Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 3 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

Implementation of Data Mining for Speech Recognition Classification of Sundanese Dialect Using KNN Method with MFCC Feature Extraction

Ery Shandy[1)*], Abdul Halim Anshor[2)], Dodit Ardiatma[3)]
[1)*2)] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa, Bekasi, Indonesia
[3)] Program Studi Teknik Lingkungan, Fakultas Teknik, Universitas Pelita Bangsa, Bekasi, Indonesia
[1)*]eryshandy22@mhs.pelitabangsa.ac.id, [2)]abdulhalimanshor@pelitabangsa.ac.id,
[3)]doditardiatma@pelitabangsa.ac.id

ABSTRACT

The importance of preservation and development of speech recognition technology for regional languages such as Sundanese, which have unique phonetic characteristics. Regional language speech recognition can assist in the development of local, educational, and cultural preservation applications to implement and evaluate the effectiveness of the combination of MFCC and KNN methods in classifying Sundanese dialect speech recognition. Methods used include trait extraction with MFCC, which converts voice data into numerical representations based on frequency characteristics, and classification with KNN, which groups data based on similarity to train data. The Dataset used consisted of speech recordings of Western and Southern Sundanese dialects. The results showed that the k-Nearest Neighbors (KNN) method can classify Sundanese dialect speech recognition with an accuracy of 80.00%, showing good ability in distinguishing "Western" and "southern" dialects. Mel-Frequency Cepstral Coefficients (MFCC) proved to be very effective in extracting sound features, helping KNN achieve low error rates. The combination of MFCC and KNN proved effective for speech recognition classification of Sundanese dialects, providing satisfactory results with high accuracy.

Keywords: K-Nearest Neighbor(K-NN); Data Mining; MFCC; Classification; Sundanese

1. INTRODUCTION

Advances in speech recognition technology have made a significant impact in various aspects of human life, ranging from virtual assistants to language learning software. However, most of the research and development of these technologies focuses on major languages such as English, Mandarin, and Spanish, while regional languages such as Sundanese have not received adequate attention. Sundanese, as one of the largest regional languages in Indonesia with millions of speakers, has unique phonetic and linguistic characteristics that require a special approach in its recognition.

One of the data mining processing techniques is classification. Classification is the process of finding a set of patterns or functions that describe and separate one data class from another, and is used to predict data that does not yet have a specific data class (Setio et al., 2020). Classification is one of the most commonly used data analysis techniques and has wide applications, including in the field of speech recognition. One of the uses of classification is to assist in decision-making regarding the identification and mapping of Sundanese dialects.

Speech Recognition is the ability of a machine to listen to spoken words and identify them. The ability to be able to convert the voice that enters the computer into text form (Adnan et al., 2022). In the context of Sundanese, Sundanese has a wide variety of dialects, including dialects from the Western and Southern regions. This variation creates its own challenges in speech recognition, as each dialect has different phonetic characteristics. Therefore, an effective technique for extracting voice features and a classification method capable of dealing with these variations with high accuracy are required.

One effective technique for extracting voice features is Mel-Frequency Cepstral Coefficients (MFCC). MFCC is one of the feature extraction methods based on human hearing behavior that cannot recognize frequencies greater than 1Khz (Yehezkiel & Suyanto, 2022). MFCC can effectively extract voice features that are important for speech recognition, such as frequency and cepstral coefficient patterns. This allows the system to recognize voices with a high degree of accuracy and MFCC can be used to recognize voice types, such as male and female voices, as well as

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

voices with different pitches (Ajinurseto et al., 2023). MFCC is able to capture spectral information from speech signals which is very important in speech recognition. By using MFCC, the important features of speech in each dialect can be identified and used as input in the classification process.

In this case, Data Mining is very influential in the process of extracting data from a data set in the form of knowledge that has not been known manually. Data mining is the process of finding interesting patterns and knowledge from large amounts of data (Nabila et al., 2021). There are several methods in data mining for classification such as Naïve Bayes, SVM, Decision Tree, K-Nearest Neighbor. In this study using the K-Nearest Neighbor (KNN) method for the classification of Sundanese dialect speech recognition. K-Nearest Neighbor (KNN) is a method for classifying objects based on previous data with the closest distance from the object (Rizky et al., 2021).

Dialect is a characteristic of a person's language because each person has a characteristic in speaking (Suparman, 2023). Dialects include differences in pronunciation, vocabulary, grammar, and often intonation as well as distinctive expressions not found in the standardized form of the language. Sundanese is a language of the Malayo-Polynesian branch of the Austronesian language family. It is spoken by at least 42 million people and is the second most spoken mother tongue in Indonesia after Javanese. The majority of the Sundanese language is spoken by people from West Java and Banten. (Guntara et al., 2021) (Indra Kusuma et al., 2021). Sundanese is a unique language with unique levels of language, or better known as term undak usuk which is almost not owned by other languages. by other languages. However, of the many Sundanese language, not all of us know the elements of know in depth the elements of Sundanese Sundanese language culture. Undak usuk basa can be interpreted as language strata, level of language, or manners in language. Broadly speaking, undak usuk is divided into three parts, namely language, medium language, and coarse language (Komalasari et al., 2021). Sundanese is the "mother" language of Sundanese people or people who were born and raised in West Java. West Java. Sundanese is one of the regional languages with the second largest number of speakers in Indonesia. In Indonesia. Sundanese served as an indigenous language in pre-independence Indonesia. Indonesia. The existence of Sundanese is recognized and protected by the state and is included in the 1945 Constitution Chapter XV Article 36. In the 1945 Constitution Chapter XV Article 36 (Di et al., 2021.).

2. LITERATURE REVIEW

Various methods and techniques have been used in the field of music genre classification, focusing on feature extraction from audio files to identify different genres. One common approach is to use Mel frequency cepstral coefficients (MFCC) combined with a k-Nearest Neighbors (KNN) classifier. The study by Author used this method to classify music genres, achieving an accuracy rate of 52.4% with K=13 as the optimal parameter for nearest neighbors. However, the accuracy for certain genres such as pop, dangdut, and jazz fell below 50%, indicating the need for further improvement in classification accuracy (Deski Prasetyo et al., 2022).

The two-dimensional mood model by Robert Thayer categorizes emotions based on stress (happiness and anxiety) and energy (calm and energetic) into four different quadrants. This model has been used in research to classify musical moods. The research methodology involves feature extraction using MFCC and classification using K-Nearest Neighbor (KNN). Tests were conducted to determine the impact of features and data scaling on the classification of audio signals. The results showed that the MFCC-Delta-Delta2 feature combination yielded the highest accuracy. Standard scaling was identified as the best data scaling method. The effect of the distance method in KNN was also tested, with the Manhattan method giving the highest accuracy. Precision and recall were also considered in the testing process. The KNN method with Manhattan distance gave better results compared to the Euclidean and cosine methods. Accuracy, precision, and recall were 85.5%, 87.34%, and 85.5% with k=5 (Fadlila Surenggana et al., 2022.).

This study explores the use of voice signals using Mel-Frequency Cepstral Coefficients (MFCC) and Learning Vector Quantization (LVQ) methods, achieving 92% accuracy with 165 recognized and 15 unrecognized data. Further testing on new data resulted in 46% accuracy, with 84 recognized data and 96 unrecognized data. This research makes a significant contribution to speech recognition using information technology. Previous research used MFCC and K-Nearest Neighbor (KNN) methods for voice feature extraction and classification. Audio data was recorded from six participants using the keywords "open" and "close." The previous study used LVQ with the highest accuracy at α 0.1. The current study aims to test the accuracy of the audio data in .wav format. The MFCC method involves several stages such as DC Release, Initial Suppression, Frame Blocking, Windowing, FFT, Mel Frequency Warping, DCT,

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

and Cepstral Filtering. K-NN algorithm is used for data classification by finding k nearest neighbors and selecting the class with the highest count (Putu et al., 2022).

3. METHOD

The data used in this research is quantitative data. Quantitative data is data that can be measured and expressed in numbers. This data is numerical and allows for statistical analysis. Quantitative data can be in the form of measurements, numbers, or other numerical values that can be operated mathematically. The sources of data collection carried out by the author to obtain valid data are secondary data and literature studies. Data that has been collected, processed, and presented by other parties or sources other than the researchers conducting the research. This data is usually already available in the form of publications, reports, databases, or other documents and can be reused by researchers for their research purposes. Data collection is done by means of literature studies obtained from the website in the form of voice recording data of western and southern Sundanese dialects. Here is the dataset link https://www.kaggle.com/datasets/fitrohamri08/suara-dialek-sunda.

Feature Extraction with Mel Frequency Cepstral Coefficients

The block diagram for MFCC can be described as follows:

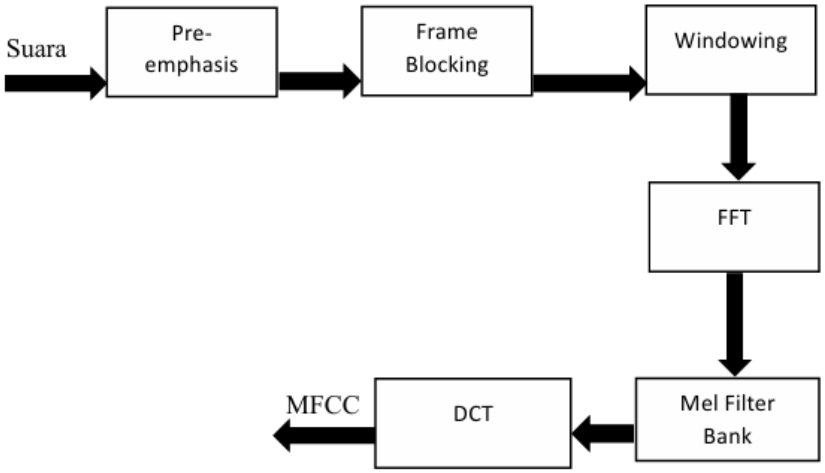


Fig 1. MFCC feature extraction stages

The following is an explanation of the stages of the MFFC feature extraction process:

1. Pre-emphasis
   Pre-emphasis is the first stage in MFCC feature extraction. At this stage the original speech signal goes through a filter to emphasize the high frequencies, which helps to improve the quality of the extracted features. Pre-emphasis aims to reduce the noise ratio in the signal so as to improve the quality of a signal and so that the baseband level in the high frequency part still has good signal quality. Systematically, pre-emphasis can be expressed using the following equation :

   $$y(n) = s(n) - \alpha\, s(n-1) \quad (1)$$

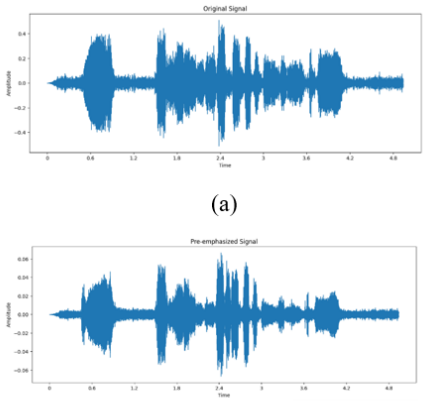
   Description:
   y(n): pre-emphasis result signal
   s(n): signal before pre-emphasis
   α: pre-emphasis filter constant (between 0.9-1.0)
   s: signal

   When plotting the data from this pre-emphasis stage, it will produce a plot like the following :

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024



(a)

(b)

Fig 2. example of pre-emphasis plot (a) Before pre-emphasis (b) after pre-emphasis

2. Frame Blocking
Frame Blocking is the stage where the speech signal is segmented into frames. At this stage, the speech signal is divided into frames with a certain shorter time. In general, in the frame blocking process, each frame is 20-25 milliseconds in size with the size of the overlapping part (M) between one frame and another frame of 30-50% of the frame length. The framing of the sound signal can be seen in the following figure :

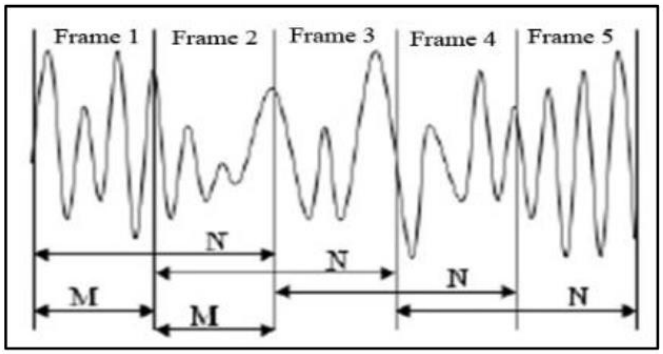


Fig 3. framing example

3. Windowing
Windowing is a weighting stage on each frame that has been formed in the previous process using a window function. windowing aims to minimize signal discontinuities at the beginning and end of each frame due to frame blocking. The equation used in the windowing process is as follows:

$$x(n) = x_i\,(n)w(n), n = 0,1,2, \dots, N \quad (2)$$

Description:
x(n): Signal value of windowing result
xi: Signal value of the i-th signal frame
w(n): Window function
N: Frame size

4. Fast Fourier Transform (FFT)

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

Fast Fourier Transform is the process of converting each frame of N samples from the time domain to the frequency domain. The FFT is an implementation of the Discrete Fourier Transform (DFT) operated on a discrete-time signal. The Fast Fourier Transform process is done by implementing the Discrete Fourier Transform (DFT) using the following equation:

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

Description:
S(k): The result of the kth Fast Fourier Transform calculation
x(n): The result of the nth windowing calculation
k: index of frequency (0,1,2,...,N)
N: Number of samples to be processed

When plotting the data from this FFT stage, it will produce a plot like the following:

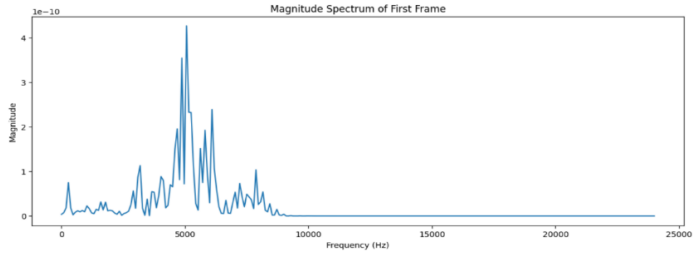


Fig 4. example plot of FFT result

5. Mel Filter Bank
A filterbank is a form of filtering that is done with the aim to determine the energy size of a particular frequency band in a sound signal. The power spectrum passes through a mel filter bank, which consists of a number of triangular filters distributed non-linearly along the mel frequency. The mel scale is a logarithmic frequency scale that is more in line with human auditory perception. The following equation is used to calculate the mel scale so as to obtain the lowest and highest values of the sound frequency in mel frequency and distribute them into N filter banks.

$$Mel(f) = 2595 \times Log_{10}(1 + \frac{f}{700}) \quad (4)$$

Description:
Mel(f): Mel Scale function
f: Frequency

6. Discrete Cosine Transform (DCT)
Discrete Cosine Transform (DCT) is the last step of feature extraction with MFCC. The basic concept of DCT is to decorrelate the mel spectrum to produce a good representation of the local spectral property. At this stage, the Mel spectrum value in the frequency domain will be converted into the time domain in order to obtain the coefficient value. DCT is applied to the log-energy to compress the information into a number of cepstral coefficients (Dwi & Candra, n.d.). Usually, the first 12-13 coefficients are used for the final MFCC representation. The DCT process can be expressed in the following equation :

$$C(k) = 2 \sum_{n=0}^{N-1} x(n) cos \frac{\pi(2n+1)k}{2N} \quad (5)$$

Description:
x(n): Output of the filterbank process

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

N: Expected number of coefficients

Data mining processing

The data mining processing carried out in this study, namely, following the stages in Knowladge Discovery in Database (KDD), to produce information in accordance with a predetermined order, the following stages:

1. Data Selection
   Before information retrieval in KDD, data must be selected from operational datasets stored in files different from the operational database.
2. Preprocessing
   Before data mining, data related to KDD needs to be cleaned, which consists of deduplication, checking inconsistent data, and evaluating data errors. At this stage, the data integration process will be carried out to merge data from different databases, then data cleaning is carried out to produce a clean dataset so that it can be used in the next stage. The following is an explanation of both processes:

   a. Data Integration
      This stage is the process of merging data from different databases, so that the data integrates with each other. Data integration is done on attributes that identify unique entities. At this stage there is no data merging because the data taken comes from one database.
   b. Data Cleaning
      This stage is the initial stage of the KDD process. At this stage irrelevant, missing value, and redundant data must be cleaned. This is because relevant data, not missing values, and not redundant are the initial requirements in doing data mining. A data is said to be missing value if there are attributes in the dataset that do not contain values or are empty, while data is said to be redundant if in one dataset more than one record contains the same value.
3. Transformation
   Encoding is the stage of data transformation that is chosen to be suitable for the data mining stage. The coding stage in KDD is a creative process and depends on the type or pattern of information in the database. The Transformation stage is the stage of changing the data that has been selected, so that the data is suitable for the data mining process. The transformation process in KDD is a creative process and is highly dependent on the type or pattern of information to be sought in the database. At this stage, from all operational data, the attribute data of the attribute grouping used for the data mining transformation process is obtained, namely the MFCC and Dialect Feature attributes as data criteria that are targeted in the data mining process.
4. Data Mining
   Data mining is the stage that finds interesting patterns in selected data using certain techniques. Methods in data mining vary. Choosing the right method has to do with the overall goal and process of KDD. This stage is the process of finding interesting patterns or information in the selected data using certain techniques or methods based on the overall KDD process. The method used in this research is the K-Nearest Neighbor (KNN) method where this method has an attribute initialized as k, which is the number of neighbors used as a reference in KNN, the value of k is a positive integer, small and odd number.
5. Evaluation
   The pattern of information by the data mining process must be shown in a form that is easily understood by those who have an interest. This phase of KDD is called interpretation. This phase involves checking whether the cycle or information decided contradicts existing facts or assumptions (Handayani, n.d.).

Process Stages of the K-Nearest Neighbor (KNN) Method

The stages of working on the KNN method in this study are as follows:

1. Determination of the k value. Determining the value of k used in classification does not have a standard rule, but in this study the value of k used is 3
2. Calculation of the distance between training data and test data. The distance calculation technique used in this KNN method is the Euclidean Distance. Euclidean Distance is the matrix most often used to calculate the similarity of two vectors. The Euclidean Distance formula is the root of square differences between two

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

vectors. Two feature vectors can be compared to each other by calculating the distance between them, or alternatively, determining their degree of similarity. There are many distance measurements used in visual pattern classification. With the following equation:

$$\sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (6)$$

Description:
x_i: Value in the training data
y_i: Value in the testing data
n: Number of dimensions or features in the dataset

3. The distances that have been obtained are then sorted from the closest to the farthest ascending
4. Determine the test data group based on the majority label of the k nearest neighbors (Dewi et al., 2022).

4. RESULT

After passing the MFCC feature extraction process and data mining processing, the following data will be obtained:

Table 1
The results of mfcc feature extraction after passing the data mining processing process

Voice recording	MFCC Features 1	MFCC Features 2	MFCC Features 3	…….	MFCC Features 13	Dialect
1	-30.592.072	19.048.987	-11.232.471	…….	17.895.751	West
2	-3.225.896	18.630.025	-1.644.849	…….	0.190.044	West
3	-3.247.438	19.591.693	-5.007.981	…….	1.577.572	West
4	-33.368.457	17.907.031	13.588.304	…….	-20.210.373	West
5	-3.354.729	19.246.504	-7.352.086	…….	39.016.023	West
6	-31.121.466	18.178.156	30.424.871	…….	0.487.063	West
7	-33.910.632	18.296.077	-9.557.931	…….	2.497.453	West
8	-3.064.007	18.064.787	8.289.159	…….	0.419.405	West
9	-32.516.586	18.623.717	0.095.455	…….	17.717.685	West
10	-26.932.578	18.869.408	2.234.495	…….	-0.561.695	West
…….	…….	…….	…….	…….	…….	…….
100	-38.785.645	14.822.371	7.441.543	…….	-19.110.253	South

Table 1 is a table of results from MFCC feature extraction that has passed the data mining processing, where there are 100 voice recording data of Sundanese dialects, 13 MFCC features and 2 dialect labels "west" and "south".

Test results of the K-Nearest Neighbor method on RapidMiner

The analysis is carried out descriptively in order to obtain an overview of the classification of speech recognition of Sundanese dialects. Based on the results of the research that has been obtained, accompanied by existing data, the authors will then analyze the results of the research that has been presented by calculating the percentage of the analysis data.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

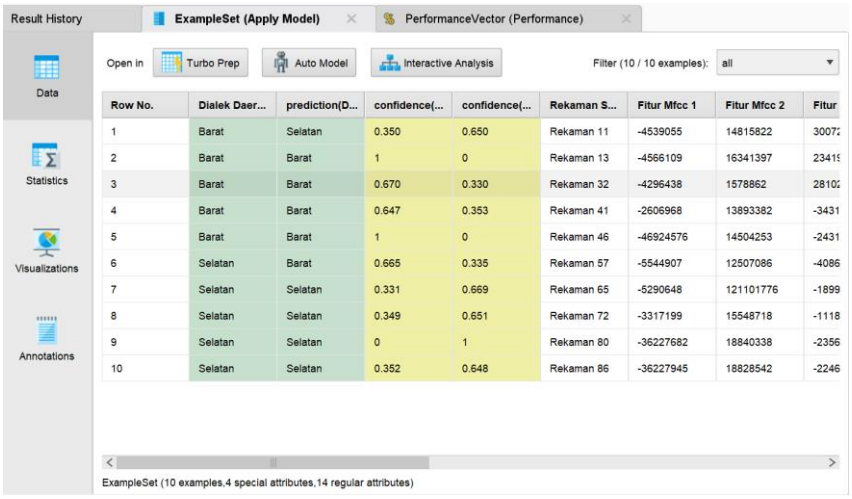


Fig 5. Apply model results

In Figure 5, testing data get prediction results from 10 testing data records that are read, resulting in a prediction decision "west" 5 data and decision " south" 5 data.

A Performance vector is a collection of performance metrics used to evaluate and measure how well a predictive model is working. Performance vectors typically include a variety of Statistics and values that provide a complete picture of a model's performance in classification, regression, or other predictive tasks.

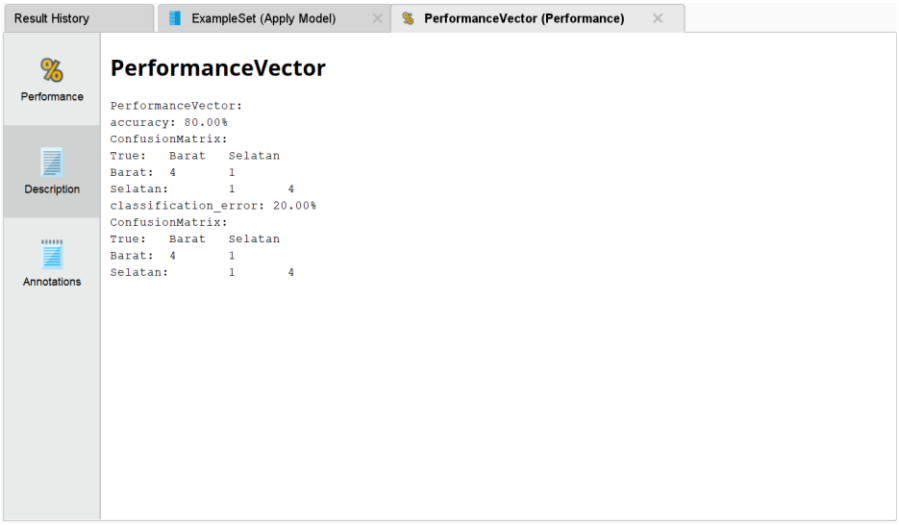


Fig 6. Performance vector results

Confusion Matrix is a tool used as a classification model evaluation to estimate the correct or wrong object. A matrix of predictions that will be compared with the actual class or in other words contain the actual value information and predictions on the classification.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

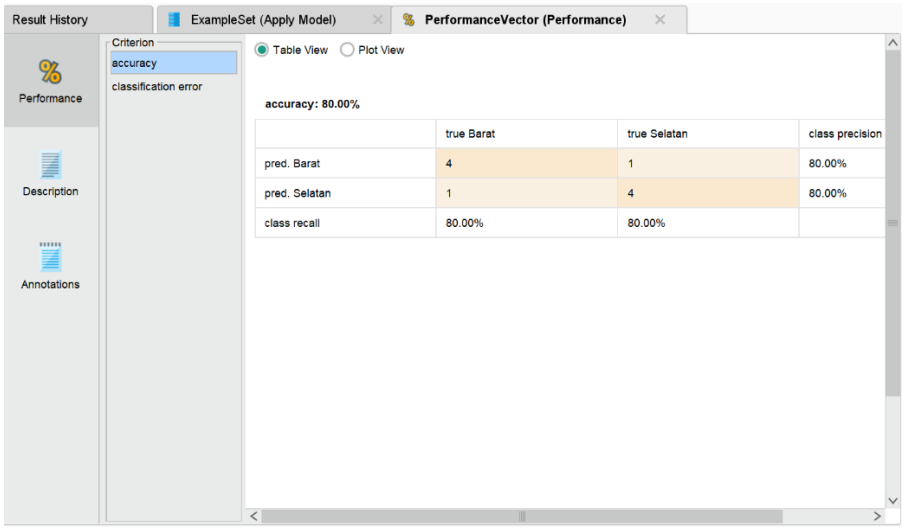


Fig 7. Performance level accuracy

Accuracy is a measure of how well a classification model predicts the correct classes of all predictions made. Based on Figure 7 , an accuracy of 80.00% was obtained, which means that of all the predictions made by the model, 80.00% of them are correct according to the actual class. High accuracy indicates that the classification model has a good ability to distinguish between "Western" and "southern" dialects based on the given features.

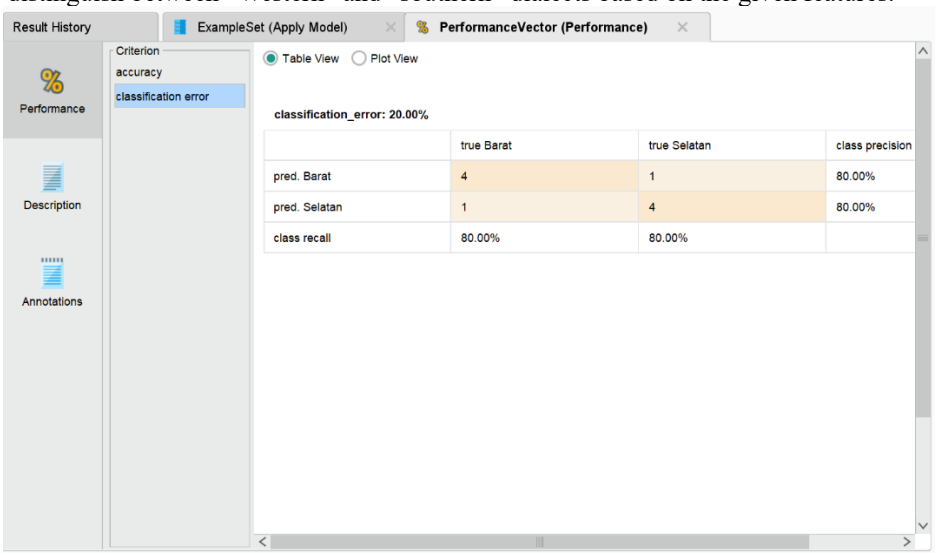


Fig 8. Performance level error

The error rate is the inverse measure of accuracy. It measures how often the model makes wrong predictions. Based on Figure 8, obtained an error rate of 20.00% means that of all the predictions made by the model, about 20.00% of them are wrong. The low error rate indicates that the model has a good ability to avoid making wrong predictions.

5. DISCUSSIONS

Calculation of feature extraction MFCC

1. Pre-emphasis
   Suppose that in this case, the pre-emphasis filter constant used is 0.97, then the calculation in the pre-emphasis stage using the formula above is as follows:

   y[0] = 5.5348501

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

y[1] = y[1]-0,97(y[0])
y[1] = (-2.7390197)-0,97(5.5348501)
y[1] = -8.1078242

2. Windowing
Suppose in the study, the window function used is hamming window then the nth window function of the signal data with hamming window can be calculated using the calculation formula of the nth signal data window function is as follows:

$$w(0) = 0{,}54 - 0{,}46\left(\frac{2\pi(0)}{520}\right)$$

$$w(0) = 0{,}54 - 0{,}46\cos(0)$$

$$w(0) = 0{,}08$$

After obtaining the value of the nth window function from the signal data, then the windowing formula is used to calculate the windowing result x(n). The result of windowing x(n) can be calculated by multiplying the signal value of the nth signal frame(y(n)) by its window function w (n). The windowing calculation process is as follows:

$$x(0) = y(0) \times w(0)$$

$$= (7.13790068) \times 0{,}08$$

$$x(0) = 5{,}71032054$$

3. Fast Fourier Transform (FFT)
Suppose in the study, the number of values calculated at the FFT stage is 512 (N=512), then the calculation process in the Fast Fourier Transform process is as follows:

$$s(0) = \sum_{n=0}^{512-1} x(n) e^{\frac{-j2\pi n(0)}{512}}$$

$$s(0) = (5{,}71032054)e^{\frac{-j2\pi(0)(0)}{512}} + (8.19981570)e^{\frac{-j2\pi(1)(0)}{512}}$$

$$+ (6.11131231)e^{\frac{-j2\pi(2)(0)}{512}} + (1.22343630)e^{\frac{-j2\pi(3)(0)}{512}}$$

$$+ (7.83836882)e^{\frac{-j2\pi(4)(0)}{512}} + (7.12290406)e^{\frac{-j2\pi(511)(0)}{512}}$$

$$s(0) = 7.08910173$$

4. Mel Filter Bank
For example, in the study, the results of the FFT stage in the form of energy density values, these values will be filtered using the mel scale with the help of the triangular filter bank. This is done to determine the energy available at each point. The above formula is used to calculate the lowest and highest frequency values in the mel frequency value and divide them into N filter banks. The calculation of the lowest and highest frequency values is as follows:

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

$$\text{Mel}(f(0)) = 2595 \log 10\ (1 + \frac{0}{700}) = 0$$

$$\text{Mel}(f(5000)) = 2595 \log 10\left(1 + \frac{5000}{700}\right) = 236.4045$$

5. Discrete Cosine Transform (DCT)
Suppose that at this stage the expected coefficient is 40 or N=40, then to calculate the value of the coefficient, the above formula is used with the following calculation:

$$c(0) = 2 \sum_{n=0}^{40-1} x_n cos \frac{\pi(2n+1)(0)}{2N}$$

$$=2[236.40458953\cos(0)+(-234.26287304\cos(0))+ (-208.0816647\cos(0))$$

$$+...+(164.66021165)\cos(0))]$$

$$= -41.2797366$$

$$c(0) = (-41.2797366)\ x \frac{1}{\sqrt{4(40)}} = -3.24568825$$

Calculation method K-Nearest Neighbor

Before performing the process stages of KNN method, the dataset results from data mining processing will be divided into training data and testing data where the division is 90:10, 90 training data and 10 testing data the process is called Split Data, The Division of training and testing data is done randomly.

Table 2
Training Data

Voice recording	MFCC Features 1	MFCC Features 2	MFCC Features 3	…….	MFCC Features 13	Dialect
1	-30.592.072	19.048.987	-11.232.471	…….	17.895.751	West
2	-3.225.896	18.630.025	-1.644.849	…….	0.190.044	West
3	-3.247.438	19.591.693	-5.007.981	…….	1.577.572	West
4	-33.368.457	17.907.031	13.588.304	…….	-20.210.373	West
5	-3.354.729	19.246.504	-7.352.086	…….	39.016.023	West
6	-31.121.466	18.178.156	30.424.871	…….	0.487.063	West
7	-33.910.632	18.296.077	-9.557.931	…….	2.497.453	West
8	-3.064.007	18.064.787	8.289.159	…….	0.419.405	West
9	-32.516.586	18.623.717	0.095.455	…….	17.717.685	West
10	-26.932.578	18.869.408	2.234.495	…….	-0.561.695	West
…….	…….	…….	…….	…….	…….	…….
90	-38.785.645	14.822.371	7.441.543	…….	-19.110.253	South

Training Data is a collection of data used to train a machine learning algorithm model. This Data helps the model learn and understand the relationship between expected input and output features. Training Data is data that is used as a reference to build a classification model (Darwis et al., n.d.).

Table 3
Testing Data

Voice recording	MFCC Features 1	MFCC Features 2	MFCC Features 3	…….	MFCC Features 13	Dialect
1	-4.539.055	14.815.822	30.072.432	…….	-8.373.011	West
2	-4.566.109	16.341.397	23.419.678	…….	-0.367.572	West

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

3	-4.296.438	1.578.862	28.102.598	…….	-6.731.068	West
4	-2.606.968	13.893.382	-34.318.665	…….	-12.805.849	West
5	-46.924.576	14.504.253	-243.111	…….	0.677.755	West
6	-5.544.907	12.507.086	-4.086.188	…….	-12.124.104	South
7	-5.290.648	121.101.776	-18.999.607	…….	-105.367.565	South
8	-3.317.199	15.548.718	-1.118.572	…….	-63.367.877	South
9	-36.227.682	18.840.338	-23.561.771	…….	-79.264.984	South
10	-36.227.945	18.828.542	-2.246.674	…….	-71.672.177	South

Testing Data is a set of data that also has a label/class that is used to test the accuracy of the pattern/model in classifying testing data (Musu & Ibrahim, n.d.). After the division of data or Split data is complete then the next process will be done calculation method KNN:

1. Determination of the k value. Determining the value of k used in classification does not have a standard rule, but in this study the value of k used is 3
2. Calculation of the distance between training data and test data. To calculate the distance between training data with testing data, suppose 6 training data and 1 testing data are taken for calculation using the Euclidean Distance formula.

Table 4
Calculation of euclidean distance

Training Data	Calculation of Euclidean Distance (Testing Data)
D1	$\sqrt{((-30-(-4)^2+(19-14)^2+(-11-30)^2+(34-825)^2+(28-10)^2+(-11-(-8))^2+(18-(-10))^2+(8-(-43)^2+(-11-(-10)^2+(-0-(-10)^2+(-64-(-496)^2+(-33-(-6)^2+(17-(-8)^2)} = \sqrt{944,952} = \mathbf{972}$
D2	$\sqrt{((-3-(-4)^2+(18-14)^2+(-1-30)^2+(4-825)^2+(32-10)^2+(-11-(-8))^2+(21-(-10))^2+(8-(-43)^2+(-1-(-10)^2+(-19-(-10)^2+(-6-(-496)^2+(-27-(-6)^2+(0-(-8)^2)} = \sqrt{919,473} = \mathbf{958}$
D3	$\sqrt{((-3-(-4)^2+(19-14)^2+(-5-30)^2+(43-825)^2+(3-10)^2+(-10-(-8))^2+(19-(-10))^2+(7-(-43)^2+(-13-(-10)^2+(-12-(-10)^2+(-70-(-496)^2+(-44-(-6)^2+(1-(-8)^2)} = \sqrt{938,491} = \mathbf{968}$
D51	$\sqrt{((-5-(-4)^2+(1-14)^2+(1-30)^2+(62-825)^2+(10-10)^2+(-9-(-8))^2+(126-(-10))^2+(-1-(-43)^2+(-0-(-10)^2+(-6-(-10)^2+(-18-(-496)^2+(4-(-6)^2+(-740-(-8)^2)} = \sqrt{1,423,001} = \mathbf{1,192}$
D52	$\sqrt{((-31-(-4)^2+(15-14)^2+(-7-30)^2+(4-825)^2+(23-10)^2+(-7-(-8))^2+(-675-(-10))^2+(-14-(-43)^2+(69-(-10)^2+(-12-(-10)^2+(-18-(-496)^2+(10-(-6)^2+(-79-(-8)^2)} = \sqrt{1,425,250} = \mathbf{1,193}$
D53	$\sqrt{((-5-(-4)^2+(122-14)^2+(-9-30)^2+(54-825)^2+(10-10)^2+(-9-(-8))^2+(9-(-10))^2+(-15-(-43)^2+(-18-(-10)^2+(-4-(-10)^2+(-13-(-496)^2+(60-(-6)^2+(-8-(-8)^2)} = \sqrt{874,594} = \mathbf{935}$

3. The distances that have been obtained are then sorted from the closest to the farthest ascending
   The distance that has been obtained is then sorted from the closest (small) distance to the farthest (large) or called the ascending order

Table 5
Ascending order

Rank	Calculation of Euclidean Distance Result
1	D53 = 935
2	D2 = 958
3	D3 = 968
4	D1 = 972
5	D51 = 1,192
6	D52 = 1,193

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

4. Determine the test data group based on the majority label of the k nearest neighbors
   Classification results by finding the majority label of k nearest neighbors

Table 6
Result of classification of majority labels from k nearest neighbors

Rank	Calculation of Euclidean Distance Result	Dialect	K=3
1	D53 = 935	South	1
2	D2 = 958	West	2
3	D3 = 968	West	3
4	D1 = 972	West	
5	D51 = 1,192	South	
6	D52 = 1,193	South	

Based on Table 6 is the result of the prediction of the calculation of Testing Data with the value of k=3 obtained 1 Southern dialect and 2 Western dialects, where the results of the prediction show the largest number of neighbors in accordance with the label testing data is western dialect. The suggestions for further research development are that it can be developed using other regional language dialects and can be developed by applying different feature extraction and classification methods.

6. CONCLUSION

Based on the results of research that shows the classification accuracy of 80.00% and the error rate of 20.00%, it can be concluded that the k-Nearest Neighbors (KNN) method can be done in classifying speech recognition Sundanese dialect with a fairly high accuracy. This shows that KNN can well distinguish between the "Western" and "southern" dialects of Sundanese. Mel-Frequency Cepstral Coefficients (MFCC) proved to be very effective in extracting sound features from Sundanese dialect speech datasets for classification purposes. The effectiveness of MFCC in extracting sound characteristics is reflected in the high classification accuracy achieved. With an accuracy of 80.00%, MFCC managed to capture the unique characteristics of each dialect that allows KNN to perform classification with a relatively low error rate. Overall, the combination of KNN and feature extraction methods using MFCC can be well applied to the task of speech recognition classification of Sundanese dialects, providing satisfactory results with fairly high accuracy and acceptable error rates.

7. REFERENCES

Adnan, F., Amelia, I., Sayyid ', D., & Shiddiq, U. (2022). Implementasi Voice Recognition Berbasis Machine Learning. *Edu Elektrika Journal*, *11*(1).

Ajinurseto, G., Bakrim, L. O., & Islamuddin, N. (2023). Penerapan Metode Mel Frequency Cepstral Coefficients pada Sistem Pengenalan Suara Berbasis Desktop. *Infomatek*, *25*(1), 11–20. https://doi.org/10.23969/infomatek.v25i1.6109

Darwis, D., Siskawati, N., & Abidin, Z. (2021). *Penerapan Algoritma Naive Bayes untuk Analisis Sentimen Review Data Twitter BMKG Nasional*. *15*(1).

Deski Prasetyo, P., Gede Pasek Suta Wijaya, I., & Yudo Husodo, A. (n.d.). *KLASIFIKASI GENRE MUSIK MENGGUNAKAN METODE MEL FREQUENCY CEPSTRUM COEFFICIENTS (MFCC) DAN K-NEAREST NEIGHBORS CLASSIFIER (Classification of Music Genres Using The Mel-Frequency Cepstrum Coefficients (MFCC) and K-Nearest Neighbors Classifier Methods)*. http://jtika.if.unram.ac.id/index.php/JTIKA/

Dewi, S. P., Nurwati, N., & Rahayu, E. (2022). Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor. *Building of Informatics, Technology and Science (BITS)*, *3*(4), 639–648. https://doi.org/10.47065/bits.v3i4.1408

Di, S., Bandung, K., Al, A., Ramadhani, F., Melga, B., & Nastiti, N. E. (2021). *Perancangan Media Pembelajaran Interaktif Berbahasa Sunda Untuk Anak Pra*.

Dwi, S., & Candra, P. (2021). *KLASIFIKASI SUARA DENGAN EKSTRAKSI CIRI MEL FREQUENCY CEPSTRAL COEFFICIENTS MENGGUNAKAN MACHINE LEARNING*.

Fadlila Surenggana, F., Aranta, A., & Bimantoro, F. (2022). *KLASIFIKASI MOOD MUSIK MENGGUNAKAN K-NEAREST NEIGHBOR DENGAN MEL FREQUENCY CEPSTRAL COEFFICIENTS (Mood Music Classification using K-Nearest Neighbor with Mel Frequency Cepstral Coefficients)*. http://jtika.if.unram.ac.id/index.php/JTIKA/

Guntara, R. G., Nuryadin, A., & Hartanto, B. (2021). *Pemanfaatan Google Speech to Text Untuk Aplikasi*

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 6, Number 3, July 2024
https://doi.org/10.47709/cnapc.v6i3.4226

Submitted : 2 July 2024
Accepted : 8 July 2024
Published : 22 July 2024

Pembelajaran Kamus Bahasa Sunda Pada Platform Mobile Android. *4*(1), 10–19. https://doi.org/10.31764/justek.vXiY.ZZZ

Handayani, F. (n.d.). Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokkan Mahasiswa Berdasarkan Gaya Belajar. *Jurnal Teknologi Dan Informasi*. https://doi.org/10.34010/jati.v12i1

Indra Kusuma, A., Sularsa, A., & Zani, T. (n.d.). *PEMBUATAN ASSET GAME EDUKASI BAHASA SUNDA "SI ASEP NYASAB DI LABIRIN" BERBASIS ANDROID*.

Komalasari, N., Hidayat, E. W., & Aldya, A. P. (2022). *APLIKASI PENGENALAN BAHASA SUNDA BERBASIS MULTIMEDIA* (Vol. 9, Issue 1).

Musu, W., & Ibrahim, A. (n.d.). *Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5*.

Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). ANALISIS DATA MINING UNTUK CLUSTERING KASUS COVID-19 DI PROVINSI LAMPUNG DENGAN ALGORITMA K-MEANS. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(2), 100. http://jim.teknokrat.ac.id/index.php/JTSI

Putu, G., Widano, A., Agung, A., Ngurah, I., & Karyawati, E. (2022). Perintah Menggunakan Sinyal Suara dengan Mel-Frequency Cepstral Coefficients (MFCC) dan K-Nearest Neighbor (KNN). In *JNATIA* (Vol. 1, Issue 1).

Rizky, S. A., Yesputra, R., & Santoso, S. (2021). PREDIKSI KELANCARAN PEMBAYARAN CICILAN CALON DEBITUR DENGAN METODE K-NEAREST NEIGHBOR. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *7*(2), 195–202. https://doi.org/10.33330/jurteksi.v7i2.1078

Setio, P. B. N., Saputro, D. R. S., & Winarno, B. (2020). *PRISMA, Prosiding Seminar Nasional Matematika Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C4.5*. *3*, 64–71. https://journal.unnes.ac.id/sju/index.php/prisma/

Suparman, S. (2023). POSISI KEMUNCULAN VOKAL KONSONAN DALAM BAHASA RAMPI DAN BAHASA TAE'. *Bahtera Indonesia; Jurnal Penelitian Bahasa Dan Sastra Indonesia*, *8*(2), 490–497. https://doi.org/10.31943/bi.v8i2.445

Yehezkiel, S. Y., & Suyanto, Y. (2022). Music Genre Identification Using SVM and MFCC Feature Extraction. *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, *12*(2), 115. https://doi.org/10.22146/ijeis.70898

* Corresponding author