ISSN: 2581-1274

Voice Recognition Performance Analysis on Smart Speakers Using the Random Forest Method

Muhammad Yani [a,1,*], Muhammad Fikry [a,2], Arnawan Hasibuan [a,3], Nudin [a,4] , Munirul Ula [a,5] , Husaini [a,6]

[a]*Teknologi Informasi, Universitas Malikussaleh, Lhokseumawe, 24355, Indonesia*
muhammad.237110201009@mhs.unimal.ac.id[1,]; muh.fikry@unimal.ac.id[2]; arnawan@unimal.ac.id[3] ; nurdin@unimal.ac.id[4] ; munirulula@unimal.ac.id[5]; husainizulkifli@uinsuna.ac.id[6]*
** corresponding author*

ARTICLE INFO

Article history:
Published
May 7, 2026

Keywords:
Voice Rrecognition
Smart Speaker
Far Field Speech Recognition
Random Forest
Amazon Alexa

ABSTRACT

The development of Internet of Things (IoT) and artificial intelligence technology has driven the increasing use of voice user interfaces (VUI) as a more natural form of human-computer interaction. One widely used VUI implementation is voice recognition-based smart speakers. Despite its widespread adoption, voice recognition performance on smart speakers is not necessarily optimal when used in real-world conditions, particularly in far-field scenarios that are influenced by user distance, environmental noise, and system response time. This study aims to analyse and compare the voice recognition performance of Amazon Alexa smart speakers and the Interactive Speaker System as a non-vendor comparison system. Testing was conducted at varying user distances in a non-soundproof room to represent real-world operational conditions.The obtained performance data was analyzed using the Random Forest method as a classification tool due to its ability to handle multivariate data and nonlinear relationships between variables. The results showed that variations in user distance significantly affected the voice recognition performance of both systems, with a tendency for performance to decrease as distance increased. In addition, differences in system architecture characteristics also influenced the level of resilience to environmental conditions. The application of the Random Forest method also enabled the identification of dominant factors that influence the success of voice recognition. This research is expected to provide theoretical contributions in the study of voice recognition performance in far-field scenarios, as well as practical contributions as a basis for consideration in the selection and development of more reliable voice-based interaction systems in real environments.

I. Introduction

The development of digital technology, the Internet of Things or IoT, and artificial intelligence has changed the pattern of human interaction with computing systems [1].While previously interaction was dominated by graphical interfaces through touch and text, there has now been a shift toward Voice User Interfaces, which allow users to give commands or obtain information through voice conversations.Voice interaction is considered more natural because it mimics human communication and supports both hands-free and eyes-free scenarios, such as cooking, driving, or performing household tasks. In modern voice assistants, increased language processing capabilities and digital service integration are also driving the increasing use of VUIs in everyday life [2].

In the home environment, the adoption of smart speakers has shown that voice interactions do not always take place under ideal conditions such as laboratory testing [3]. Interactions often occur spontaneously and are influenced by social context, such as other conversations in the room, noise from electronic devices, and changes in speaking strategy when the system fails to understand a command. Studies of home voice assistant device usage confirm that user practices are dynamic and

situational, including attempts to repeat commands, modify phrases, or change intonation to get the device to respond [4]. The term voice recognition refers to a system's ability to perform Automatic Speech Recognition (ASR), which is the process of converting speech signals into text representations or labels that can be processed as commands. Modern ASR has evolved rapidly with advances in machine learning and deep learning, but its performance remains influenced by the quality of the signal received by the microphone and the processing conditions [5]. Voice recognition technology enables computer systems to automatically recognize, process, and understand voice commands through speech signal modeling and machine learning approaches [2].

This capability makes human-computer interaction more efficient, especially on devices designed for hands-free and context-based use. Smart speakers are a key application of voice recognition technology, integrating microphones, speakers, network connectivity, and cloud-based AI, and have become an important part of the IoT ecosystem. However, in real-world use, their speech recognition performance often degrades due to user distance, background noise, and room acoustics, particularly in far-field conditions where increased distance and reverberation reduce signal quality and require more robust recognition systems [6].

Amazon Alexa is one of the most widely used and researched commercial smart speaker platforms. Alexa is designed to support far-field speech recognition through the use of microphone array technology and cloud-based processing. Despite optimizations, Alexa's performance is still affected by environmental conditions and user usage patterns [7]. In addition to commercial smart speakers, a conversational artificial intelligence-based voice interaction system designed as a communication interface between the user and the system is also being developed. In this research, this system is referred to as the Interactive Speaker System. This system allows for more flexible and controlled voice recognition performance testing because it is not tied to a specific vendor ecosystem [8]. A performance comparison between a commercial smart speaker and an Interactive Speaker System is essential for gaining a more comprehensive understanding of the performance characteristics of voice recognition on different system architectures. This comparative analysis allows for the identification of the advantages and limitations of each system under comparable test conditions [9].

The Interactive Speaker System was used as a non-vendor system for controlled comparative testing, so that the effects of distance, acoustic conditions, and system response could be analyzed more rigorously and in line with conversational agent evaluation studies [10]. Speech Recognition (ASR) is a core technology that converts speech into text or commands, with modern developments based on deep learning that improve the ability to understand complex variations in voice signals [11]. A systematic review of ASR shows a research trend towards more robust techniques across domain and environment differences, especially for applications in consumer products such as smart speakers and voice assistants [12].

Far-field speech recognition is speech recognition from a distance that causes the signal to weaken and is susceptible to interference, and is the main operational condition of smart speakers because the interaction is hands-free [13]. Microphone arrays enable multi-channel far-field ASR with integrated neural beamforming and end-to-end models that improve robustness to variations in sound source position and room conditions, so performance evaluations must consider user distance, room acoustics, and environmental interference in an integrated manner [14].

Environmental noise degrades ASR performance in far-field smart speakers by damaging the spectral quality of speech, decreasing the SNR, and causing acoustic mismatch, thereby increasing the speech recognition error rate [15]. Increasing distance degrades signal quality so that acoustic features such as MFCC become unstable and distorted, which ultimately reduces the model's ability to recognize speech consistently [16]. Increasing user distance limits the effective range of speech recognition, with performance degrading beyond the optimal distance due to limitations in microphone sensitivity and signal attenuation that cannot be fully compensated for by software [17]. Latency in speech recognition systems is the time delay between a user command and the system's

response, spanning the acquisition to recognition process, and is a key factor in the quality of interaction and real-time response on smart speakers [18].

In real-time ASR, VAD and streaming architectures such as incremental decoding are used to reduce latency by allowing the system to respond immediately without waiting for the entire utterance to complete [19]. Random Forest is effectively used for classification in audio and speech domains, including far-field and noisy conditions, so it is appropriately chosen as an analysis tool for evaluating voice recognition performance without focusing on the development of the ASR algorithm [20].

II. Method

A. Research Flow

The research process (figure 2) began with identifying voice recognition performance issues on smart speakers in real-world use, particularly the influence of user distance. Based on these initial findings, the research focused on analyzing and comparing voice recognition performance between Amazon Alexa and the Interactive Speaker System. This was followed by a literature review, experimental design, and preparation of test devices and scenarios. The testing phase involved participants, followed by data analysis and system performance evaluation as the basis for drawing conclusions.

B. Research Stages

This research followed a systematic process beginning with problem identification and objective formulation, supported by a literature review. An experimental design with user distance scenarios was implemented, followed by data collection, labeling, feature extraction, and preprocessing. The data were then analyzed and classified using Random Forest, and the results were evaluated to address the research objectives and draw conclusions.

C. Research Variables

This study defines variables to maintain the focus of the analysis, with the main independent variable being the user's distance from the device (1, 2, and 3 meters) and the dependent variables including the success of voice recognition and the system's response time. Recognition success is determined by the appropriateness of the meaning of the voice command, while response time assesses the responsiveness of the interaction. The environmental noise level is also measured as a control variable to ensure the testing is conducted under documented and consistent conditions, thus focusing the analysis on the effect of distance on voice recognition performance.

D. Research Instruments and Tools

The research instruments and devices were selected to support systematic testing and reliable data. The primary devices were the Amazon Alexa smart speaker, representing a commercial cloud-based system, and the Interactive Speaker System, a flexible and controlled benchmark system, allowing testing to reflect real-world conditions.

Fig 2. Amazon Alexa

Fig 2. Research Flow

Amazon Alexa technical specifications are presented to provide an overview of the device's ability to capture and process voice commands at various user distances tested in the study.

Table 1. Amazon Alexa Specifications

Parameter	**Specifications**
Device name	Amazon Echo Dot (5th Generation)
Device type	Smart speaker
Virtual assistant	Amazon Alexa
Microphone technology	Far-field microphone array
Number of microphones	4 microphones
Connectivity	Wi-Fi
Speech processing	Cloud-based
Speaker	In-ear speaker
Additional features	Temperature sensor and accelerometer
Main functions	Voice recognition and virtual assistant

The comparison device specifications are the Interactive Speaker System Interface

Table 2. Interactive Speaker System Specifications

Parameter	Specifications
System Type	Vendor-free voice interaction system
Processing Device	Windows laptop operating system
Microphone	Laptop internals
Interface Platform	Web (Flask REST API)
Input Method	Voice → STT → query text
Speech Processing	Local Speech-to-Text
Model/AI Engine	OpenAI (ChatCompletion, GPT-4)
Data Sent to AI	Text (system prompt + user query)
Network Connection	Internet (OpenAI API access)
System Output	Response text
Logging/Database	MySQL (conversation history, JSON)
Performance Parameters	Response time (seconds)
Main Functions	AI-based voice recognition and responses

Voice Input and Response Log is a web interface on an interactive speaker system that allows users to issue voice commands and view a history of the system's responses. This interface displays a list of questions and answers, along with response times and durations, and also serves to monitor and evaluate system performance, including response consistency, processing speed, and interaction patterns.

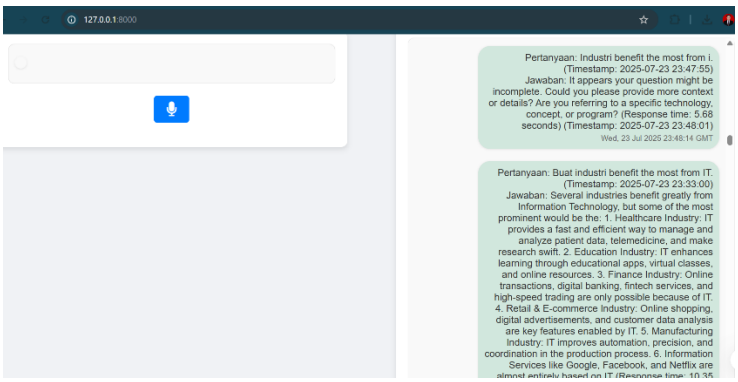

Fig 3. Interface Voice Input & Response Log

In addition to the main device, the study used a smartphone with the Sensor Logger application to record noise levels and event times during testing, so that environmental conditions were objectively documented and system performance analysis could take the acoustic context into account.

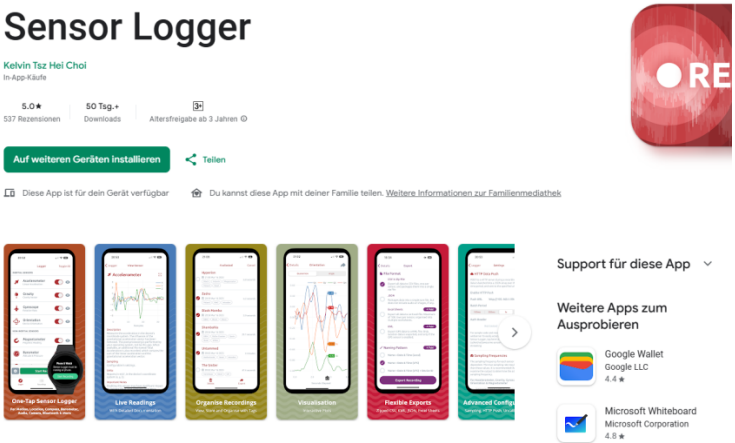

Fig 4. Sensor Logger Application

The Sensor Logger application is used to record noise levels and event times during testing as a supporting research instrument.

Table 3. Spesifikasi Aplikasi Sensor Logger

Parameters	Specifications
Application name	Sensor Logger
Platform	Android
Application type	Sensor logging application
Sensors used	Device microphone
Recorded parameters	Environmental noise level
Units of measurement	dB
Main function	Noise recording and event time recording
Role in research	Environmental control variables

Internal logging systems as well as audio and video recording are used as verification tools to ensure the accuracy of voice recognition results and response times during testing.

E. Testing and Data Collection Procedures

The testing procedure was structured to ensure data consistency and validity. Testing began with recording environmental noise using a Sensor Logger to document baseline conditions. Participants then provided voice commands according to the scenario at distances of 1, 2, and 3 meters in a consistent manner. Recorded data included speech recognition results, system response times, and environmental noise levels, which were systematically collected for analysis.

F. Data Processing and Analysis

Test data was processed by checking for completeness and consistency before being analyzed. Voice recognition performance analysis was performed using the Random Forest method to classify test results into success and failure categories based on attributes such as user distance, response time, and environmental conditions. This method was used as a classification analysis tool to systematically and objectively identify system performance patterns, not to build large-scale predictive models.

G. System Performance Evaluation

The performance evaluation of the speech recognition system was conducted based on data analysis to assess performance at various user distances and compare Amazon Alexa and the Interactive Speaker System. The evaluation included speech recognition accuracy and response time as indicators of interaction quality. The evaluation results serve as the basis for drawing conclusions and developing recommendations for future voice recognition system development.

III. Results and Discussion

1. Results

A. Data collection

Research data was obtained through direct testing involving users giving voice commands to a smart speaker, supported by automatic recording using a sensor logger. Each command was treated as a single trial and systematically documented. The total data amounted to 1020 trials, consisting of 305 successes and 715 failures, with varying user distances to represent different usage conditions. The high failure rate was primarily influenced by the use of a foreign language (English), indicating that linguistic factors play a significant role in speech recognition performance.

B. Data Labeling

Each trial result was first labeled into failure (0) and success (1) categories based on the suitability of the smart speaker's response to the voice command. After labeling, the data was summarized to describe the distribution of test results, as shown in Table 4. with 715 failed trials and 305 successful trials. The predominance of failure results indicates the presence of factors that influence the success of the system, so this summary provides an overview of the characteristics of the dataset and the initial context in interpreting the research results.

Table 4. Trial Label Distribution

Sistem	Label		Total
	0 (Fail)	1 (Pass)	
Alexa	230	280	510
Interactive Speaker System	485	25	510
TOTAL	715	305	1020

C. Feature Extraction

Audio and sensor logger data are collected as raw data representing field conditions, then extracted into statistical features to summarize signal characteristics. The features used include average, standard deviation, stability, and maximum and minimum values to describe signal quality, variation, and interference. This extraction aims to simplify large-dimensional data to make it more structured and reliable to support voice recognition performance analysis.

Table 5. Fitur Statistik

No	Feature	Description	Relevance to Voice Recognition
1	Mean_dBFS	Average user sound level	Indicates how loud/soft a voice is; affects command detection
2	Std_dBFS	Standard deviation of sound level	Measures voice stability; unstable voices can decrease accuracy
3	Var_dBFS	Sound level variance	Measures voice fluctuations compared to the mean; high variance is used for reduced performance
4	Max_dBFS	Highest sound level	Indicates peaks in noise; important to avoid false recognition
5	Min_dBFS	Lowest sound level	Indicates very soft parts of a voice; too low can lead to unrecognition
6	Median_dBFS	Mean of sound level	Provides a stable representation of a voice without being affected by outliers
7	Range_dBFS	Difference between Max_dBFS and Min_dBFS	Measures voice dynamics; a large range is used for fluctuations, a small range is used for stability
8	Q25_dBFS	First quartile (25th percentile) of sound level	Indicates consistently low voice levels
9	Q75_dBFS	Third quartile (75th percentile) of sound level	Indicates frequently high voice levels without extreme outliers
10	Skew_dBFS	Asymmetry of sound distribution	Positive skew is used for low voice dominance, negative is used for high voice dominance
11	Kurtosis_dBFS	Spiciness of sound distribution	High kurtosis is used for many extreme values, can affect predictions
12	Stability_dBFS	Ratio of standard deviation to mean of sound	Measures relative stability; smaller is used for more consistent voices
13	Noise_Level	10th percentile of sound level	Measures low background noise; low noise is used for more accurate predictions

14	Signal_Level	90th percentile of sound level	Measures the strength of the main signal; a high signal is used for clearer recognition
15	SNR_Proxy	Difference between 90th and 10th percentiles	Measures signal quality relative to noise; a high SNR is used for more easily recognized voices
16	Mean_Level	Average user speed	High mobility is used for unstable voices, affecting performance.
17	Noise	Standard deviation of horizontal position	Measuring position uncertainty; inaccurate location can affect interactions.
18	Mean_Speed	Average Wi-Fi signal quality	Slow/unstable connections are used for degraded voice recognition.
19	Mean_Horizontal_ Accuracy	Horizontal accuracy of Wi-Fi/GPS sensor	Low accuracy is used for incorrect speaker/user positioning, which can lead to degraded recognition.
20	DISTANCE (Meters)	Physical distance between user and speaker	Long distance is used for weakened voices, which can lead to reduced recognition accuracy.
21	system_code	Numerical code of smart speaker type	Helps the model distinguish the characteristics of each speaker.

Presents a summary of the results of audio statistical feature extraction for each system and user combination, including Mean, Std, Var, Max, Min, Median, and Range dBFS. These features represent the intensity, variation, and dynamics of the audio signal used to describe the sound quality as input in the speech recognition process.

Table 6. Audio Feature Summary Example

System	User	Mean_dBFS	Std_dBFS	Var_dBFS	Max_dBFS	Min_dBFS	Median_dBFS	Range_dBFS
Alexa	Asrul	-86.31	41.09	1688.52	-24.0	-159.0	-94.0	135.0
Alexa	Asrul	-86.31	41.09	1688.52	-24.0	-159.0	-94.0	135.0
Alexa	Asrul	-86.31	41.09	1688.52	-24.0	-159.0	-94.0	135.0
Alexa	Asrul	-86.31	41.09	1688.52	-24.0	-159.0	-94.0	135.0
Alexa	Azmi	-92.36	27.17	738.26	-27.0	-139.0	-94.0	112.0
Alexa	Azmi	-92.36	27.17	738.26	-27.0	-139.0	-94.0	112.0
Alexa	Azmi	-92.36	27.17	738.26	-27.0	-139.0	-94.0	112.0
Alexa	Azmi	-92.36	27.17	738.26	-27.0	-139.0	-94.0	112.0
…….	……	……..	…….	…….	……...	……...	……	……...
Alexa	Karina	-93.184	27.10	734.80	-28.0	-160.0	-95.5	132.0
Alexa	Karina	-93.184	27.10	734.80	-28.0	-160.0	-95.5	132.0
Alexa	Karina	-93.184	27.10	734.80	-28.0	-160.0	-95.5	132.0

Audio features represent sound quality and fluctuations that affect speech recognition accuracy, while network, location, and Wi-Fi features are analyzed using their respective average and stability indicators. All features from various sources are combined into a multimodal table, then normalized and the speaker system is numerically encoded. This process produces a dataset with 20 predictor features ready for Random Forest training.

D. Analysis and Classification

The boxplot of Mean_dBFS against distance shows that Alexa maintains a higher and more stable sound level than the Interactive Speaker System, especially at distances above 2 meters.

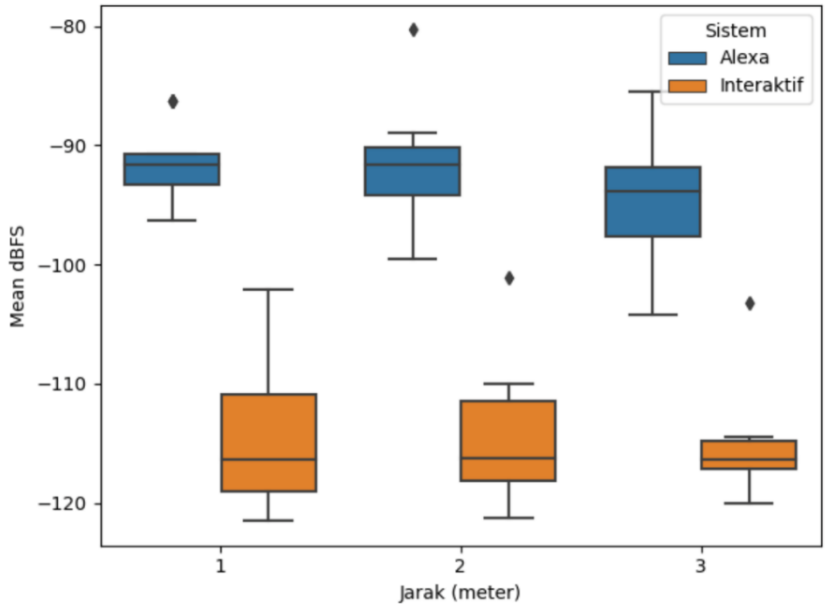

Fig 5. Audio Signal Degradation with Distance

Audio signal degradation shows that the Mean dBFS value decreases with increasing user distance, indicating attenuation of the voice signal. On Alexa, the Mean dBFS decreases from −91.75 dBFS (1 m) to −94.73 dBFS (3 m) with increasing signal variation, indicating instability due to noise and voice attenuation. Meanwhile, the Interactive system has a lower Mean dBFS (approximately −114 dBFS) at all distances, indicating a weaker signal and potentially lower signal-to-noise ratio. This difference indicates that device design and microphone sensitivity influence the system's ability to maintain signal quality over long distances.

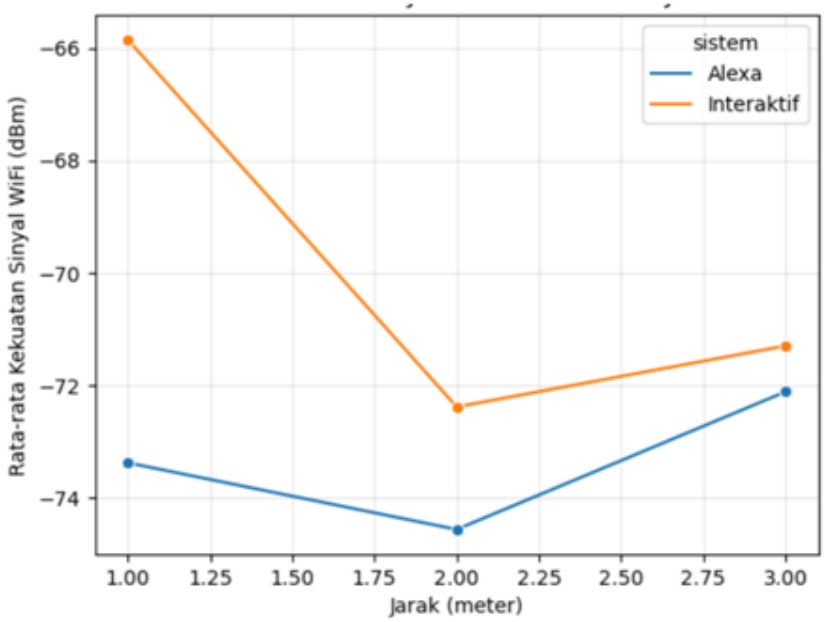

Fig 6. WiFi Signal Strength Trends by Distance

WiFi signal strength trends show that the Alexa system maintains a relatively stable Mean_Speed in the range of −73 to −72 dBm across all distances with zero variation, indicating signal stability. The Interactive system has a stronger signal at 1 meter, but with greater variation, then decreases and stabilizes and approaches Alexa at 2–3 meters. In general, WiFi signal strength decreases slightly with distance, with Alexa being more stable and Interactive being stronger but fluctuating at closer distances.

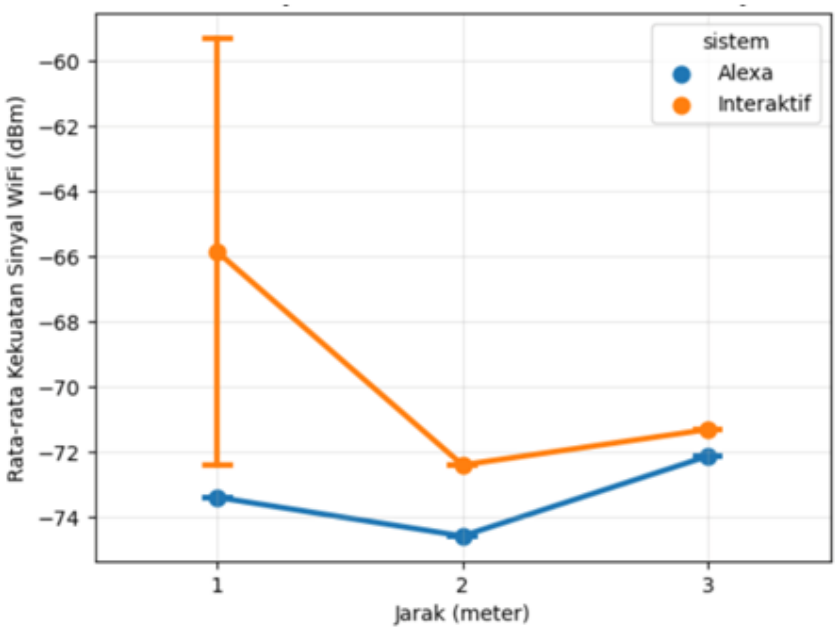

Fig 7. Average WiFi Signal Strength ± SD based on Distance

The Mean ± SD summary shows that at 1 meter, the Interactive system has a stronger WiFi signal than Alexa, but with higher variability, while Alexa is more stable despite its weaker signal. At 2–3 meters, the signal strength of both systems becomes similar and stable with an SD of zero. In general, distance causes a decrease in WiFi signal, with the Interactive system excelling at short distances but fluctuating, while Alexa is consistent across distances.

E. Model Evaluation and Interpretation of Results

Random Forest code is used for classification with Stratified K-Fold Cross-Validation evaluation and performance measurement using accuracy, F1-macro, and recall-macro metrics.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import StratifiedKFold, cross_validate

features = [
   "Mean_dBFS", "Std_dBFS", "Var_dBFS", "Max_dBFS", "Min_dBFS", "Median_dBFS", "Range_dBFS",
    "Q25_dBFS", "Q75_dBFS", "Skew_dBFS", "Kurtosis_dBFS", "Stability_dBFS",
    "Noise_Level", "Signal_Level", "SNR_Proxy",
    "Mean_Level", "Noise",
    "Mean_Speed", "Mean_Horizontal_Accuracy",
    "JARAK (Meter)", "sistem_code"
]

X = df_merge[features]
y = df_merge["Label"]

for col in features + ["sistem_code"]:
    X[col] = pd.to_numeric(X[col], errors="coerce")

mask = X.notna().all(axis=1) & y.notna()
X_clean = X[mask]
y_clean = y[mask]

cv = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)

scoring = {
    "accuracy": "accuracy",
    "f1_macro": "f1_macro",
    "recall_macro": "recall_macro"
}

rf = RandomForestClassifier(
    n_estimators=200,
    random_state=42,
    class_weight="balanced"
)
scores = cross_validate(
    rf,
    X_clean,
    y_clean,
    cv=cv,
    scoring=scoring,
    return_train_score=False
)

for metric in scoring.keys():
    print(metric, ":", round(np.mean(scores[f"test_{metric}"]), 3))
```

Fig 10. Code Random Forest

The dataset was evaluated using stratified k-fold (3 folds), with a Random Forest model containing 200 estimators and assessment metrics in the form of accuracy, F1-macro, and recall-macro.

Table 7. Random Forest Cross-Validation Results

Evaluasi	Skor
Accuracy	0.98
F1_macro	0.98
Recall_macro	0.98

The confusion matrix shows that the Random Forest model is very reliable, with the classification of failed trials having a precision of 0.99 and a recall of 1.00, and successful trials with a precision of 1.00 and a recall of 0.98. Feature importance analysis shows that audio features such as Mean_dBFS, SNR_Proxy, and Stability_dBFS provide the largest contribution, followed by network and Wi-Fi features, confirming that sound quality and connection stability are the main factors in voice recognition performance.

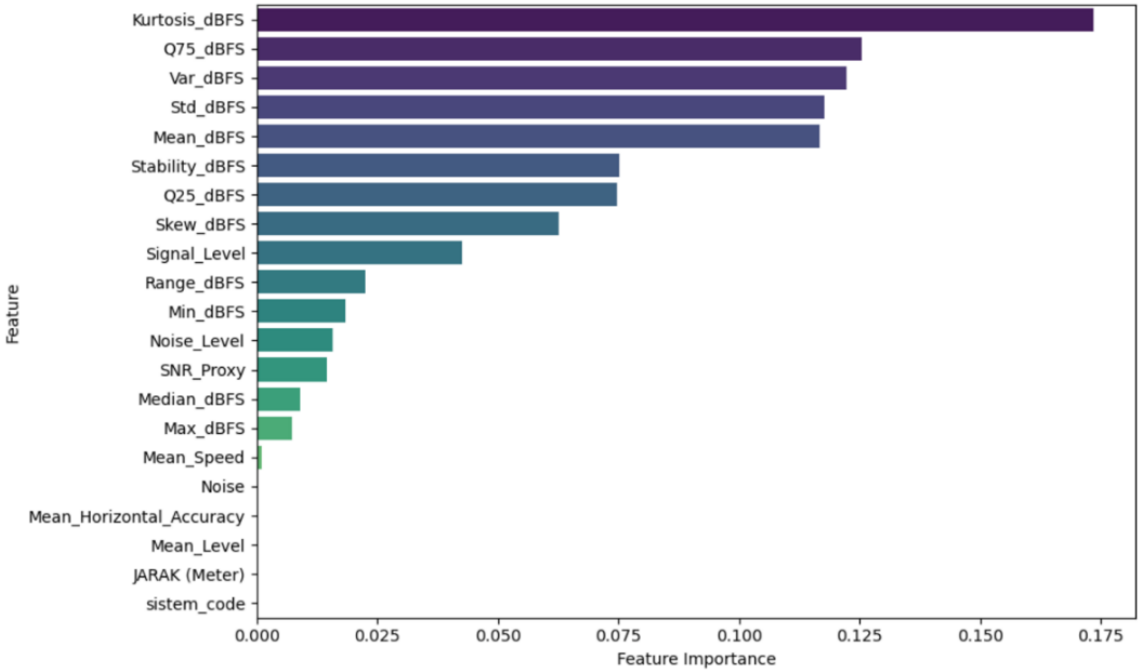

Fig 11. Feature Importance Random Forest

2. Discussion

The feature importance results show that audio features dominate the determination of voice recognition performance, with Kurtosis_dBFS as the most significant factor, followed by Q75_dBFS, Var_dBFS, Std_dBFS, and Mean_dBFS, which confirms the importance of quality, consistency, and distribution of sound levels. Other audio features contribute less, while location, speed, distance, and system code features have minimal influence. In addition to classification, Random Forest Regressor is able to predict Mean_dBFS very well ($R^2 = 0.982$; MAE = 0.05; RMSE = 0.082), indicating the model is accurate and precise in representing variations in audio signals.

```python
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
import numpy as np

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)

pipeline = Pipeline([
    ("imputer", SimpleImputer(strategy="median")),
    ("model", RandomForestRegressor(
        n_estimators=300,
        random_state=42
    ))
])

pipeline.fit(X_train, y_train)

y_pred = pipeline.predict(X_test)

print("R² :", round(r2_score(y_test, y_pred), 3))
print("MAE:", round(mean_absolute_error(y_test, y_pred), 3))
print("RMSE:", round(np.sqrt(mean_squared_error(y_test, y_pred)), 3))
```

Fig 12. Code R^2, MAE (Mean Absolute Error) RMSE (Root Mean Squared Error)

The results of the study indicate that voice recognition performance is influenced by user distance, audio quality, and network and Wi-Fi stability. The Random Forest algorithm is able to integrate all these variables and produce high classification accuracy (accuracy, F1-macro, and recall-macro each 0.98), effectively distinguishing success from failure at various distances and network conditions. These findings confirm the achievement of the research objectives and the effectiveness of Random Forest as an analytical approach in evaluating voice recognition performance based on multimodal data.

IV. Conclusion

Penelitian ini menganalisis kinerja pengenalan suara pada Amazon Alexa dan Interactive Speaker System menggunakan 1020 data percobaan melalui pendekatan Random Forest. Hasil menunjukkan bahwa kinerja sistem dipengaruhi oleh jarak pengguna, kualitas audio, kestabilan jaringan, dan sinyal Wi-Fi, dengan Interactive Speaker System memiliki sinyal lebih kuat namun lebih fluktuatif pada jarak dekat, sedangkan Alexa lebih stabil. Random Forest terbukti sangat efektif dalam klasifikasi performa (accuracy, F1-macro, dan recall-macro = 0,98), dengan fitur audio sebagai faktor dominan, khususnya Kurtosis_dBFS. Selain itu, model regresi mampu memprediksi Mean_dBFS dengan akurasi sangat tinggi (R^2 = 0,982), menyatakan bahwa kualitas dan karakteristik audio merupakan penentu utama keberhasilan pengenalan suara.

References

[1] A. F. Salsabila and R. Rehnaningtyas, "Pengaruh Revolusi Industri 4.0 terhadap hubungan komunikasi antarmanusia dalam implikasi perubahan sosial di era digital," *J. Pendidik. dan Ilmu Sos. (JUPENDIS)*, vol. 2, no. 1, pp. 68–87, Nov. 2023, doi: 10.54066/jupendis.v2i1.1180.

[2] K. Valendra, Tasmi, and C. Setiawan, "Pengembangan sistem pendeteksi kebisingan otomatis pada perpustakaan menggunakan Google Assistant dan ESP32 berbasis voice recognition," *J. Intell. Netw. IoT Glob.*, vol. 2, no. 1, pp. 8–17, Jul. 2024, doi: 10.36982/jinig.v2i1.4436.

[3] K. S. B. Prakoso and B. H. Prasetio, "Implementasi speech recognition berbasis Raspberry Pi 5 pada ekosistem smart-home menggunakan algoritma gated recurrent unit (GRU)," unpublished.

[4] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples, "Voice interfaces in everyday life," in *Proc. 2018 CHI Conf. Human Factors Comput. Syst.*, Montreal, QC, Canada, Apr. 2018, pp. 1–12, doi: 10.1145/3173574.3174214.

[5] K. Amin, L. Elvitaria, and L. Trisnawati, "Artificial intelligence automatic speech recognition (ASR) untuk pencarian potongan ayat Al-Qur'an," unpublished.

[6] L. A. Syahputra, F. R. Prasetya, and A. F. Laila, "Optimalisasi pengendalian smart home melalui teknologi prototipe hand tracking dan speech recognition," *Repeater Publ. Tek. Inform. dan Jar.*, vol. 3, no. 1, pp. 146–159, Jan. 2025, doi: 10.62951/repeater.v3i1.366.

[7] A. Palanica, P. Flaschner, A. Thommandram, M. Li, and Y. Fossat, "Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey," *J. Med. Internet Res.*, vol. 21, no. 4, p. e12887, Apr. 2019, doi: 10.2196/12887.
[8] C. J. K. Fouodo, L. L. Kronziel, I. R. König, and S. Szymczak, "Effect of hyperparameters on variable selection in random forests," arXiv:2309.06943, Jan. 25, 2025, doi: 10.48550/arXiv.2309.06943.
[9] M. Nayeem, M. S. Tabrej, K. J. Deb, S. Goswami, and M. A. Hakim, "Automatic speech recognition in the modern era: Architectures, training, and evaluation," arXiv:2510.12827, Oct. 11, 2025, doi: 10.48550/arXiv.2510.12827.
[10] A. B. Kocaballi *et al.*, "The personalization of conversational agents in health care: Systematic review," *J. Med. Internet Res.*, vol. 21, no. 11, p. e15360, Nov. 2019, doi: 10.2196/15360.
[11] N. A. Yardi, "Survei algoritma pemrosesan bahasa pada Bisindo," vol. 2, 2023.
[12] F. Fitroh, J. Nurhidayah, and Z. Zulfiandri, "Tren dan tantangan arsitektur komputasi neuromorfik: Tinjauan literatur sistematis," *J. Teknol. Sist. Inf.*, vol. 6, no. 1, pp. 103–113, Apr. 2025, doi: 10.35957/jtsi.v6i1.10046.
[13] Z. Tang and D. Manocha, "Scene-aware far-field automatic speech recognition," arXiv:2104.10757, Apr. 21, 2021, doi: 10.48550/arXiv.2104.10757.
[14] D. Zhao *et al.*, "A unified multichannel far-field speech recognition system: Combining neural beamforming with attention-based end-to-end model," arXiv:2401.02673, Jan. 05, 2024, doi: 10.48550/arXiv.2401.02673.
[15] T. Niu, Y. Chen, D. Qu, and H. Hu, "Enhancing far-field speech recognition with mixer: A novel data augmentation approach," *In Review*, Oct. 14, 2024, doi: 10.21203/rs.3.rs-5188489/v1.
[16] H. Dzulfikar, "Perbandingan tingkat kemiripan antara suara langsung dan suara buatan menggunakan metode MFCC, DTW dan KNN untuk mendukung analisa audio forensik," 2021.
[17] Feriman and B. Santoso, "Sistem kontrol elektronik pada rumah pintar dengan input suara pada module pengenalan suara V3 berbasis IoT," *Indones. J. Comput. Sci.*, vol. 13, no. 5, Oct. 2024, doi: 10.33022/ijcs.v13i5.4398.
[18] F. H. Putra, A. R. Albar, A. S. N. Akbar, and A. Kurniawan, "Analisis efektivitas pengenalan perintah suara menggunakan algoritma machine learning untuk mendukung pembelajaran digital di wilayah Bengkulu," vol. 4, no. 4, 2025.
[19] A. W. S. Nuraris and K. Prawiroredjo, "Sistem kendali smart home berbasis Android voice command melalui Bluetooth," no. 1, 2020.
[20] S. Amaliah, M. Nusrang, and A. Aswi, "Penerapan metode random forest untuk klasifikasi varian minuman kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 4, no. 3, pp. 121–127, Dec. 2022, doi: 10.35580/variansiunm31.