu.ng

Abstract— Analysis of feature importance and signal to noise ratio (SNR) impact on Gradient Boost model-based multilingual speaker recognition in noisy environment is presented. Specifically, the study examines to what extent can signal to noise ratio and also the feature set used in the model training effect performance of Gradient Boost model in detecting speaker in noisy environment. Empirically collected speech of 15 persons from different languages and cultural backgrounds were used. White Gaussian noise was systematically introduced into the clean speech samples at predefined SNR levels, ranging from high (30 dB) to low (0 dB). The results showed that in terms of accuracy, the model performed better with the composite (speech signal with white noise) dataset having maximum accuracy of 92 % while the clean dataset (speech signal only, without noise) recorded accuracy of 86 %. In terms of precision, the model performed better with the composite dataset having maximum precision of 93 % while the clean dataset recorded precision of 85 %. In terms of F1 score, the model performed better with the composite dataset having maximum F1 score of

92 % while the clean dataset recorded F1 score of 83 %. In terms of recall, the model performed better with the composite dataset having maximum recall of 92 % while the clean dataset recorded recall of 86 %. In all, the model performance when trained with composite dataset is recommended as it gives a robust model that can perform well when trained with clean and composite dataset.

Keywords—*Feature Importance, Signal to Noise Ratio, Gradient Boost Model, White Gaussian Noise, Multilingual Speaker Recognition*

1. Introduction

The era of smart systems has prompted numerous research across the globe [1,2,3]. This is propelled by the advancements in sensors technologies, Artificial Intelligence (AI) and telecommunications technologies [4,5,6]. AI models are making it possible to detect, classify and manage things automatically without human intervention [7,8]. The AI-based approaches have been applied at industrial scales across the globe [9,10].

In this research, the focus is on the application of machine learning model to recognize speaker in a multilingual scenario [11,12]. Speaker identification is essential in security mechanisms, in customized solution

development and also in criminal investigation cases [13,14]. However, background noise can affect the performance of the model. As such, this work specifically focuses on evaluating the impact of the signal to noise ratio on the performance of the machine learning model used [15,16]. The work also examined the effect of feature set uses. In one hand, reducing the features will reduce the execution time but it also reduces the model accuracy. Therefore, the study seeks to ascertain the impact of feature set selection on the performance of the model used for the speaker recognition mechanism.

2. Methodology

The study examines the effect of signal to noise ratio and also the feature set used on the performance of Gradient Boost model used for speaker recognition in noisy environment. Empirically collected dataset was used.

2.1 Data Collection and Noise Augmentation

Speech of 15 persons from different languages and cultural backgrounds were recorded. The records lasted average of 120 seconds. The record were conducted in a controlled environment, first with minimal background noise, which is taken as the clean dataset without background noise. White Gaussian noise was systematically introduced into the clean speech samples at predefined SNR levels, ranging from high (30 dB) to low (0 dB). This process created a composite dataset comprising both clean and noisy speech samples, enabling the evaluation of the system's performance under varying noise conditions.

2.2 Feature Extraction

Feature extraction was performed using advanced signal processing techniques to derive meaningful representations. The following feature models were implemented:

i. Mel-Frequency Cepstral Coefficients (MFCCs)

These features model human auditory perception by mapping the frequency spectrum onto a mel scale and compressing the information using the Discrete Cosine Transform (DCT).

MFCCs were derived by performing the following steps:

1. Compute the magnitude spectrum of the short time fourier transform (STFT) as represented in Equation 3.3.

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N} \quad (1)$$

2. Apply the mel filter bank as represented in Equation 3.4 as represented in Equation 2.

$$M(f) = \log(\sum_{m=1}^M H_m |X(f)|^2) \quad (2)$$

3. Compute the Discrete Cosine Transform (DCT) of the log power spectrum to derive MFCCs as represented in Equation 3.

$$M(f) = \sum_{k=1}^K \log(M(f_k)) \cos\left(\frac{\pi n(k-0.5)}{K}\right) \quad (3)$$

ii. Gammatone Frequency Cepstral Coefficients (GFCCs)

GFCCs were derived using the gammatone filter bank, modelled to simulate the human auditory system. The output of each filter is processed to generate cepstral coefficients, emphasizing low-frequency components. The output of the m th gammatone filter is given by Equation 4.

$$y_m(t) = x(t) * g_m(t) \quad (4)$$

Where $g_m(t)$ the impulse response of the gammatone is filter and $*$ denotes convolution.

iii. Continuous Wavelet Transform (CWT)

3. CWT offers time-frequency localization for non-stationary signals. The continuous wavelet transforms for a signal $x(t)$ is given by 5.

$$CWT(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^*\left(\frac{t-b}{a}\right) dt \quad (5)$$

Where a and b are the scaling and shifting parameters, and ψ^* is the complex conjugate of the mother wavelet.

iv. Short-Time Fourier Transform (STFT)

STFT analyzes frequency components over time, crucial for detecting transient speech features using:

$$X(k, l) = \sum_{n=0}^{N-1} x(n)w(n-l)e^{-j2\pi kn/N} \quad (6)$$

2.3 Gradient Boosting (GB)

Gradient Boosting is an ensemble technique where weak learners (usually decision trees) are added sequentially, each correcting the residual errors of the previous one. At each step, the algorithm fits a new model to the residual errors of the previous model, updating the predictions incrementally.

The final prediction $F_M(x)$ after M boosting iterations is given in Equation 7 as:

$$F_M(x) = \sum_{m=1}^M \gamma_m h_m(x) \quad (7)$$

Where $h_m(x)$ is the m -th weak learner (usually a decision tree), γ_m is the weight associated with the learner, x is the input feature vector.

The Gradient Boost model was trained first using the clean dataset. The trained model was validated first with clean dataset and then with composite dataset of varying SNR. Afterwards, the Gradient Boost model was trained using the composite dataset. The trained model was validated first with clean dataset and then with composite dataset of varying SNR. Furthermore, the model was also trained using different number of features ranging from 5 to 48 features.

3. Results and discussion

4.1.4 Performance Evaluation of Gradient Boost Model

The performance evaluation of the Gradient Boost model under varying SNR conditions is presented in Figure 2, Figure 3, Figure 4 and Figure 5 while the performance evaluation of the Gradient Boost model under varying feature selection conditions is presented in Figure 6 and Figure 7. The results in Figure 2, Figure 3, Figure 4 and Figure 5 highlight the model's performance at specific Signal-to-Noise Ratio (SNR) levels, ranging from 0 dB to 30 dB at 5 dB intervals, and also at 70 that shows the upper value of the SNR used. The results in Figure 2, Figure 3, Figure 4 and Figure 5 also compare the models

performance when trained with cleaned data and composite data. In terms of accuracy, the model performed better with the composite dataset having maximum accuracy of 92 % while the clean dataset recorded accuracy of 86 %. In terms of precision, the model performed better with the composite dataset having maximum precision of 93 % while the clean dataset recorded precision of 85 %. In terms of F1 score, the model performed better with the composite dataset having maximum F1 score of 92 % while the clean dataset recorded F1 score of 83 %. In terms of recall, the model performed better with the composite dataset having maximum recall of 92 % while the clean dataset recorded recall of 86 %.

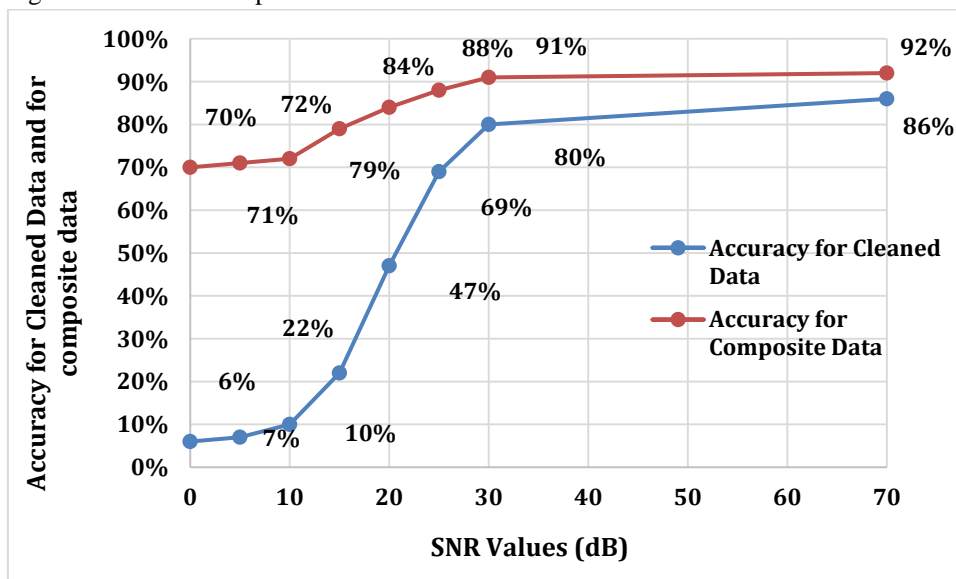


Figure 2 The accuracy for Cleaned Data and composite data

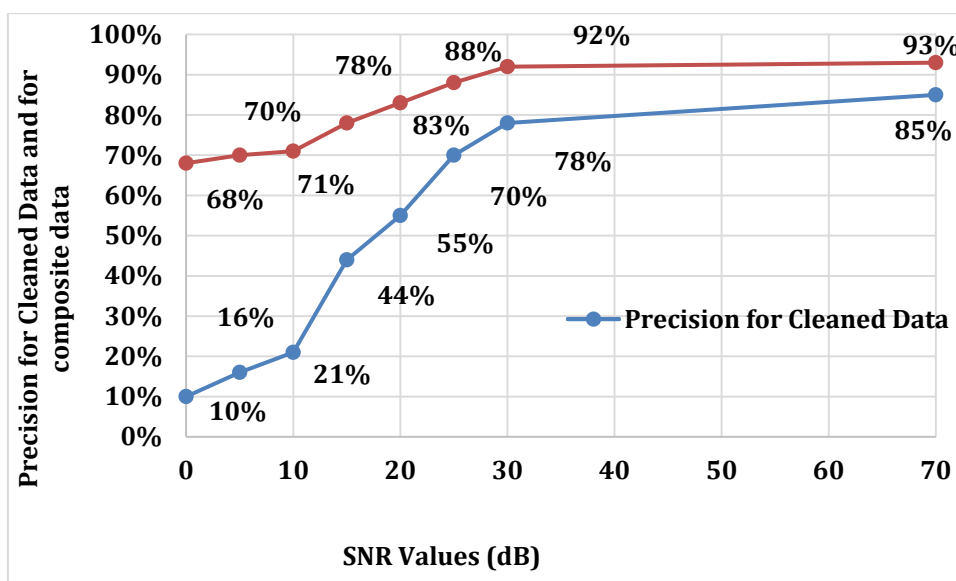


Figure 3 The Precision for Cleaned Data and composite data

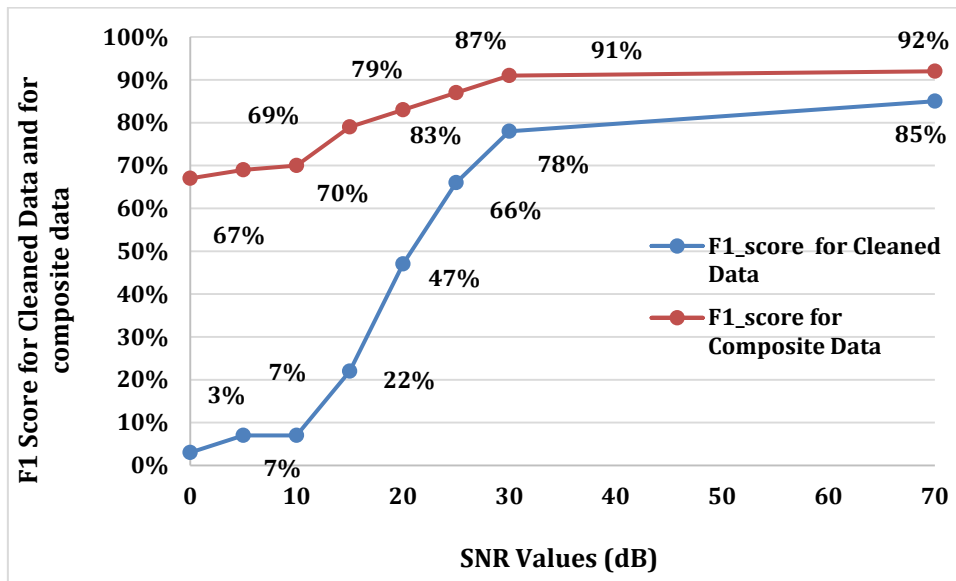


Figure 4 The F1 Score for Cleaned Data and composite data

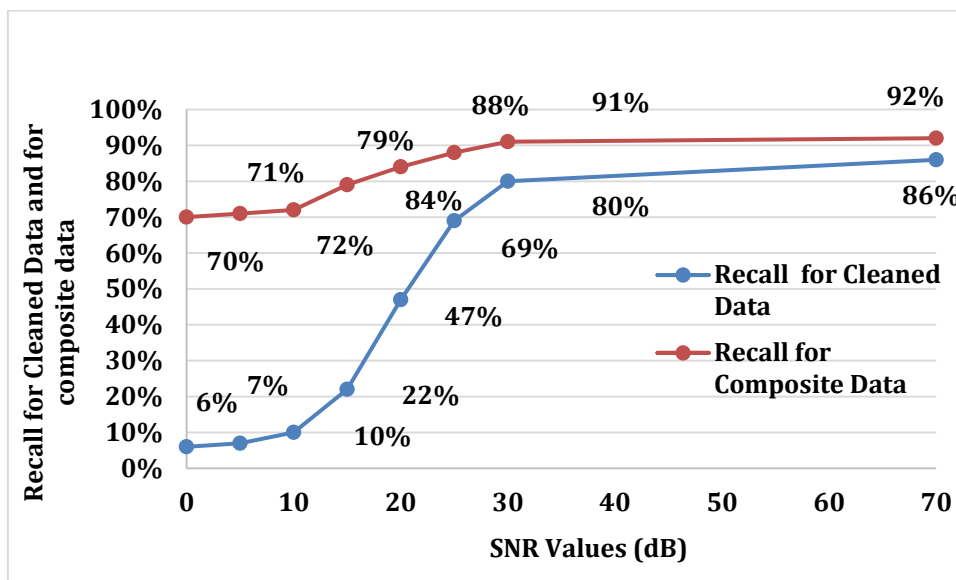


Figure 5 The Recall for Cleaned Data and composite data

Again the results in Figure 6 and Figure 7 show the model's performance based on the number of selected features, which range from 5 to 48 features at intervals of 5, for both clean and composite datasets.

The clean data trained model performed better when validated with clean dataset but performed poorly when validated with composite dataset, recording

maximum accuracy of 47 % with composite dataset and maximum accuracy of 86 % with clean dataset. On the other hand, the composite data trained model performed well when validated with both clean dataset and validated with composite dataset, recording maximum accuracy of 84 % with composite dataset and maximum accuracy of 94 % with clean dataset.

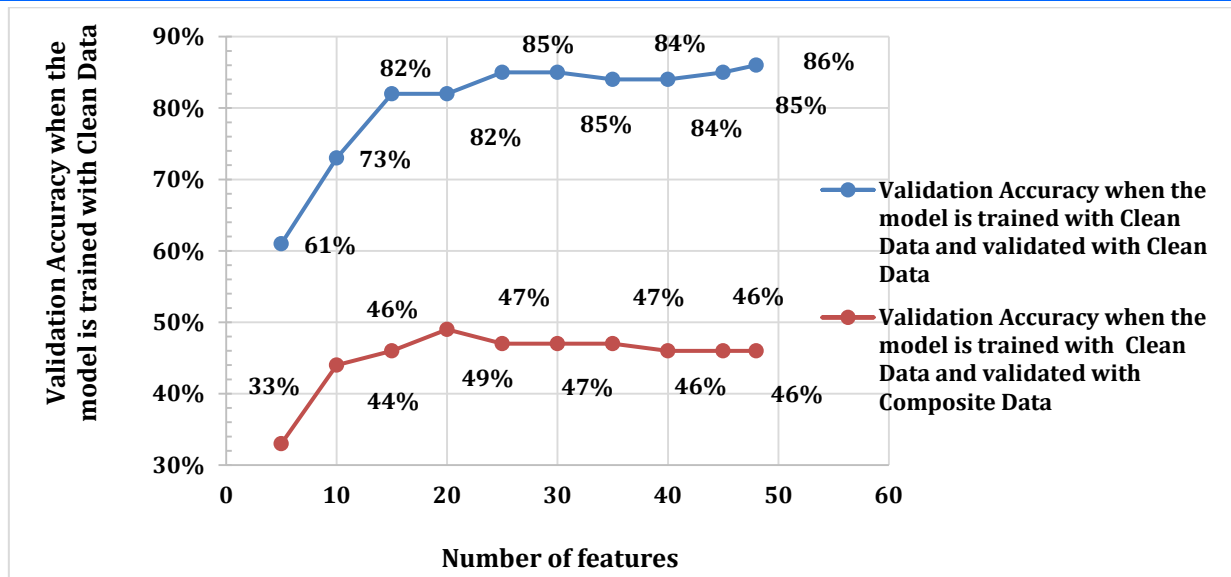


Figure 6 Validation Accuracy when the model is trained with Clean Data

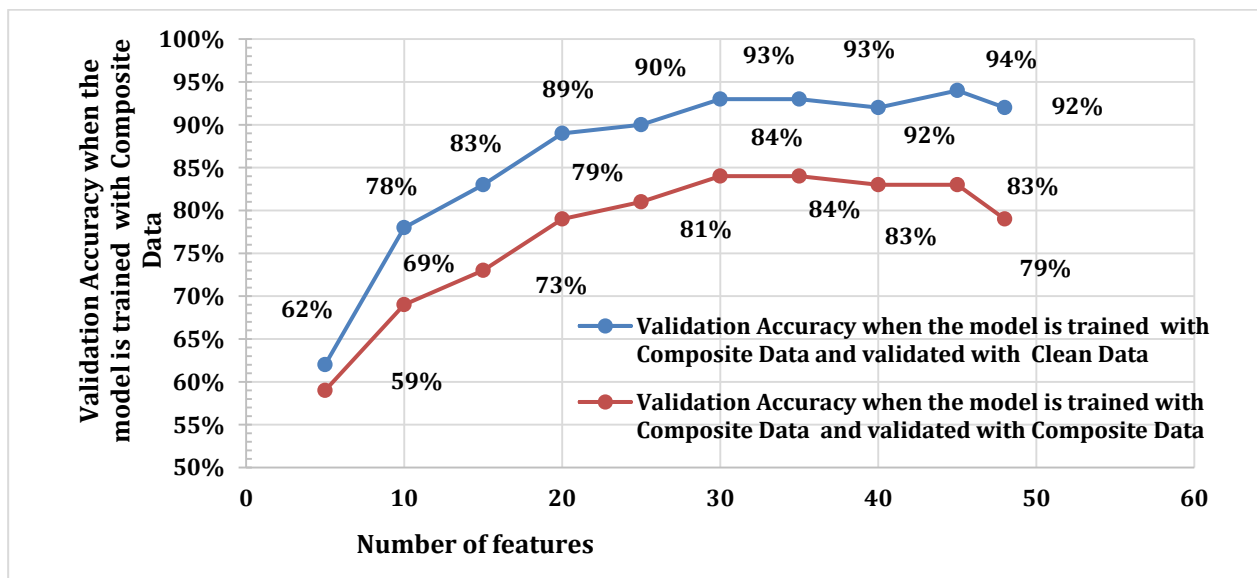
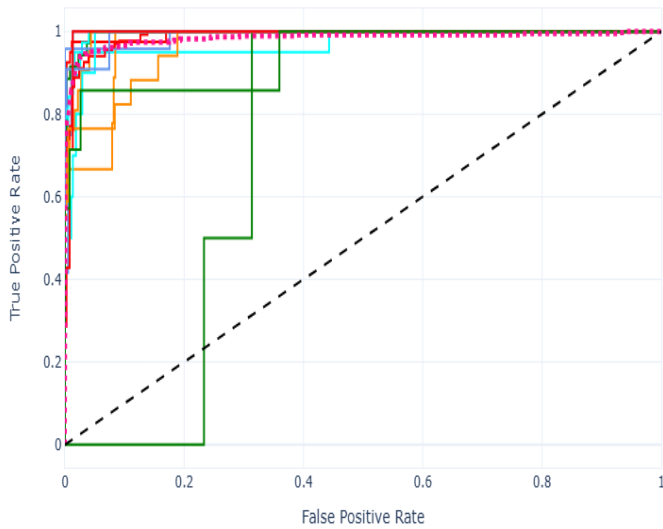


Figure 7 Validation Accuracy when the model is trained with Composite Data

The results in Figure 8 illustrate the Receiver Operating Characteristic (ROC) curve and confusion matrix for the Gradient Boost model developed using the clean training dataset, while Figure 9 shows the ROC curve and confusion matrix for the model trained with the composite dataset. Collectively, these results offer a comprehensive evaluation of the Gradient Boost model's effectiveness

across various SNR levels, feature sets, and training data types, demonstrating its robustness and adaptability under different experimental conditions.

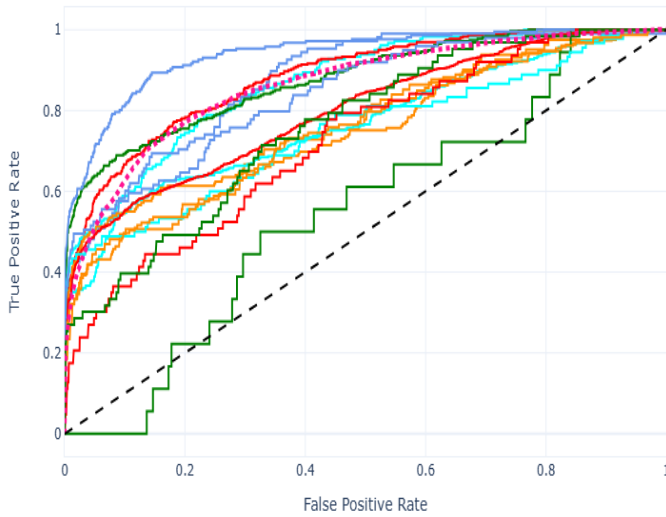
In all, the model performance when trained with composite dataset is recommended as it gives a robust model that can perform well when trained with clean and composite dataset.



(a) ROC with Clean Validation Dataset

Aaron	16	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0
Bawa	0	5	0	0	0	1	0	0	2	0	0	1	0	0	0	0
Becky	0	0	11	0	0	0	0	0	0	0	0	1	0	0	0	0
EmmaUbong	0	0	0	36	0	3	0	0	1	0	0	0	0	0	0	0
Etty	0	0	1	0	32	0	0	0	1	0	0	0	0	0	1	0
FRED	1	0	0	0	0	34	0	0	3	0	0	0	0	0	0	0
Izyee	0	1	0	0	0	0	16	0	3	0	0	0	0	0	0	1
JIME	0	0	1	0	0	1	0	21	1	0	0	0	0	0	0	0
JIMOH	0	0	1	0	0	0	1	0	132	0	1	0	0	0	0	0
Jonathan	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Mbre	0	0	0	0	0	0	0	0	4	0	5	0	0	1	0	0
Megee	0	0	1	1	0	0	1	0	3	0	0	11	0	0	0	0
MmaDickson	0	0	0	0	0	0	0	0	0	1	0	2	0	7	1	0
Nkanowo	0	0	0	0	0	0	2	0	2	0	0	0	0	0	3	0
Uduma	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	4

(b) Confusion Matrix with Clean Val Dataset

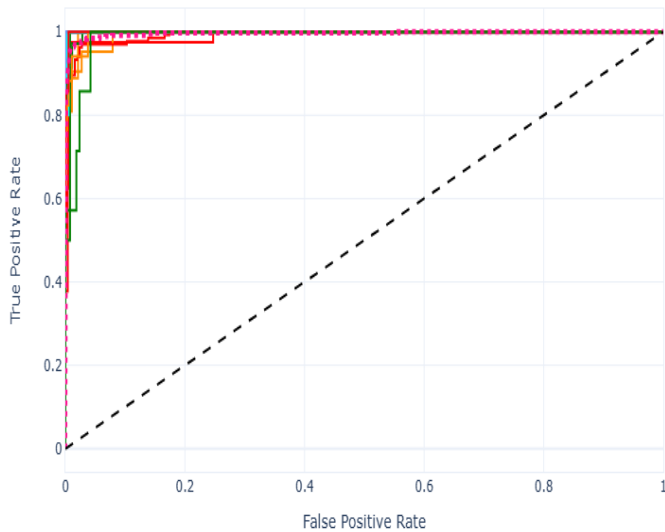


(c) ROC with Composite Validation Dataset

Aaron	51	0	51	50	0	4	1	9	12	0	0	1	1	0	0	0
Bawa	0	15	31	17	1	3	6	1	4	0	0	3	0	0	0	0
Becky	1	0	92	9	1	0	0	1	0	0	1	0	3	0	0	0
EmmaUbong	0	0	96	227	1	9	0	17	8	0	0	2	0	0	0	0
Etty	3	1	97	19	146	1	0	15	27	0	0	6	0	0	0	0
FRED	0	1	110	42	0	143	0	6	34	0	1	5	0	0	0	0
Izyee	0	2	61	28	0	0	70	1	19	0	0	8	0	0	0	0
JIME	2	0	48	18	0	1	0	140	7	0	0	0	0	0	0	0
JIMOH	0	0	481	23	1	5	1	58	627	0	0	19	0	0	0	0
Jonathan	0	0	8	0	0	1	0	0	9	0	0	0	0	0	0	0
Mbre	0	0	32	4	0	0	3	1	19	0	26	5	0	0	0	0
Megee	0	0	62	11	2	1	0	6	18	0	0	53	0	0	0	0
MmaDickson	0	0	35	13	3	0	4	4	8	0	2	1	29	0	0	0
Nkanowo	0	3	19	2	5	0	3	2	19	0	0	9	0	1	0	0
Uduma	0	4	16	11	0	0	13	2	1	0	0	11	1	0	0	4

(d) Confusion Matrix with Composite Val Dataset

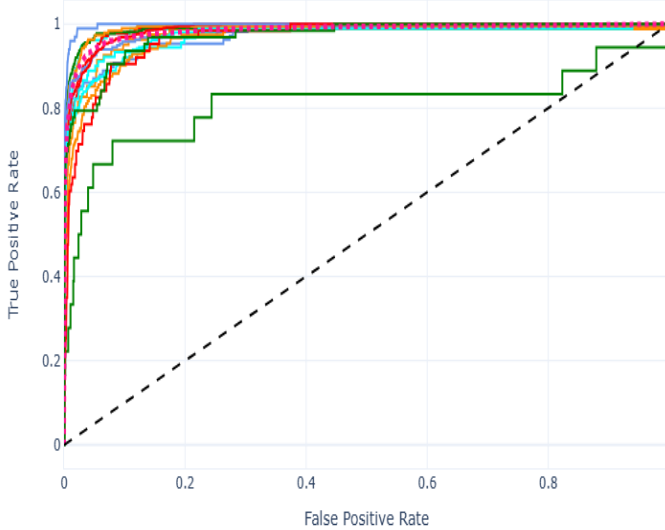
Figure 8: Receiver Operating Characteristic Curve and Confusion Matrix for Gradient Boost Model Developed with Clean Training Dataset.



(a) ROC with Clean Validation Dataset

Aaron	16	0	0	0	3	0	0	0	1	0	0	0	0	0	0	0
Bawa	0	8	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Becky	0	0	10	0	0	0	1	0	0	0	0	1	0	0	0	0
EmmaUbong	1	0	0	37	0	0	0	0	2	0	0	0	0	0	0	0
Etty	0	0	0	0	34	0	0	0	1	0	0	0	0	0	0	0
FRED	0	0	0	0	0	37	0	0	1	0	0	0	0	0	0	0
Izyee	0	0	0	0	0	0	18	0	0	0	0	0	0	0	1	2
JIME	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0
JIMOH	0	0	0	2	0	0	0	0	130	0	1	1	0	1	0	0
Jonathan	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Mbre	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0
Megee	0	0	0	0	0	0	0	0	3	0	0	14	0	0	0	0
MmaDickson	0	0	0	0	0	0	1	0	0	0	0	0	10	0	0	0
Nkanowo	0	0	0	0	0	0	0	0	1	0	0	0	0	0	6	0
Uduma	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	4

(b) Confusion Matrix with Clean Val Dataset



(c) ROC with Composite Validation Dataset

Aaron	135	0	0	20	2	7	1	0	13	0	0	2	0	0	0	0
Bawa	0	43	0	3	5	4	13	0	1	1	1	0	7	1	2	0
Becky	1	0	71	0	2	3	4	1	17	0	0	7	2	0	0	0
EmmaUbong	17	0	0	304	6	9	0	2	19	0	0	1	0	1	1	0
Etty	4	1	2	3	280	2	0	1	15	1	0	5	0	0	1	0
FRED	4	1	0	18	8	263	1	5	39	0	0	3	0	0	0	0
Izyee	0	3	1	0	2	2	158	0	10	1	0	3	4	2	3	0
JIME	0	0	2	6	0	10	0	168	29	0	0	0	0	1	0	0
JIMOH	1	0	1	2	4	13	1	11	1177	0	1	3	0	1	0	0
Jonathan	0	0	0	0	0	1	1	0	8	1	6	0	1	0	0	0
Mbre	0	0	0	0	0	4	1	0	22	1	56	3	1	2	0	0
Megee	1	1	0	3	7	4	2	1	29	0	0	100	2	2	1	0
MmaDickson	0	0	0	0	1	0	8	0	2	0	1	1	84	0	2	0
Nkanowo	0	2	1	0	5	2	1	0	23	0	2	8	0	19	0	0
Uduma	0	4	0	0	4	4	5	1	0	0	0	3	1	1	40	0

(d) Confusion Matrix with Composite Val Dataset

Figure 9: Receiver Operating Characteristic Curve and Confusion Matrix for Gradient Boost Model Developed with Composite Training Dataset.

4. Conclusion

Gradient Boost model is presented for speaker recognition. The model was applied in different scenarios; one the model was trained with clean dataset without significant noise and then validated with clean dataset and also with dataset having noise components with varying signal to noise ratio (SNR) values. Also, the model was trained with composite dataset and then validated with clean dataset and with dataset having noise of varying SNR. Furthermore, the model was trained with varying number of features which have been sorted according to feature importance ranking. The results show that the model performed best in all cases whet it was trained with composite dataset. Also, the more the features selected, the better the model accuracy. Although, the training time was

not reported but the impact of more features is that the model execution time increases with the number of features adopted. In all, the study effectively confirmed that training the model with composite dataset makes it more robust in identifying the speaker even in noisy environment.

References

1. Saleem, Y., Crespi, N., Rehmani, M. H., & Copeland, R. (2019). Internet of things-aided smart grid: technologies, architectures, applications, prototypes, and future research directions. *Ieee Access*, 7, 62962-63003.
2. Shafique, K., Khawaja, B. A., Sabir, F., Qazi, S., & Mustaqim, M. (2020). Internet of things (IoT) for next-generation smart systems: A

- review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. *IEEE access*, 8, 23022-23040.
3. Sarker, I. H. (2022). Smart City Data Science: Towards data-driven smart cities with open research issues. *Internet of Things*, 19, 100528.
 4. Alahi, M. E. E., Sukkuea, A., Tina, F. W., Nag, A., Kurdthongmee, W., Suwannarat, K., & Mukhopadhyay, S. C. (2023). Integration of IoT-enabled technologies and artificial intelligence (AI) for smart city scenario: recent advancements and future trends. *Sensors*, 23(11), 5206.
 5. Monteiro, A. C. B., França, R. P., Arthur, R., & Iano, Y. (2021). An overview of artificial intelligence technology directed at smart sensors and devices from a modern perspective. *Smart Sensor Networks: Analytics, Sharing and Control*, 3-26.
 6. Chander, B., Pal, S., De, D., & Buyya, R. (2022). Artificial intelligence-based internet of things for industry 5.0. In *Artificial intelligence-based internet of things systems* (pp. 3-45). Cham: Springer International Publishing.
 7. Shu, X., Yao, D., & Bertino, E. (2015). Privacy-preserving detection of sensitive data exposure. *IEEE transactions on information forensics and security*, 10(5), 1092-1103.
 8. Nozari, H., & Sadeghi, M. E. (2021). Artificial intelligence and Machine Learning for Real-world problems (A survey). *International Journal of Innovation in Engineering*, 1(3), 38-47.
 9. Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Artificial intelligence applications for industry 4.0: A literature-based study. *Journal of Industrial Integration and Management*, 7(01), 83-111.
 10. Khang, A., Shah, V., & Rani, S. (Eds.). (2023). *Handbook of Research on AI-Based Technologies and Applications in the Era of the Metaverse*. IGI Global.
 11. Jani, M. M., Panchal, S. R., Patel, H. H., & Raiyani, A. (2023, December). Multilingual speech recognition: An in-depth review of applications, challenges, and future directions. In *International Conference on Communication and Intelligent Systems* (pp. 1-13). Singapore: Springer Nature Singapore.
 12. Fan, P., Guo, D., Zhang, J., Yang, B., & Lin, Y. (2024). Enhancing multilingual speech recognition in air traffic control by sentence-level language identification. *Applied Acoustics*, 224, 110123.
 13. Saxena, N., & Varshney, D. (2021). Smart home security solutions using facial authentication and speaker recognition through artificial neural networks. *International Journal of Cognitive Computing in Engineering*, 2, 154-164.
 14. Sudharsan, B., Corcoran, P., & Ali, M. I. (2022). Smart speaker design and implementation with biometric authentication and advanced voice interaction capability. *arXiv preprint arXiv:2207.10811*.
 15. Lacy, F., Ruiz-Reyes, A., & Brescia, A. (2024). Machine learning for low signal-to-noise ratio detection. *Pattern Recognition Letters*, 179, 115-122.
 16. Yan, T., Wang, D., Kong, J. Z., Xia, T., Peng, Z., & Xi, L. (2021). Definition of signal-to-noise ratio of health indicators and its analytic optimization for machine performance degradation assessment. *IEEE Transactions on Instrumentation and Measurement*, 70, 1-16.