Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 4 (2023) 832 - 836 ISSN Media Electronic: 2580-0760

Multi-Accent Speaker Detection Using Normalize Feature MFCC Neural Network Method

Kristiawan Nugroho[1], Edy Winarno[2], Eri Zuliarso[3], Sunardi[4]
[1,2,3]Master of Information Technology, Faculty of Information Technology and Industry, Universitas Stikubank
[4]Graphic Multimedia Engineering Technology Departement, Vocational Faculty, Universitas Stikubank
[1]kristiawan@edu.unisbank.ac.id, [2]edywin@edu.unisbank.ac.id, [3]eri299@edu.unisbank.ac.id, [4]sunardi@edu.unisbank.ac.id

Abstract

Speaker recognition is a field of research that continues to this day. Various methods have been developed to detect the human voice with greater precision and accuracy. Research on human speech recognition that is quite challenging is accent recognition. Detecting various types of human accents with different accents and ethnicities with high accuracy is a research that is quite difficult to do. According to the results of the research on the data preprocessing stage, feature extraction and the selection of the right classification method play a very important role in determining the accuracy results. This study uses a preprocessing approach with normalizing features combined with MFCC as a method for performing feature extraction and Neural Network (NN) which is a classification method that works based on the workings of the human brain. Research results obtained using the normalize feature with MFCC and Neural Network for multi-accent speaker recognition, the accuracy performance reaches 82.68%, precision is 83% and recall is 82.88%.

Keywords: speaker recognition; classification; multi accent; MFCC; neural network.

1. Introduction

Artificial intelligence technology in the field of speech recognition is growing rapidly nowadays. Various companies have produced smart tools for voice recognition such as Alexa, Siri and Google Assistant. These various products have become part of people's lives like personal assistants who help humans in every activity of their lives such as translating languages, playing entertainment in the form of music to recommending paths that are free from traffic jams as well as recommendations for places to eat, travel and the nearest gas station.

The development of speech recognition technology began in 1940 by a company by the name of the American Telephone and Telegraph Company (AT & T) by building a tool to recognize human speech. The research has progressed to date with the discovery of various methods and the results of the real contribution of speech recognition technology in the health, education, automotive and military fields. Voice recognition technology continues to help various areas of human life so that they can carry out various life activities well. This technology is proven to have helped humans in various fields of work, making it easier for humans to complete all their work faster.

One of the most intriguing research topics is speech recognition, a field of study that is constantly expanding. Various methods that have been frequently used to achieve maximum accuracy include Support Vector Machine (SVM), Random Forest (RF), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Neural Network (NN). SVM is a machine learning method that is also used in voice classification, such as the research conducted by Zade[1] which combines SVM with the feature extraction method of MFCC (Mel-Frequency Cepstral Coefficient) and LPC (Linear Predictive Coding) in the Azerbaijani DataSet classification. In another study, Pradana[2] also used SVM which was also combined with MFCC to detect Arabic speech resulting in an accuracy rate of 61.16%.

Another study on speech recognition was also conducted using the Random Forest method in research conducted by Rao[3] using the VoxCeleb dataset, which achieved a voice recognition accuracy rate of 84.53%. Nivetha[4] also uses Random Forest and MFCC LPC to detect sounds in Tamil. Another method that is often used in speech recognition is GMM such as research conducted by Chauhan[5], Nayana[6], and Rajan[7] which has achieved a fairly good level of accuracy. HMM is also a method used in speech

Accepted: 07-11-2022 | Received in revised: 12-06-2023 | Published: 12-08-2023

recognition such as research by Chamidy[8], Nada[9] and Chen[10]. The ability to distinguish between speakers with various accents is one of the most difficult subjects in speaker recognition. Recognizing speakers with different accents is a complex task that is not easy to do[11], so an appropriate method is needed to recognize the speaker's accent so as to produce a high level of accuracy. Various studies have been carried out to achieve a better level of speech recognition accuracy, including using the MFCC (Mel-Frequency Cepstral Coefficient) method such as the research conducted by Maurya[12] with the MFCC and GMM methods, Widyowati[13] which combines MFCC with Convolutional Neural Network (CNN) and Nugroho[14] which use MFCC and Deep Neural Network (DNN). In these various studies, MFCC has been proven to help improve speech recognition accuracy optimally.

Speaker Accent Recognition (SAR) is part of the topic of speech recognition. SAR is a technology used to identify or recognize someone's voice accent or dialect based on a recording of their voice. Research on speaker accent recognition has been carried out by several researchers, including Odulio[15] who used the Convolutional Neural Network (CNN) to detect Bikol and Tagalog accents in the Philippines. This research achieved an accuracy rate of between 78.33% to 79.28%. In another study Ahmet[16] used MFCC and several other machine learning algorithms such as K-Nearest Neighbor (KNN) to recognize 329 speakers with 6 accents which achieved an accuracy of 80.49%. Duduka[17] has also done research on speaker accent recognition used MFCC and CNN to detect speakers with Arabic, English and Mandarin accents which achieved an accuracy rate of 62.81%. Zhang[18] also contributed to SAR research used 2 datasets, namely Librispeech and AESRC which consisted of 8 Chinese (CN), Indian (IN), Japanese (JP), Korean (KR), American (US), English (UK), Portuguese (PT) and Russian (RU) by using Hybrid Phonetic Features which achieve an accuracy rate of up to 99.6%.

Research on SAR in Indonesia was also carried out by several researchers, such as the research conducted by Sakti [19] who have done research with several accents including Sundanese, Javanese, Acehnese, Batak, Balinese and Bugis which achieved model accuracy of 73.33%. In another study, Idwal[20] using Linear Predictive Coding (LPC) and Vector Quantization (VQ) to recognize Malay and Sundanese accents which produce an accuracy of 70%. These research have helped the detection of multi-accent speakers with various levels of performance accuracy but can still be improved by choosing the right algorithm.

Feature extraction is the process of converting raw data into a numerical representation that can be used to analyze data. In speech recognition, feature extraction involves converting speech signals into a series of numerical features that represent important information in the signal. In the field of speaker recognition research the MFCC (Mel Frequency Cepstral Coefficients) method is the method most often used. MFCC is a method that is often used because it has a faster extraction time and a higher level of accuracy than other methods. As for the classification technique, the Neural Network (NN) method is also a method that is often used by researchers because it has the advantage of parallel processing which allows several tasks to be carried out simultaneously.

This research used MFCC which has previously been preprocessed using the Normalize Feature combined with the Neural Network (NN) method so that it can achieve a high level of accuracy for recognizing different accent voices. NN is used because it has various advantages, including advantages in predicting nonlinear cases, having good performance in parallel processing and the ability to tolerate errors[21] so that this approach is suitable for use in this accent speech recognition research.

2. Research Methods

Research on the detection of speakers with multiple accents is a challenging research topic. Research on the sspeaker accents recognition who use various accents is carried out in stages as shown in Figure 1.

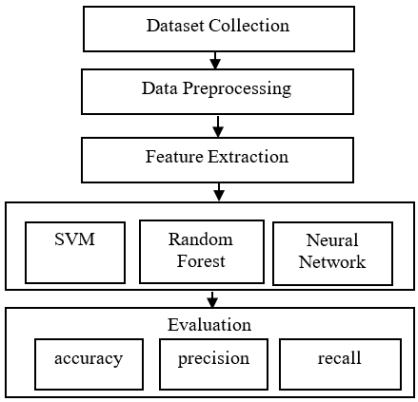

Figure 1. Research Stages

The research stages in figure 1 are the stages in detecting speakers who have various kinds of accents.

2.1 Dataset Collection

The dataset is the main data source used in a study. This research on the detection of speakers who use many accents takes a dataset from the UCI Dataset with the following website address https://archive.ics.uci.edu/ml/datasets/Speaker+Accent+Recognition# This dataset consists of 12 attributes with 1 label named language which contains 6 different types of speaker accents, including those from France, England, America and Germany.

2.2 Data Preprocessing

This stage is a very important stage because it relates to the preparation of valid data so that it can be used

DOI: https://doi.org/10.29207/resti.v7i4.4652

properly in the next stage. In this study, a preprocessing approach is used in the form of deleting data that has empty values and the Normalize Feature which is an integral part of the Neural Network method[22] in the Orange application with a display as shown in figure 2.

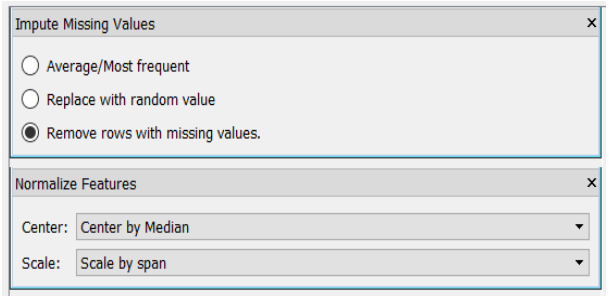

Figure 2. Preprocessing Stage

In the preprocessing process on figure 2, the multi-accent speaker data will be deleted which contains empty data and each feature will be normalized so that better data will be obtained and ready for feature extraction.

2.3 Feature Extraction

In this study, the MFCC method is used as one of the superior methods in extracting voice signal data. The MFCC method is often used because it has the advantage of being able to identify the characteristics of the voice signal properly[23] and is a method that is often used in speech recognition. In the MFCC method, the steps taken to extract the voice signal are:

Pre-emphasis: The initial stage of the MFCC method is carried out after the sound sampling stage, Pre-emphasis aims to obtain a smoother form of speech signal frequency spectral so that the quality of the voice signal will be better for processing in the next hold.

Frame Blocking: This stage is the process of dividing the voice signal in the form of shorter segments so that the time period changes.

Windowing: This function aims to manipulate the amplitude of the sound signal by using a mathematical formula.

Fast Fourier Transform (FFT): FFT is an approach to perform DFT/Discrete Fourier Transform calculations quickly.

Mel Frequency Wrapping (MFW): a step to determine the size of the frequency band in the voice signal that is needed before the next stage.

Discrete Continues Transform (DCT): The DCT approach is needed in sound processing in changing from the frequency domain to the time domain[24] and performing spectrum compression.

Cepstral Lifting (CL): In the MFCC method, CL is the last technique used to improve the quality of speech recognition signals.

2.4 Classification

In this study, the Neural Network (NN) method is used, which is a method that works like a neural network in the human brain. The architectural of the NN method[25] can be seen in figure 3.

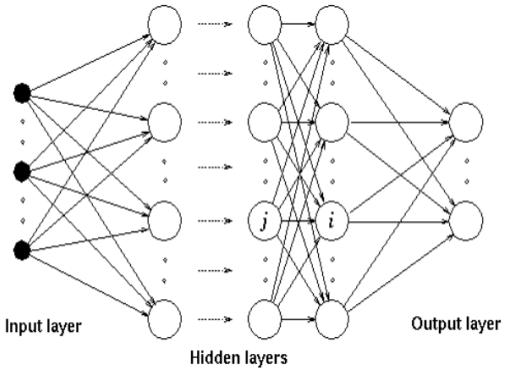

Figure 3. Neural Network architecture

In the NN architecture in Figure 3, it can be seen that there are 3 main parts of this method, namely the input layer, hidden layer and output layer which consists of several circular neurons. The input layer is the part to perform input in the form of features that will be processed in this algorithm, while the hidden layer consists of several network layers that are used to process data from the input layer and the results will be displayed in the output layer. The calculation of the value at the output layer (y) on the NN can be calculated using Equation 1.

$$y = \sum (xiwi) + b \quad (1)$$

x is the input value, w is the weight and b shows the bias.

This study also compares the results of research using NN with 2 other Machine Learning methods, namely SVM and Random Forest. These two methods are also used in this multi-accent voice detection research because they also have several advantages, SVM has the advantage of producing a good classification model even though it is trained to use only a small amount of data[26] while Random Forest is a robust method to overcome the problem of overfitting and data that is lacking. non-linear[27].

E. Evaluation

The evaluation stage is an important process in determining the performance of an algorithm. In research that uses Machine Learning, Measuring the level of performance of each method's accuracy, precision, and recall is a common evaluation technique. The calculation to determine the performance level of each evaluation is formulated in Equation 2 until 4.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100\% \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{FP+TP}} * 100\% \quad (3)$$

DOI: https://doi.org/10.29207/resti.v7i4.4652

$$\text{Recall} = \frac{\text{TP}}{\text{FN+TP}} * 100\% \quad (4)$$

Where TP is the abbrevation from True Positive, TN is the abbrevation from True Negative, FP is the abbrevation from False Positive, and FN is the abbrevation from False Negative

3. Results and Discussions

Based on the results of research that has been carried out on multi-accent speaker recognition, after the multi-accent speaker dataset is preprocessed using the normalize feature technique, the next step is the feature extraction stage with MFCC resulting in 330 lines of extracted speech data. Then the classification process is carried out using the Orange application which is one of the processing applications in Data Mining.

In this study, the Neural Network method was used with random sampling with a composition of 70% training set and 30% test set. To perform data training, 100 neurons in the hidden layer are used using the Relu (Rectified Linear Units) activation function, which is a simple non-linear activation function in a Neural Network.

The results of this multi-accent speech processing classification use a Neural Network (NN) which is then compared the results with 2 other methods, namely SVM and Random Forest. The results of measuring the performance of each model using the cross validation sampling technique with 10 folds were tested 5 times. resulting in a performance comparison table as shown in table 1,

Table 1. Model Performance Test

Methods	Acc(%)	Prec(%)	Rec(%)
Test-1			
SVM	79.6	82.2	79.6
Random Forest	72.3	72.3	72.3
NN	**82.7**	**82.7**	**82.7**
Test-2			
SVM	79.6	82.4	79.6
Random Forest	74.5	74.0	74.5
NN	**83.6**	**84.0**	**83.6**
Test-3			
SVM	79.6	82.4	79.6
Random Forest	73.9	74.1	79.3
NN	**82.4**	**82.3**	**82.4**
Test-4			
SVM	79.6	82.4	79.6
Random Forest	74.8	75.4	74.8
NN	**83.0**	**83.4**	**83.0**
Test-5			
SVM	79.6	82.4	79.6
Random Forest	74.8	74.8	74.8
NN	**82.7**	**82.8**	**82.7**

If we observe in table 1, the Neural Network (NN) method is a method that produces the best level of performance compared to the other 2 methods. Comparison of the results can be seen in table 2 of the average performance of the model.

Table 2. Avarage Result of Model Test

Methods	Acc(%)	Prec(%)	Rec(%)
SVM	79.60	82.30	79.60
Random Forest	74,06	74,12	75,14
NN	**82,68**	**83,04**	**82,88**

According to table 2, it can be concluded that the Neural Network method is the best method for its performance in recognizing multi-accent speakers compared to other methods. Neural Network achieves an accuracy rate of 82.68 %, precision 83.04% and recall of 82.88% outperforms other methods where the performance level of the method is less than 80%. A more detailed explanation of the performance of each method used in multi-accent speaker recognition can be described in a comparison chart so that it can be seen more clearly about the performance of each model when compared to one another. Graphics regarding the performance of these models can be seen in figure 4.

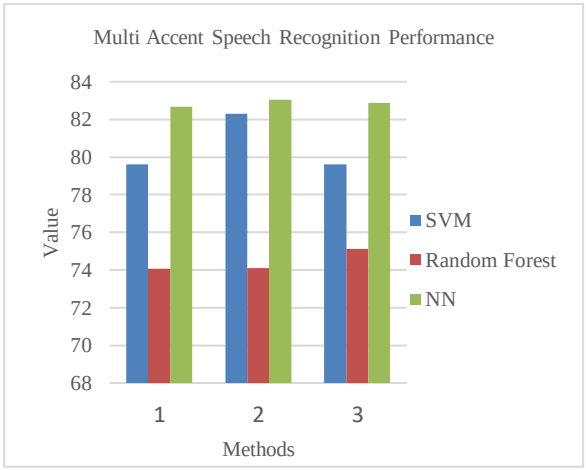

Figure 4. Model Performance Comparation

In figure 4, it can be seen that the Neural Network method dominates the level of excellence in terms of accuracy, precision and recall. A more complete comparison of the level of accuracy can be seen in figure 5, visualization and calculations on the Confusion Matrix.

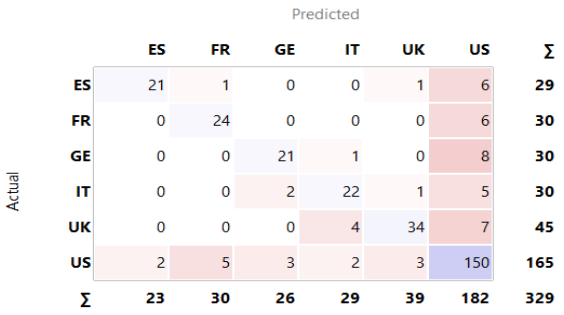

Figure 5. Confusion Matrix Result

Based on figure 5, the calculation of the accuracy of the Neural Network method on the confusion matrix is as Equation 5.

$$\text{Accuracy} = \frac{(21+24+21+22+34+150)}{329} * 100\% \quad (5)$$

$$\text{Accuracy} = 82.68\%$$

DOI: https://doi.org/10.29207/resti.v7i4.4652

The results of measuring the level of accuracy using NN which reached 82.68% showed that the NN method was superior to other methods in detecting multi-accented speech sounds.

4. Conclusion

Recognition of speakers with different accents is an interesting topic for research in speech recognition research. Various methods have been developed in recognizing speakers who use various accents. One method that is often used in speech recognition is the Neural Network (NN) which is a method that works based on the neural performance of the human brain.

In this research, a preprocessing approach using normalize feature and Neural Network is used to identify multi-accented speakers. The outcomes demonstrate that the Neural Network approach yields the greatest results. with an average accuracy of 82.68%, precision of 83.04% and recall of 82.88%, outperforming other methods such as Random Forest and Support Vector Machine.

References

[1] K. Aida-Zade, A. Xocayev, and S. Rustamov, "Speech recognition using Support Vector Machines," *Appl. Inf. Commun. Technol. AICT 2016 - Conf. Proc.*, vol. 1, 2017, doi: 10.1109/ICAICT.2016.7991664.

[2] W. A. Pradana, Adiwijaya, and U. N. Wisesty, "Implementation of support vector machine for classification of speech marked hijaiyah letters based on Mel frequency cepstrum coefficient feature extraction," *J. Phys. Conf. Ser.*, vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012050.

[3] M. S. Rao, G. B. Lakshmi, P. Gowri, and K. B. Chowdary, "Random Forest Based Automatic Speaker Recognition System," *Int. J. Anal. Exp. Model Anal.*, vol. 12, no. 4, pp. 526–535, 2020, [Online]. Available: http://www.ijaema.com/gallery/63-ijaema-april-3748.pdf

[4] N. S, D. R. A, and M. G. S, "Speech Recognition System for Isolated Tamil Words using Random Forest Algorithm," *Int. J. Recent Technol. Eng.*, vol. 9, no. 1, pp. 2431–2435, 2020, doi: 10.35940/ijrte.a1467.059120.

[5] V. Chauhan, S. Dwivedi, P. Karale, and P. S. M. Potdar, "Speech to Text Converter Using Gaussian Mixture Model (GMM) of Electronics and Telecommunication Engineering," *Int. Res. J. Eng. Technol.*, pp. 125–129, 2016.

[6] P. K. Nayana, D. Mathew, and A. Thomas, "Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods," *Procedia Comput. Sci.*, vol. 115, pp. 47–54, 2017, doi: 10.1016/j.procs.2017.09.075.

[7] R. R. K and A. P. Joseph, "Domestic Language Accent Detector Using MFCC and GMM," *Int. J. Appl. Eng. Res.*, vol. 15, no. 8, p. 800, 2020, doi: 10.37622/ijaer/15.8.2020.800-803.

[8] T. Chamidy, "Metode Mel Frequency Cepstral Coeffisients (MFCC) Pada klasifikasi Hidden Markov Model (HMM) Untuk Kata Arabic pada Penutur Indonesia," *Matics*, vol. 8, no. 1, p. 36, 2016, doi: 10.18860/mat.v8i1.3482.

[9] P. Huruf, Q. Nada, C. Ridhuandi, P. Santoso, and D. Apriyanto, "Speech Recognition dengan Hidden Markov Model untuk," *J. AL-AZHAR Indones. SERI SAINS DAN Teknol.*, vol. 5, no. 1, pp. 19–26, 2019.

[10] Y. Chen, "A hidden Markov optimization model for processing and recognition of English speech feature signals," *J. Intell. Syst.*, vol. 31, no. 1, pp. 716–725, 2022, doi: 10.1515/jisys-2022-0057.

[11] Y. Singh, A. Pillay, and E. Jembere, "Features of speech audio for accent recognition," *2020 Int. Conf. Artif. Intell. Big Data, Comput. Data Commun. Syst. icABCD 2020 - Proc.*, 2020, doi: 10.1109/icABCD49160.2020.9183893.

[12] A. Maurya, D. Kumar, and R. K. Agarwal, "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach," in *Procedia Computer Science*, 2018, vol. 125, pp. 880–887. doi: 10.1016/j.procs.2017.12.112.

[13] D. S. Widyowaty, A. Sunyoto, and H. Al Fatta, "Accent Recognition Using Mel-Frequency Cepstral Coefficients and Convolutional Neural Network," *Proc. Int. Conf. Innov. Sci. Technol. (ICIST 2020)*, vol. 208, no. Icist 2020, pp. 43–46, 2021, [Online]. Available: https://doi.org/10.2991/aer.k.211129.010

[14] K. Nugroho, E. Noersasongko, D. R. Ignatius, and M. Setiadi, "Enhanced Indonesian Ethnic Speaker Recognition using Data Augmentation Deep Neural Network," *J. King Saud Univ. - Comput. Inf. Sci.*, no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.04.002.

[15] B. Odulio *et al.*, "A speaker accent recognition system for filipino language," *Proc. 33rd Pacific Asia Conf. Lang. Inf. Comput. PACLIC 2019*, no. 2013, pp. 511–515, 2019.

[16] A. A. Ayrancõ, "Makine Ö ÷ renmesi AlgoritmalarÕ Kullanarak KonuúmacÕ AksanÕ TanÕma Speaker Accent Recognition Using Machine Learning Algorithms," 2020.

[17] P. M. C. Saiprasad Duduka, Henil Jain, Virik Jain, Harsh Prabhu, "A Neural Network Approach to Accent Classification," *Irjet*, vol. 8, no. 3, pp. 1775–1777, 2021.

[18] Z. Zhang, Y. Wang, and J. Yang, "Accent Recognition with Hybrid Phonetic Features," *Sensors (Basel).*, vol. 21, no. 18, 2021, doi: 10.3390/s21186258.

[19] S. Sakti, P. Hutagaol, A. A. Arman, and S. Nakamura, "Indonesian speech recognition for hearing and speaking impaired people," *8th Int. Conf. Spok. Lang. Process. ICSLP 2004*, no. February 2015, pp. 1037–1040, 2004, doi: 10.21437/interspeech.2004-366.

[20] A. Y. P. Idwal, Y. I. Nurhasanah, and D. B. Utami, "Sistem Pengenalan Suara Bahasa Indonesia Untuk Mengenali Aksen Daerah," *J. Tek. Inform. dan Sist. Inf.*, vol. 3, no. 3, pp. 461–471, 2017, doi: 10.28932/jutisi.v3i3.661.

[21] M. Badrul, "Optimasi Neural Network dengan Algoritma Genetika untuk Prediksi Hasil Pemilukada," *Bina Insa. ICT J.*, vol. 3, no. 1, pp. 229–242, 2016.

[22] B. Li, F. Wu, S. N. Lim, S. Belongie, and K. Q. Weinberger, "On feature normalization and data augmentation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 12378–12387, 2021, doi: 10.1109/CVPR46437.2021.01220.

[23] M. Susanti, B. Susilo, and D. Andreswari, "Aplikasi Speech-To-Text Dengan Metode Mel Frequency Cepstral Coefficient (Mfcc) Dan Hidden Markov Model (Hmm) Dalam Pencarian Kode," *J. Rekursif*, vol. 6, no. 1, pp. 48–58, 2018, [Online]. Available: https://ejournal.unib.ac.id/index.php/rekursif/article/view/6497%0Ahttps://ejournal.unib.ac.id/index.php/rekursif/article/download/6497/3102

[24] C. G. K. Leon, "Robust computer voice recognition using improved MFCC algorithm," *Proc. - 2009 Int. Conf. New Trends Inf. Serv. Sci. NISS 2009*, pp. 835–840, 2009, doi: 10.1109/NISS.2009.12.

[25] B. S. Santoso, J. P. Tanjung, U. P. Indonesia, B. Gandum, and A. N. Network, "Classification of Wheat Seeds Using Neural Network Backpropagation," *JITE (Journal Informatics Telecommun. Eng. Available*, vol. 4, no. January, pp. 188–197, 2021.

[26] M. Ichwan, I. A. Dewi, and Z. M. S, "Klasifikasi Support Vector Machine (SVM) Untuk Menentukan TingkatKemanisan Mangga Berdasarkan Fitur Warna," *MIND J.*, vol. 3, no. 2, pp. 16–23, 2019, doi: 10.26760/mindjournal.v3i2.16-23.

[27] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer ' s Disease : A Systematic Review," vol. 9, no. October, pp. 1–12, 2017, doi: 10.3389/fnagi.2017.00329.

DOI: https://doi.org/10.29207/resti.v7i4.4652