JISA (Jurnal Informatika dan Sains)
Vol. 09, No. 01, June 2026

e-ISSN: 2614-8404
p-ISSN:2776-3234

Accuracy Comparison of Support Vector Machine, Random Forest, and K-Nearest Neighbors for Sundanese Speech Classification

Laela Nur Rohmah[1*)], Abdul Halim Anshor[2], Wahyu Hadikristanto[3]
[1]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[2]Program Studi Sistem Informasi, Fakultas Teknik, Universitas Pelita Bangsa
[3] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
Email: [1]laelanurrohmah85@gmail.com, [2]abdulhalimanshor@pelitabangsa.ac.id, [3]wahyu.hadikristanto@pelitabangsa.ac.id

Abstract – To support the preservation of the Sundanese language, speech recognition systems based on machine learning canbe developed. This study aims to evaluate and compare the classification performance of Support Vector Machine, Random Forest, and K-Nearest Neighbors which represent margin-based, ensemble-based, and distance-based classification approaches that have been widely applied in speech classification tasks. A secondary dataset consisting of 100 voice recordings was utilized in this research. The study followed the Knowledge Discovery in Database framework, which encompasses data selection, preprocesing, transformation, data mining, and evaluation phases. Feature extraction was performed using the Mel-Frequecy Cepstral Coefficients method. Experimental result demonstrate that the Random Foret algorithm achieved superior performance, reaching 100% accuracy and an Area Under Curve (AUC) of 100%. Meanwhile, K-Nearest Neighbors achieved 87% accuracy with an AUC of 100%, and Support Vector Machine yielded the lowest performance with 67% accuracy and an AUC of 72.89%. Although Random Forest achieved the highest metrics, futher research is required as a perfect 100% score raises concerns regarding model overfitting. To address this issue, utilizing a large dataset is recommended for future studies. Consequently, K-Nearest Neighbors can be considered a more reliable choice in this study, demonstrating robust and stable performance for MFCC based speech classification on smaller dataset.

Keywords – Speech Classification, Sundanese Language, Support Vector Machine, Random Forest, K-Nearest Neighbors, MFCC, Machine Learning.

I. INTRODUCTION

Indonesia is a country rich in linguistic diversity, possessing a total of 718 indigenous languages spread across its region. These local languages server not only as communication tools but also as vital cultural heritages that reflect regional identity. However, these languages currently face serious threats of extinction. According to the United Nations (UN) and UNESCO, one indigenous language disappears every two weeks due to the declining number of active speakers capable of maintaining its usage [1].

The Sundanese language, a major regional language in West Java, has experienced a notable decline in the number of speakers. According to data from the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek), reffering to the 2010-2020 Central Bureau of Statistics (BPS) census, indicates a significant decline in the use of the Sundanese language [2]. Although the Language Development and Fostering Agency has initiated various revitalization programs, their long-term sustainability remains heavily dependent on the community's commitment to daily usage.

Technological advancement offer new opportunities for language preservation. Artificial Intelligence has significantly contributed to various sectors, including healthcare, education, industry, and information systems [3][4]. One of the important implementations of artifacial intelligence is speech recognition systems, wich enable machine to understand, recognize, and respond to verbal instructions from users. This technology is widely utilized in virtual assistants, navigation systems, Internet of Things (IoT) devices, and automated service systems [5].

Nevertheless, the development of current speecg recognition technology remains focused on major global languages such as English, Mandarin, and Spanish. Local dialect like Sundanese, particulary the West Sundanese and South Sundanese dialect has recive disproportionately less attention in speech recognition development[6]. Current models face limitations in distinguishing acoustic variations between dialect, resulting in low accuracy and a lack of inclusivity in voice based systems [7][8]. This condition presents a challenge in creating speech recognition systems that can accurately accommodate linguistic diversity.

Machine learning is a branch of artificial intelligence that plays a ceucial role in developing voice classification systems. Classification is the process of grouping data based on specific patterns to distinguish characteristic between object [9][10]. In speech recognition, machine learning is used to identify voice patterns and distinguish dialect characteristic based on features. Various machine learning algorithms can be used in the classification, such as Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN). Each algoritm has different characteristic, advantages, and disadvantages in handling voice data.

Support Vector Machine (SVM) is a supervised learning algorithm that functions by identifying the optimal hyperlane to separate data classes [11]. SVM has advantages in handling complex data and producing high accuracy, especially for high dimensional data[12]. Random forest (RF) is an ensemble learning algorithm that construct multiple decision trees to produce more stable and accurate prediction[13]. It is highly effective in managing large dataset, handling missing values, and performing automated feature selection. Meanwhile, K-Nearest Neighbors 9KNN0 is a non-parametic algorithm that classifies data based on proximity using Euclidean Distance[14][15]. KNN is recognized for its simplicity, ease of implementation, an effectiveness in pattern similarity based classification[16].

In speech clasiffication research, features extraction is an important stage. One of the most widely adopted methods is Mel-Frequency Cepstral Coefficients (MFCC), which is a technique that transform raw audio signals into numerical representations relevant to human auditory perception[17].MFCC efectively captures essential voice signal characteristics making it highly suitable for machine learning based speecg recognition[18].

Previous studies have explored the classification of Sundanese speech using machine learning. Anshor and Wiyatno[19] compared Random Forest and SVM for Sundanese dialect classificatin using MFCC features and ROC-AUC evaluation. Similarly, Shandy et al [20] utilized KNN for MFCC based Sundanese dialect classification. However, this studies were limited to the evaluation of only one or two algorithm and lacked a comprehensive comparison between SVM, Random Forest, and KNN on the same dataset

Based on this background, this study aims to perform a accuracy comparison of Support Vector Machine, Random Forest, and K-Nearest Neighbors for Sundanese speech classification. The research utilizes a public dataset titled "Suara Dialek Sunda", containing recordings of West and South Sundanese dialect. Following feature extraction via MFCC, the three algorithms are tested and compared based on accuracy, precision, recall, f1-score, and ROC-AUC metrics.

The result of this research are expected to contribute to identifying the most effective classification method for Sundanese speech recognition and support the development of indigenous language-based speech technology. In addition, this study is also expected to become one of the efforts to utilized technology in preserving local culture in the digital era.

II. RESEARCH METHODOLOGY

The research methodology follows a data mining approach based on the Knowledge Discovery in Database (KDD) framework. This method was selected for its systematic stages in data processing, beginning with data selection and proceeding through preprocessing, transformation, modeling, to the evaluation of classification results[21]. The KDD framework is particulary suitable for this research, as raw audio data requires a comprehensive feature extraction process before it can be effectively utilized by machine learning models.

This research aims to compare the performance of three classification algorithm, Support Vector Machine (SVM), Random Forest, and K-nearest Neighbors (KNN), in classifying Sundanese speech based on the West Sundanese and South Sundanese dialect. the main focus is to identify which classification method yields the highest accuracy for Sundanese speech recognition.

The research is use a public dataset titled "Suara Dialek Sunda' which is available on the Kaggle platform. The dataset contains 100 voice sample in .flac format representing two dialect classes, West Sundanese dialect and South Sundanese dialect. this data was selected because it is relevant to the research objective of identifying differences in voice characteristics based on dialect.

The overall research flow can be illustrated through the flowchart in Figure 1. Each step is designed to ensure the research process runs systematically and objectively, in orde to obtain accurate and accountable results.

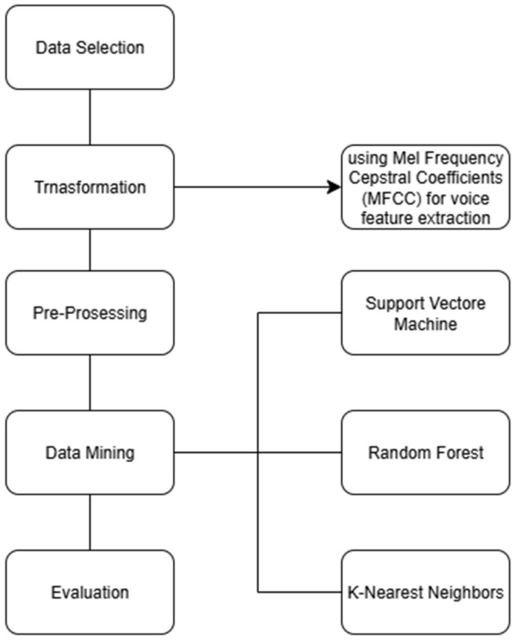

Figure 1. Research flow diagram

The detailed explanation for each stage in the research flowchart is as follows:

1. **Data selection:** This first stage is selection and collection data of Sundanese voice recording. The dataset comprises 100 audio files obtained from a public dataset on the Kaggle platform.
2. **Transformation (Feature Extraction):** This stage is conducted to transform the data into a format suitable for the research requirements. In this research, raw audio data is converted into a numerical representations. The Mel-Frequency Cepstral Coefficients (MFCC) method is utilized at this stage due to its capability to capture the essential characteristics of speech signal. The ouput of this transformation will be divided into training data and testing data, which are subsequently processed in the machine learning classification phase. The MFCC feature extraction process consists of several sequential stages, as follows[18]:

e-ISSN: 2614-8404
p-ISSN:2776-3234

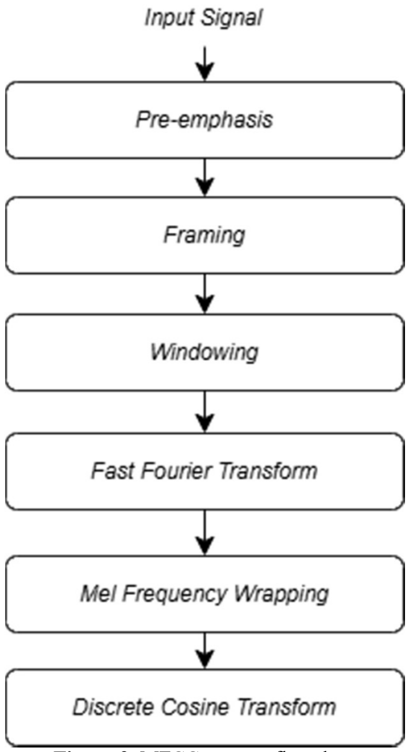

Figure 2. MFCC process flowchart

a) Pre-emphasis: is a filtering process aimed at smoothing the spectral form of the speech signal and reducing noise interference during signal acquisition. This process is defined in equation (1).

$$y[n] = x[n] - ax[n-1] \quad (1)$$

Description:
$y[n]$ = sinyal output
$x[n]$ = sinyal input
a = konstanta fitur dengan karakteristik 0.9 < a < 1.0

b) Farming: Framing, or Frame Blocking, is the process of segmenting the continuous signal into small, overlapping sections. The speech signal is converted into several short-duration frames to ensure the signal remains quasi-stationary.

c) Windowing: Windowing involves applying a weighting function to each frame generated in the previous stage. This process aims to minimize signal discontinuities at the beginning and end of each frame caused by frame blocking. The equation used for the windowing stage is as follows (2).

$$x(n) = x\_i\ (n)w(n), \;\; n = 0,1,2, \ldots, N \quad (2)$$

Description:
$x(n)$ = Signal value after windowing
x_i = Signal value from the i-th frame
$w(n)$ = Windowing function (typically Hamming window)
N = sample index within each frame

d) Fast Fourier Transform (FFT)
FFT is widely used in digital signal processing to convert a signal from the time domain to the frequency domain. The mathematical formula used is as follows (3).

$$S(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \quad (3)$$

Description:
S(k) = k-th FFT calculation result
x(n) = n-th windowing output
k = Frequency index (0, 1, 2, … , N)
N = Number of samples to be processed

e) Mel-Frequency Wrapping: This stage involves a filtering process to determine the energy magnitude of specific frequency bands within the speech signal. The power spectrum is passed through a Mel filter bank, which consists of several triangular filters distributed non-linearly along the Mel frequency scale. The Mel scale is a logarithmic frequency scale that more closely aligns with human auditory perception.

f) Discrete Cosine Transform (DCT): DCT is the final step in MFCC feature extraction. The fundamental concept of DCT is to decorrelate the Mel spectrum to produce a robust representation of local spectral properties. In this stage, the Mel spectrum values in the frequency domain are converted back to the time domain to obtain the coefficients. DCT is applied to the log-energy to compress the information into a set of cepstral coefficients. Typically, the first 12–13 coefficients are used for the final MFCC representation. The DCT process can be expressed in Equation (4).

$$c(k) = 2 \sum_{n=0}^{N-1} x(n) \cos \frac{\pi(2n+1)k}{2N} \quad (4)$$

Description:
x(n) = Final MFCC coefficients
N = Number of desired coefficients

3. Pre-Processing
The objective of this stage is to improve data quality and produce a more accurate and efficient analytical model. In this study, pre-processing is conducted to handle inaccurate, incomplete, or inconsistent data, such as addressing missing values, reducing noise, and resolving data inconsistencies.

4. Data Mining
At this stage, classification models are training using three machine learning algorithms: Support Vector Machine, Random Forest, and K-Nearest Neighbors. Prior to the training process, the

e-ISSN: 2614-8404
p-ISSN:2776-3234

preprocessed dataset is partitioned into training data and testing data with a 70:30 ratio. The training data is utilized to construct the model and learn the characteristic patterns of the sundanese dialect, while the testing data is used to evaluate the model clasiffication performance on unseen data that was not involved in the training phases. This data splitting is implemented to measure the algorithms capability to recognize and distinguish sundanese dialect.

These three algorithms employed in this study utilize different classification approaches. Support Vector Machine (SVM) is a learning algorithm that fungtions by finding an optimal hyperplane with a maximum margin between classes, enabling effective classification in high-dimensional feature spaces. Random Forest is an ensamble classification algorithm that construct multiple decision tree from randomly selected subsets of data and feature. The final Random Forest prediction is determined through a majority voting mechanism,which contribute to enhancing model stability and reducing the risk of overfitting. Lastly, K-Nearest Neighbirs (KNN) is an instance-based algorithm that perform classification based on the labels of the nearest neighbors using the euclidean distance metric. This three algorithm were selected to compare the performance of margin-based, ensemble-based, and distance -based classification approaches in identifying sundanese speech dialect based on acoustic feature extracted through the Mel-Frequency Cepstral Coefficients (MFCC) method.

5. Evaluation

This final stage is conducted to analyze and assess the performance of the constructed models. Each model is evaluated using the test data based on accuracy as the primary metric, while also considering precision, recall, and the F1-score[22]. Furthermore, this study employs the Receiver Operating Characteristic – Area Under Curve (ROC-AUC) to provide a detailed evaluation of the models' ability to distinguish between the two classes. The AUC value ranges from 0 to 1, where a value closer to 1 indicates superior classification performance.

The evaluation process is implemented using Google Colab, a cloud-based programming environment that supports Python for machine learning[23]. Google Colab is utilized for model training, classification testing, and the visualization of confusion matrices, classification reports, and ROC-AUC analysis. The results from all three algorithms are then compared to determine the most optimal method for Sundanese speech classification. The final conclusion is drawn from these experimental results to support the development of more accurate and efficient indigenous language-based speech recognition systems.

III. RESULTS AND DISCUSSION

Bassed on the classification models implemented using Google Collaboratory (Google Colab), model performance was analuzed using several evaluation metrics, including accuracy, precision, recall, f1-score, ROC-AUC, and confusion matrix analysis. Accuracy was utilized to measure the proportion of correctly classified samples, while precision, recall, and f11-score were employed to assess the classification quality for each dialect class. Additionally, ROC-AUC was used to evaluate the model capability to differentiate between dialect classes across various decision threshold. The comparative performance results of the three machine learning algorithm are presented in Table 1.

Tabel 1. Comparison of Machine Learning Algoritm Performance

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
SVM	67%	69%	60%	64%	72.89%
RF	100%	100%	100%	100%	100%
KNN	87%	100%	73%	85%	100%

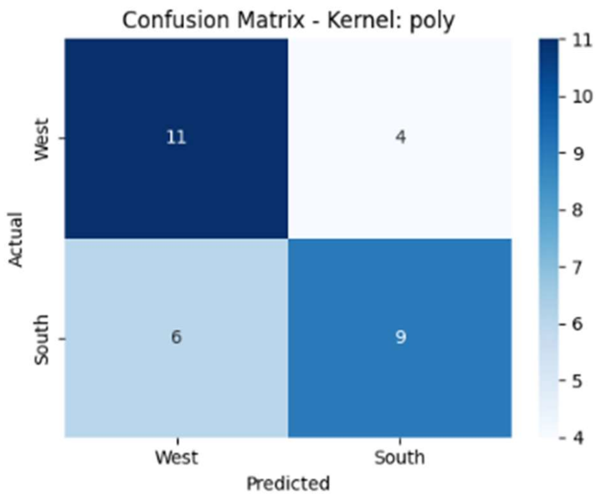

Figure 3. Confusion Matrix of Support Vector Machine

In figure 3, model SVM accurately classified 11 West Sundanese dialect samples correctly. However, 4 sample were misclassified as South Sundanese dialect. regarding the South dialect, model successfully indentified 9 instances correctly, while 6 samples were missclassified as the West dialect. The confusion matrix results indicate that while the SVM model is capable of correctly identifying a portion of the dialect samples, significant misclassifications still occur across both classes. This finding demonstrates the model's limited capability to differentiate between Western and Southern Sundanese dialects, which aligns with its lower accuracy and ROC-AUC scores compared to those of Random Forest and K-Nearest Neighbors (KNN).

e-ISSN: 2614-8404
p-ISSN:2776-3234

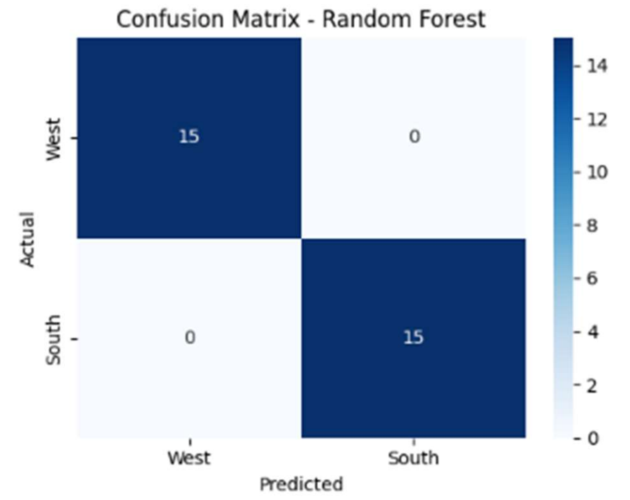

Figure 4. Confusion Matrix of Random Forest

In figure 4, model RF show that all data were succesfully predicted correctly. Out of the 15 class data sample, all were accurately classified as west dialect. similary for the 15 south class data sample, all were successfully predicted as south. These results indicate that the model's excellent capability to differentiate between Western and Southern Sundanese dialects based on the extracted MFCC features. Such flawless classification performance also aligns with the obtained accuracy and ROC-AUC scores, which are higher than those of SVM and K-Nearest Neighbors (KNN).

The following method is K-Nearest Neighbors. This algorithm works by classifiying data based on the nearest distance between features. Within thw KNN framework, the following steps are performed after the data splitting process:

1. Determining the k-value: There are no fixed rules for determining the k-value,ini this research the k-value is set to 10.
2. Euclidean Distance Calculation: This step involves calculating the distance between the training data and the testing data. In this research, 15 training data and 1 testing data sample are utilizwd as a sample.

Table 2. Euclidean Distance Calculation

Data training	Euclidean Distance Calculation
D86	$\sqrt{(-254.00989) - (-163.97807)^2} + ((88.732315) - (34.690945))^2 + ((-40.110630) - (-27.213692))^2 + ((30.947512) - (22.054520))^2 + ((-1.890005) - (-13.354515))^2 + ((-2.493251) - (14.781217))^2 + ((7.844913) - (-12.342719))^2 + ((-10.974412) - (1.776014))^2 + ((1.200290) - (-10.514988))^2 + ((-7.606719) - (0.159154))^2 + ((-2.069135) - (-0.354056))^2 + ((1.658952) - (3.970661))^2 + ((-5.120046) - (-7.359786))^2 = 111.724$
D54	$\sqrt{((-326.57794) - (-163.97807))^2} + ((51.159830) - (34.690945))^2 + ((-29.254915) - (-27.213692))^2 + ((45.570140) - (22.054520))^2 + ((-6.977783) - (-13.354515))^2 + ((1.615259) - (14.781217))^2 + ((8.016255) - (-12.342719))^2 + ((-11.957951) - (1.776014))^2 + ((8.280593) - (-10.514988))^2 + ((-6.925526) - (0.159154))^2 + ((-1.302201) - (-0.354056))^2 + ((4.698872) - (3.970661))^2 + ((-4.145867) - (-7.359786))^2 = 168.817$
....	....
D92	$\sqrt{((-266.93310) - (-163.97807))^2} + ((74.309430) - (34.690945))^2 + ((-31.016413) - (-27.213692))^2 + ((42.155600) - (22.054520))^2 + ((-12.889493) - (-13.354515))^2 + ((3.149026) - (14.781217))^2 + ((12.862194) - (-12.342719))^2 + ((-11.102824) - (1.776014))^2 + ((7.251139) - (-10.514988))^2 + ((-3.911420) - (0.159154))^2 + ((-2.561542) - (-0.354056))^2 + ((-1.291445) - (3.970661))^2 + ((-1.718913) - (-7.359786))^2 = 117.987886$

Subsequently, the result of euclidean distance calculation are sorted in ascending order from the shortest to the farthest distance.

Table 3. Ascending Result of Euclidean Distance

Rank	**X_train_index**	Result of Euclidean Distance	**Actual_Dialect_Name**	***K=10***
1	33	9.520783	West	1
2	31	10.680736	West	2
3	39	19.087149	West	3
4	12	83.056314	West	4
5	62	90.482413	South	5
6	65	91.426549	South	6
7	11	95.208741	West	7
8	64	98.675702	South	8
9	9	100.898593	West	9
10	6	103.368218	West	10
11	86	111.724792	South	
12	83	114.864435	South	
13	92	117.987880	South	
14	72	137.177135	South	
15	48	144.635470	West	

Based on table 3, which present the prediction results from the calculation of one testingd ata sample using a value of k=10, identified 7 of the west dialect an 3 of the south dialect among the nearest neighbors. This indicates that the prediction result is consistent with the actual label of testing data, which is the west dialect.

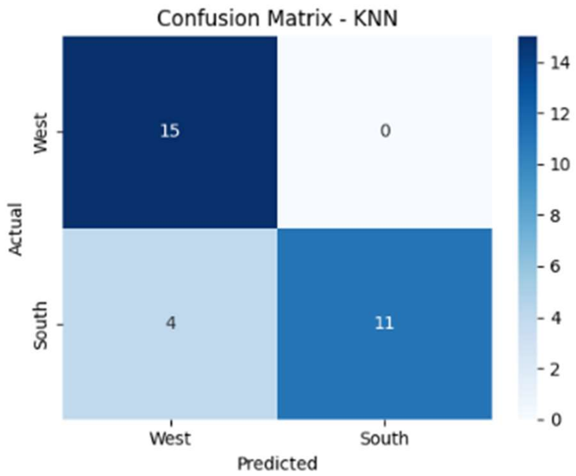

Figure 5. Confusion Matrix of K-Nearest Neighbors

In figure 5, model KNN show that successfully classified all 15 west dialect sample correctly. However for the south dialect, the model succesfully identified only 11 sample correctly. These results indicate that the KNN model has excellent performance in recognizing the west class, but it still encounters classification errors within in a porpotion of the south class data.

IV. CONCLUSION

Based on the experimental result of the three algorithms, random forest achieved the highest performance with accuracy and ROC-AUC of 100%. However achieved perfect accuracy and ROC-AUC score, the random forest model need to be analyzed futher since the dataset were used in this research relatively small. Such verification is essential to ensure that the model does not suffer from overfitting and maintains its generalizability to new unseen data.

In the second position the KNN model yielded an accuracy of 87% with an AUC of 100%. Although some misclassification occurred dureng testing, the AUC value suggest that the model has very strong performance because ROC-AUC evaluates not only the classification result at a single decision threshold but also the overall performance across various thresholds.

Finally, the SVM algorithm showed the lowest performance with an accuracy of 67% and an ROC-AUC of 72.89%. This indicates that the polynomial kernel utilized in this research was not yet optimal in mapping the MFCC features to effectively separate the classes. Overall, the comparison of these three algorithms demostrates that the choice of algorithm significantly influences the success rate of speech classification models.

For the future research, it is recommended to use a larger speech dataset to more comprehensively evaluate the model generalizability to unseen data. The application of additional validation technuiques is also recommended to obtain more reliable evaluation results and minimize the potential risk of overfitting. Futhermore, future studies could explore alternative speech feature extraction methods or compare the performance of the three algorithms used in this study with other classification model for sundanese dialect classification.

REFERENCES

[1] E. Rusyana and R. U. N. Rohmah, “Interferensi bahasa Indonesia terhadap bahasa Sunda dalam karangan berbahasa Sunda siswa SMP,” *Diglosia J. Kaji. Bahasa, Sastra, dan Pengajarannya*, vol. 7, no. 2, pp. 237–246, 2024, doi: 10.30872/diglosia.v7i2.954.

[2] Pun, “UNESCO: Setiap Dua Minggu, Satu Bahasa Daerah Punah di Dunia,” *Berita*. p. https://jabarprov.go.id/berita/unesco-setiap-dua-m, 2024.

[3] M. Azhar and H. F. Pardede, “Klasifikasi Dialek Pengujar Bahasa Inggris Menggunakan Random Forest,” *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 439, 2021, doi: 10.30865/mib.v5i2.2754.

[4] A. M. Afida, “Klasifikasi Jenis Burung Berdasarkan Suara Menggunakan Algoritme Support Vector Machine,” *MATHunesa J. Ilm. Mat.*, vol. 8, no. 1, pp. 1–6, 2020, doi: 10.26740/mathunesa.v8n1.p1-6.

[5] U. B. Nusantara, “SPEECH RECOGNITION DENGAN WHISPER DALAM BAHASA INDONESIA,” vol. 9, no. 2, pp. 386–397, 2025.

[6] M. A. As Sarofi, I. Irhamah, and A. Mukarromah, “Identifikasi Genre Musik dengan Menggunakan Metode Random Forest,” *J. Sains dan Seni ITS*, vol. 9, no. 1, pp. 79–86, 2020, doi: 10.12962/j23373520.v9i1.51311.

[7] G. Drougkas, E. M. Bakker, and M. Spruit, “Multimodal machine learning for language and speech markers identification in mental health,” *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, 2024, doi: 10.1186/s12911-024-02772-0.

[8] T. Kessler, D. Schneider, and M. Shadaydeh, “A Comparison of Machine Learning Algorithms and Feature Sets for Automatic Vocal Emotion Recognition in Speech,” 2022.

[9] Desi Permatasari and Suliadi, “Algoritma Pemilihan Variabel untuk Klasifikasi dan Penerapannya pada Klasifikasi Desa-Kelurahan,” *J. Ris. Stat.*, pp. 49–56, 2024, doi:

10.29313/jrs.v4i1.3876.

[10] P. Rosyani, S. Saprudin, and R. Amalia, "Klasifikasi Citra Menggunakan Metode Random Forest dan Sequential Minimal Optimization (SMO)," *J. Sist. dan Teknol. Inf.*, vol. 9, no. 2, p. 132, 2021, doi: 10.26418/justin.v9i2.44120.

[11] Y. Supit, "Penerapan Algoritma Support Vector Machine (SVM) untuk Deteksi Penyakit Mata Katarak," *Indones. J. Data Sci.*, vol. 2, no. 2, pp. 87–96, 2025, doi: 10.30989/ijds.v2i2.1449.

[12] S. N. Khairil Lukman, "Komparasi Algoritma Naïve Bayes danSVM untuk Identifikasi CyberbullyingSelebriti di Media Sosial Twitter.pdf." 2025. doi: 10.33364/algoritma/v.22-1.2196.

[13] N. I. Yaman, A. R. Juwita, S. A. P. Lestari, and S. Faisal, "Perbandingan Kinerja Algoritma Decision Tree dan Random Forest untuk Klasifikasi Nutrisi pada Makanan Cepat Saji," *Jurnal Algoritma*, vol. 21, no. 2. pp. 184–196, 2024. doi: 10.33364/algoritma/v.21-2.1649.

[14] Widodo and Supatman, "Klasifikasi Transaksi Obat Puskesmas Bagelen Menggunakan Algoritma KNN," *J. Zetroem*, vol. 6, no. 1, pp. 79–85, 2024, doi: 10.36526/ztr.v6i1.3540.

[15] A. Benita Suciani, G. Gunawansyah, and K. Maulana Adhinugraha, "Klasifikasi Jenis Batik Menggunakan Algoritma K-Nearest Neighbors (Knn)," *Infotronik J. Teknol. Inf. dan Elektron.*, vol. 9, no. 2, pp. 83–89, 2024, doi: 10.32897/infotronik.2024.9.2.3846.

[16] V. Artanti, M. Faisal, and F. Kurniawan, "Klasifikasi Cardiovascular Diseases Menggunakan Algoritma K-Nearest Neighbors (KNN) Classification of Cardiovascular Diseases using K-Nearest Neighbors (KNN) Algorithm," *Techno.COM*, vol. 23, no. 2, pp. 467–479, 2024.

[17] M. Oktaviani, T. E. Sutanto, and M. Mahmudi, "Klasifikasi Usia Berdasarkan Suara Dengan Ekstraksi Ciri Mel Frequency Cepstral Coefficients Menggunakan Support Vector Machine," *Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen)*, vol. 4, no. 4, pp. 901–907, 2023, [Online]. Available: https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/240%0Afiles/5264/Oktaviani et al. - 2023 - Klasifikasi Usia Berdasarkan Suara Dengan Ekstraks.pdf

[18] R. Hidayat and A. Winursito, "A Modified MFCC for Improved Wavelet-Based Denoising on Robust Speech Recognition," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 1, pp. 12–21, 2020, doi: 10.22266/IJIES2021.0228.02.

[19] A. H. Anshor and T. N. Wiyatno, "Comparative Analysis of Random Forest and Support Vector Machine for Sundanese Dialect Classification Using Speech Recognition Features," vol. 14, pp. 269–276, 2025.

[20] E. Shandy, A. H. Anshor, and D. Ardiatma, "Implementation of Data Mining for Speech Recognition Classification of Sundanese Dialect Using KNN Method with MFCC Feature Extraction," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 3, pp. 1145–1158, 2024, doi: 10.47709/cnahpc.v6i3.4226.

[21] P. Angga, R. Herdiana, and D. Solihudin, "Klasifikasi menggunakan metode Naive Bayes.," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 8, pp. 456–460, 2023.

[22] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021.

[23] R. Andarsyah and A. Yanuar, "Sentimen Analisis Aplikasi Posaja Pada Google Playstore Untuk Peningkatan Pospay Superapp Menggunakan Support Vector Meachine," *J. Tek. Inform.*, vol. 16, no. 2, pp. 1–7, 2024, [Online]. Available: https://ejurnal.ulbi.ac.id/index.php/informatika/article/view/3533