

Improving Indonesian multiethnic speaker recognition using pitch shifting data augmentation

Kristiawan Nugroho¹, Isworo Nugroho¹, De Rosal Ignatius Moses Setiadi², Omar Farooq³

¹Department of Information Technology and Industry, Universitas Stikubank, Semarang, Indonesia

²Department of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

³Department of Electronics Engineering, Z. H. College of Engg and Technology, A. M. U, Aligarh, India

Article Info

Article history:

Received Jan 25, 2023

Revised Mar 15, 2023

Accepted Mar 27, 2023

Keywords:

Data augmentation

Deep learning

Pitch shifting

Speaker recognition

ABSTRACT

Speaker recognition to recognize multiethnic speakers is an interesting research topic. Various studies involving many ethnicities require the right approach to achieve optimal model performance. The deep learning approach has been used in speaker recognition research involving many classes to achieve high accuracy results with promising results. However, multi-class and imbalanced datasets are still obstacles encountered in various studies using the deep learning method which cause overfitting and decreased accuracy. Data augmentation is an approach model used in overcoming the problem of small amounts of data and multiclass problems. This approach can improve the quality of research data according to the method applied. This study proposes a data augmentation method using pitch shifting with a deep neural network called pitch shifting data augmentation deep neural network (PSDA-DNN) to identify multiethnic Indonesian speakers. The results of the research that has been done prove that the PSDA-DNN approach is the best method in multi-ethnic speaker recognition where the accuracy reaches 99.27% and the precision, recall, F1 score is 97.60%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Kristiawan Nugroho

Department of Information Technology and Industry, Universitas Stikubank Semarang

Jl. Tri Lomba Juang, Semarang, Indonesia

Email: kristiawan@edu.unisbank.ac.id

1. INTRODUCTION

Speaker recognition is one of the challenging research fields. Various kinds of problems need to be solved to produce the discovery of new theories that can make a positive contribution to human life. Research results in the field of speaker recognition have proven to have led to various forms of new technology applied to voice authentication, surveillance speaker recognition, forensic speaker recognition, security, and multi speaker tracking. Tech giant companies such as Google, Microsoft, and Amazon have also taken advantage of this technology, including Google Voice technology, Apple Siri, and Amazon's Alexa.

The application of this technology is proven to be able to help human work, such as in the field of security and authentication in smart homes to help people with disabilities to recognize sound patterns and images. Initially, research in the field of speaker recognition used classical methods in machine learning, gaussian mixture model (GMM), such as studies carried out by Motlicek *et al.* [1] and Veena and Mathew [2]. In speaker recognition, the hidden markov model (HMM) strategy is also utilized, as demonstrated by Maghsoodi *et al.* [3], Hussein, *et al.* [4] and the support vector machine (SVM) approach on research that has been done by Chaunan *et al.* [5].

However, along with the growth of data that is getting bigger and the complexity of the problems faced today, researchers are starting to use the deep learning method in speaker recognition research. Deep

learning (DL) is an approach developed from the neural network algorithm. This method continues to be developed by researchers to solve various problems in machine learning. DL has advantages in terms of the ability to handle computational processes that involve very large data. In addition, DL is also able to process data representations of various forms such as text, images, and sound, which makes it a good ability to process information in the form of multi modals so that this method outperforms the previous machine learning method, especially used in the field of computer vision.

Various DL methods are used in voice signal processing research using the deep neural networks (DNN) approach as on research that has been done by Guo *et al.* [6] combined with I-Vector in research on short speech. This study improves the equal error rate (EER) to 26.47%. In other research Mohan and Patil [7] using self organizing map (SOM) and latent dirichlet allocation (LDA) succeeded in increasing crop prediction accuracy by 7-23%. DNN is also used by Saleem and Khattak [8] in research on speech separation, which produces promising performance. However, behind the advantages of DL, this method also has several weaknesses, including higher accuracy performance requiring large data sets, overfitting, and computational process efforts that require large resources. In addition to the need for large data volumes, research in the field of modern classification also faces the problem of multiple classes when it comes to processing an explosion of features and limited data.

In several studies regarding speaker recognition multi ethnic, the problem of limited data is an initial challenge in conducting research such as the study conducted by Hanifa *et al.* [9], which only had 62 recorded data of Malaysian speakers, and the study of Cole [10] to identify speakers in South East England with 227 speakers. To solve the problem of limited data, various approaches are used, among others, by using data augmentation (DA). DA is a method of expanding the quantity of data and has proven to be effective in conducting training on neural network. In speech signal processing, DA is proven to increase accuracy in audio classification using DL [11]. In the field of speech signal processing, several DA methods that are often used are adding white noise (AWN), time stretching (TS), pitch shifting (PS), mixup and speech augment. Some of these DA approaches are proven to be able to increase the quantity of research data so that research using DL which requires relatively large amounts of data can be done well so as to achieve a high level of accuracy.

Currently, in machine learning, the problems faced by researchers in an effort to increase the accuracy of speaker recognition include the number of classes that must be classified in the imbalance dataset. One of the problems encountered in machine learning classification is unbalanced multiclass which will cause model inaccuracies in predicting data. This problem can cause prediction errors in machine learning algorithms, so it needs to be resolved immediately. In research conducted by Khan *et al.* [12] and Mi *et al.* [13] the DA approach used the generative adversarial network (GAN) method to solve multi-class problems which achieved an accuracy increase of 6.4%. In several studies related to the implementation of DA in speech signal processing, various methods used include AWN, such on research that has been done by Morales *et al.* [14], which seeks to increase speaker recognition accuracy by adding noise effects.

Jacques and Roebel [15] also used the approach to adding noise in the research conducted. TS is also used by several researchers in speech signal processing research, TS functions to change the speed of audio signal duration as on research that has been done by Sasaki *et al.* [16] and Aguiar *et al.* [17] to classify music genres with convolutional neural network (CNN). However, the AWN and TS methods used still cannot produce high accuracy for speech recognition because they only achieve an accuracy level of around 70% to 80%. Another approach used in voice-based augmentation data is PS. PS method is also used in several studies conducted by Morbale and Navale [18] in processing audio files, Rai and Barkana [19] in processing musical instruments and Ye *et al.* [20] with CNN using UME and TIMIT datasets. In research conducted by Ye the Pitch Shifting method has achieved an accuracy rate above 90% with the highest accuracy of 98.72%.

This paper aims to improve the performance of the multi-ethnic speaker recognition model with a pitch shifting method based on deep neural networks and consists of several parts, namely the Introduction section in Chapter 1, which describes the research problems and other research that has been carried out and the relationship with several other studies. Chapter 2 is a literature review that contains a section that contains the study of related topics. In Chapter 3, proposed methodology contains the proposed model in solving the problem. Chapter 4 is a results and discussion that contains experiment results from the research that has been done. The final part is Chapter 5, namely conclusion, which contains several conclusions and solutions for the research carried out.

2. METHOD

This study proposed a pitch shifting deep neural network (PSDA-DNN) which has the best performance when implemented in voice signal processing supported by MFCC as a feature extraction method and a DNN that processes multiethnic speaker recognition classification. This research begins with processing the dataset of multiethnic speakers followed by the preprocessing process. Data augmentation is a step that

provides a solution to the limited dataset of multiethnic speakers. The dataset is then extracted using the mel frequency cepstral coefficient (MFCC) approach and then the result of feature extraction is included in the 7 Layer DNN architecture. As the last step, measuring the performance results of the proposed framework, among others, uses measures of accuracy, precision, recall, and F1. The proposed method can be seen in Figure 1.

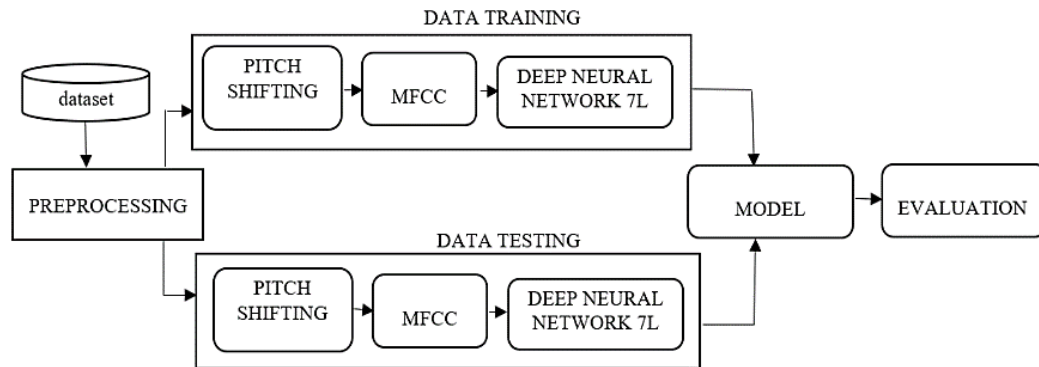


Figure. 1 Proposed method

2.1. Datasets and preprocessing

Research on multi-ethnic speaker recognition uses the sample for recognizing ethnic speakers dataset from Indonesia taken from Youtube 301 language in Indonesia [21]. The dataset was compiled using 70 ethnic male speakers from hundreds of ethnic groups in Indonesia. Sound processing using Adobe Audition CS6, the voice of tribal speakers is carried out by a sampling process by taking 10 tribal speakers each with a duration of 1 second. The sampling process uses a standard sample rate (SR) of 44.100 Hz with 32-bit bit depth, mono 32-bit Floating Point. The next stage is the reprocessing process, which is an important process in data mining processing. Preprocessing is a step that needs to be done in data mining to get good data quality to reduce processing time and get the desired results. This study uses the Adobe Audition CS6 application in preprocessing data with noise reduction facilities to remove noise in the speaker's speech data.

2.2. Data augmentation

DA is one method that is often used in increasing the quantity of dataset needed in research. DA is an approach that aims to increase the size of the data quantity and is a powerful technique used in the field of data mining and data processing for regression and classification purposes. PS is a DA method in sound signal processing by raising or lowering the original voice pitch in audio without affecting the long duration of the recorded sound. PS is used in this study because it has the advantage that the overall spectral envelope does not change so that it can achieve high-quality output. The process in the PS approach can be seen in Figure 2. In this study, PS results will be compared with 2 other DA methods, namely Awn and TS which are widely implemented in various research fields such as sound recording, music production, music learning, and foreign languages. The original speaker's voice signal is processed using the Pitch Shifting approach. The results of processing the voice signal before and after using the PS method can be seen in Figure 2.

2.3. Feature extraction and deep neural network seven layer

2.3.1. Mel-frequency cepstral coefficient

MFCC is a feature extraction method used in this study. MFCC is a robust approach used in speech recognition. The speech signal of the ethnic speaker consisting of 700 wav voices from 70 Ethnic was extracted using the python application with MFCC settings of frame 900 lengths, 25 frame shifts 10, window type hamming, preemphasis coefficient 0.97, number of cepstral coefficient 13 and number of lifters 22. In the MFCC approach, the voice signal will be processed through the following steps:

i) Preemphasis

Is a process that will be carried out after the sound sampling process, this process serves to reduce noise at the sound source. The preemphasis process in the time domain can be formulated as:

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

A indicates the constant which is at $0.9 < a < 1.0$

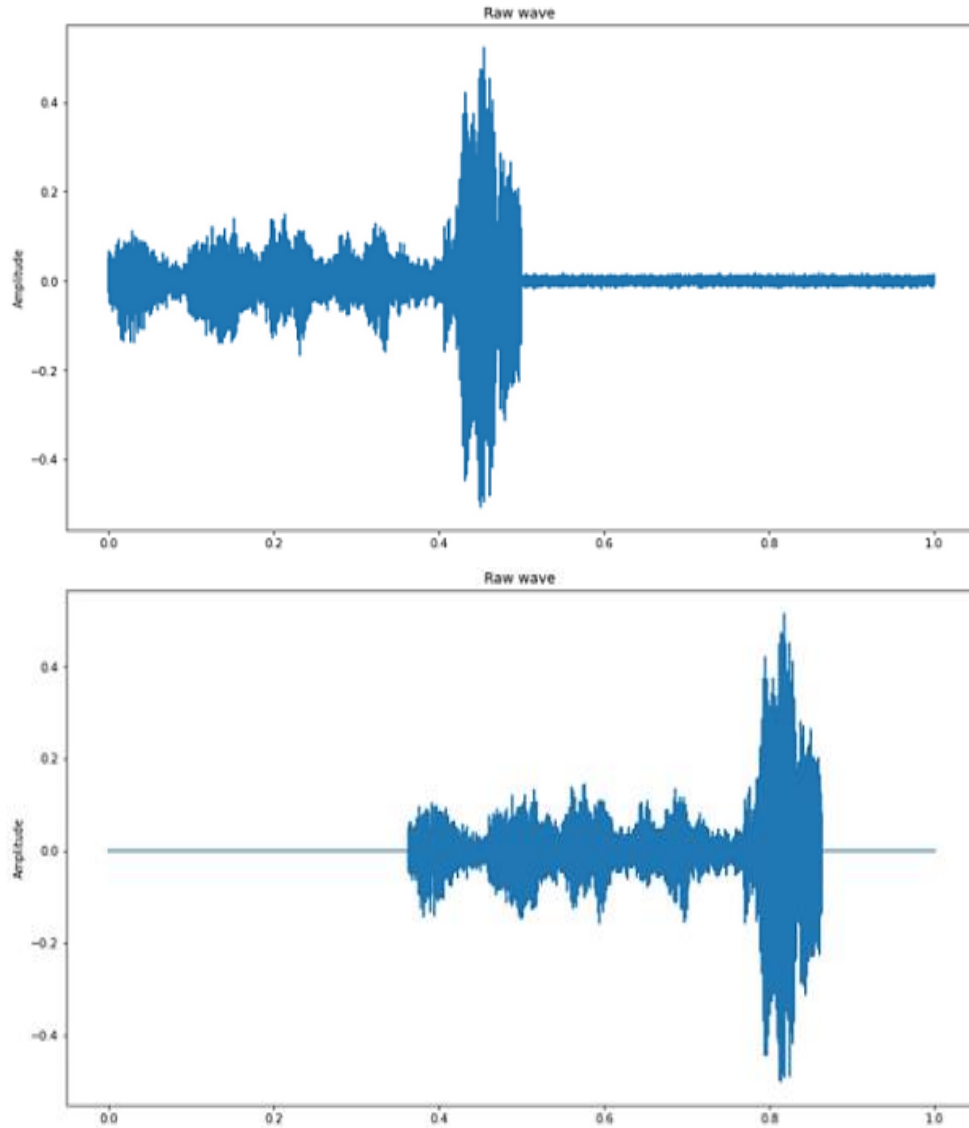


Figure 2. Original and pitch shifting speech signal

ii) Frame blocking

The speech signal will then enter the Frame Blocking process which functions to divide the sound into several parts of the frame.

iii) Windowing

Windowing is a step to analyze a long signal by taking the right part to be processed in the next stage. If the window is defined as $w(n)$, $0 < n < N - 1$ where N is the number of samples in each frame, then the windowing signal can be formulated as:

$$y_1(n) = x_1(n) w(n), 0 \leq n \leq N - 1 \quad (2)$$

iv) Fast fourier transform (FFT)

Fourier transform is used to convert the time series of signals in the form of a limited time domain into a frequency spectrum, while FFT is part of a fast Discrete Fourier Transform (DFT) algorithm that converts frames into N samples starting from the time domain to the frequency domain. The result of this processing is referred as Cepstrum which is formulated as:

$$x(n) = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N} \quad (3)$$

Where $n = 0, 1, 2, \dots, N-1$ and $j = \text{sqrt}-1$ while $X[n]$ is an n -frequency form resulting from the Fourier Transform mechanism.

v) Mel-frequency wrapping

In this step the existing FFT signals are grouped in a triangular filter which aims to multiply the FFT value with the appropriate filter gain which then the results will be summed. The wrapping process into a signal in the frequency domain can be formulated as:

$$x_i = \log_{10}(\sum_{k=0}^{N-1} |x(k) H_{i(k)}|) \quad (4)$$

Where $i = 1, 2, 3, \dots, M$, M is the sum of the triangle filters and $H_i(k)$ is the value for the triangular i -filter for the acoustic frequency k .

vi) Cepstrum

In order for the signal to be heard by humans, it is necessary to convert the signal into a time domain using a discrete cosine transform (DCT). The final result of this process is referred to as Mel Frequency Cepstral Coefficients which is formulated as:

$$c_j = \sum_{i=1}^K x_i \cos((j-1)/2 \frac{\pi}{K}) \quad (5)$$

C_j shows the MFCC coefficient. X_j is the strength of the Mel Frequency spectrum, $j = 1, 2, 3, \dots, K$ is the expected coefficient and M indicates the number of filters.

2.3.2. Deep neural network seven layer

One of the finest DL techniques is the DNN. DNN has the advantage of building more accurate models. In this work used DNN 7 layers with architectural network as in Table 1. The DNN architecture used in this study consists of seven layers consisting of dense or also called fully connected layers, indicating that the layer in which there are neurons connected to neurons in the previous layer. Layer 1 consists of 193 nodes which is an input layer that shows 193 features generated from the extracted features. Between layers is given a dropout function which is a technique used to solve overfitting problems and prediction problems in large neural networks. Layers two to seven use half of the number of nodes in the previous layer in order to reduce the complexity of calculations on each layer.

Table 1. DNN7L architecture

Layer (type)	Output shape	Layer (type)	Output shape
Dense 1 (dense)	(None,193)	Dense 5 (dense)	(None,50)
Dense 2 (dense)	(None,400)	Dropout_4 (Dropout)	(None,50)
Dropout_1 (Dropout)	(None,400)	Dense 6 (dense)	(None,25)
Dense 3 (dense)	(None,200)	Dropout_5 (Dropout)	(None,25)
Dropout_2 (Dropout)	(None,200)	Dense 7 (dense)	(None,15)
Dense 4 (dense)	(None,100)	Dropout_6 (Dropout)	(None,15)
Dropout_3 (Dropout)	(None,100)	Dense 8 (dense)	(None,8)

2.4. Evaluation

The end result of the process of Indonesian ethnic speakers recognition is an evaluation process by measuring the performance of the proposed model by evaluating the level of accuracy, precision, recall, and F1 measure. Accuracy is the ratio of the number of cases that were predicted with correct answers compared to the total number of cases, while recall is the ratio of the number of positive cases that were predicted correctly to the number of positive cases that were predicted. Accuracy, precision, recall and F1-score measure can be calculated using the (6) to (9), respectively.

$$\frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (6)$$

$$\frac{TP}{(TP+FP)} \times 100\% \quad (7)$$

$$\frac{TP}{(TP+FN)} \times 100\% \quad (8)$$

$$\frac{2 \times (\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (9)$$

Where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively. Performance measurement in this study also uses recall which is the ratio of true positive cases that are predicted to be positive.

3. RESULTS AND DISCUSSION

Dataset on multi-ethnic speakers in Indonesia were tested using DA approach, namely AWN, PS and TS then trained using DNN using a split ratio of 70:30, 80:20 and 90:10. The test results show that the PSDA-DNN has better performance than adding white noise data augmentation deep neural network (AWNDA- DNN), and time stretching data augmentation deep neural network (TSDA-DNN) methods. The following are the results of the PSDA-DNN model performance testing process using a 70:30 split ratio as depicted in Figure 3.

Testing on the 70:30 split ratio resulted in an accuracy level of 98.55%, precision, recall and F1 measure each of 94.37%. This performance result shows that PSDA-DNN will be more robust when used on larger data. Classification using various machine learning methods using many classes is not easy and cause various problems in learning [22]. An appropriate approach is needed in managing the dataset. The following is a comparison of the number of classes processed between the various types of research shown in Table 2.

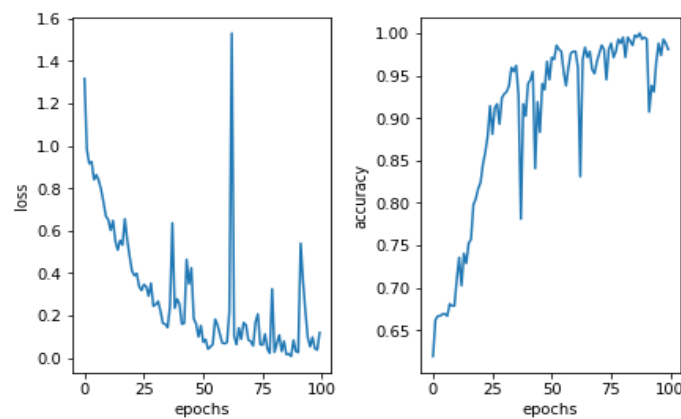


Figure 3. PSDA-DNN performance

Table 2. Comparison with another datasets

Methods	Datasets	ΣClass	Acc (%)
SVM [9]	Ethnicity of Malaysian dataset	4	57.7
CNN [23]	Urdu speakers	4	87.5
Deep Belief Network (DBN) [24]	Accented spoken English corpus	6	90.2
DNN [25]	TITML-IDN, OpenSLR	4	98.9
PSDA-DNN	Indonesian multiethnics speakers	42	99.2

Based on the comparison results presented in Table 2, even though it has more classes, the proposed model, namely PSDA-DNN, produces better model performance than other machine learning methods with an accuracy of up to 99.2%. The achievement of a high level of performance at PSDA-DNN was due to a good preprocessing process on the speech signal dataset which was then carried out by data augmentation process with PS and proper classification using the DNN 7-layer approach. The PSDA-DNN method also produces the most effective performance in comparison to various other techniques using the Indonesian Multiethnics Speakers dataset. The results of the comparison of these methods can be seen in Table 3.

Table 3. Comparison with another methods

Methods	Accuracy
K-Nearest Neighbor	92%
Random-Forest	81%
DNN	98.4%
PSDA-DNN (Ours)	99.2%

The results of this study will be compared with several methods with classical machine learning and deep learning approaches. Table 3 shows that the PSDA-DNN method has better performance compared to other methods. In several comparisons of the performance of the methods as previously presented, it can be concluded that the proposed method has the best performance.

4. CONCLUSION

Study in the area of speaker identification is an interesting topic and challenges researchers around the world to work hard to make new scientific contributions, including research on Indonesian multiethnic speaker recognition. The DL method is an approach that has been chosen, especially for processing large amounts of data, including sound signal processing. However, the problem of multiple classes and data imbalanced causes low accuracy because the model performance is not optimal. The PS Approach in Data Augmentation is a solution in increasing the quantity of data and as a solution for multiple class classification problems. Obtaining a high model accuracy performance of 99.27% through the proposed model, namely PS-DNN is one of the solutions to overcome the problem of multiple classes and imbalanced datasets in machine learning. This study proposes the PSDA-DNN approach which is a multi-ethnic speaker recognition method that uses the PSDA technique which is supported by the MFCC and DNN methods in processing speech signals. The research results show that PSDA-DNN has better performance compared to other approaches such as AWN and TS which are also DNN-based in processing speech signals. The PSDA-DNN approach produces an average level of accuracy of 99.27%, precision, recall, and F1 measure of 97.60%.

REFERENCES

- [1] P. Motlicek, S. Dey, S. Madikeri, and L. Burget, "Employment of subspace gaussian mixture models in speaker recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, Apr. 2015, pp. 4445–4449, doi: 10.1109/icassp.2015.7178811.
- [2] K. V. Veena and D. Mathew, "Speaker identification and verification of noisy speech using multitaper MFCC and gaussian mixture models," Dec. 2015, doi: 10.1109/picc.2015.7455806.
- [3] N. Maghsoodi, H. Sameti, H. Zeinali, and T. Stafylakis, "Speaker recognition with random digit strings using uncertainty normalized HMM-based i-vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1815–1825, Nov. 2019, doi: 10.1109/taslp.2019.2928143.
- [4] J. S. Hussein, A. A. Salman, and T. R. Saeed, "Arabic speaker recognition using HMM," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 1212–1218, Aug. 2021, doi: 10.11591/ijeecs.v23.i2.pp1212-1218.
- [5] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database," in *2019 IEEE 4th International Conference on Computer and Communication Systems ICCCS*, Feb. 2019, pp. 130–133, doi: 10.1109/ccoms.2019.8821751.
- [6] J. Guo *et al.*, "Deep neural network based i-vector mapping for speaker verification using short utterances," *Speech Communication*, vol. 105, pp. 92–102, Dec. 2018, doi: 10.1016/j.specom.2018.10.004.
- [7] P. Mohan and Kiran Patil, "Deep learning based weighted SOM to forecast weather and crop prediction for agriculture application," *International Journal of Intelligent Engineering and Systems*, vol. 11, no. 4, pp. 167–176, Aug. 2018, doi: 10.22266/ijies2018.0831.17.
- [8] N. Saleem and M. I. Khattak, "Deep neural networks based binary classification for single channel speaker independent multi-talker speech separation," *Applied Acoustics*, vol. 167, p. 107385, Oct. 2020, doi: 10.1016/j.apacoust.2020.107385.
- [9] R. M. Hanifa, K. Isa, and S. Mohamad, "Speaker ethnic identification for continuous speech in Malay language using pitch and MFCC," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 207–214, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp207-214.
- [10] A. Cole, "Identifications of speaker ethnicity in South-East England: multicultural London English as a divisible perceptual variety," in *Proceedings of the LREC 2020 Workshop on Citizen Linguistics in Language Resource Development*, 2020, pp. 49–57.
- [11] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, p. 101084, May 2020, doi: 10.1016/j.ecoinf.2020.101084.
- [12] M. H.-M. Khan *et al.*, "Multi-class skin problem classification using deep generative adversarial network (DGAN)," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–13, Mar. 2022, doi: 10.1155/2022/1797471.
- [13] Q. Mi, Y. Hao, M. Wu, and L. Ou, "An enhanced data augmentation approach to support multi-class code readability classification," in *Proceedings of the International Conference on Software Engineering and Knowledge Engineering SEKE*, Jul. 2022, pp. 48–53, doi: 10.18293/SEKE2022-130.
- [14] N. Morales, L. Gu, and Y. Gao, "Adding noise to improve noise robustness in speech recognition," Jan. 2007, doi: 10.21437/interspeech.2007-335.
- [15] C. Jacques and A. Roebel, "Data augmentation for drum transcription with convolutional neural networks," Sep. 2019, doi: 10.23919/eusipco.2019.8902980.
- [16] T. Sasaki *et al.*, "Time stretching: illusory lengthening of filled auditory durations," *Attention, Perception, Psychophys*, vol. 72, no. 5, pp. 1404–1421, Jul. 2010, doi: 10.3758/app.72.5.1404.
- [17] R. L. Aguiar, Y. M. G. Costa, and C. N. Silla, "Exploring data augmentation to improve music genre classification with convNets," in *2018 International Joint Conference on Neural Networks IJCNN*, Jul. 2018, pp. 1–8, doi: 10.1109/ijcnn.2018.8489166.
- [18] P. R. Morbale and M. Navale, "Design and implementation of real time audio pitch shifting on FPGA," *International Journal of Innovative Trends in Engineering*, vol. 4, no. 2, pp. 81–88, 2015.
- [19] A. Rai and B. D. Barkana, "Analysis of three pitch-shifting algorithms for different musical instruments," in *2019 IEEE Long Island Systems, Applications and Technology Conference LISAT*, May 2019, pp. 1–6, doi: 10.1109/lisat.2019.8817334.
- [20] Y. Ye, L. Lao, D. Yan, and R. Wang, "Identification of weakly pitch-shifted voice based on convolutional neural network," *International Journal of Digital Multimedia Broadcasting*, vol. 2020, pp. 1–10, Jan. 2020, doi: 10.1155/2020/8927031.





- [21] "301 languages in Indonesia-regional dialects #2 (In Indonesia: 301 languages in Indonesia-bahasa Logat Dialek Daerah #2)." Indonesia Ideas, 2018, [Online]. Available: <https://www.youtube.com/watch?v=FkwXbCY1rWg>.
- [22] Y. Xue and M. Hauskrecht, "Active learning of multi-class classification models from ordered class sets," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5589–5596, Jul. 2019, doi: 10.1609/aaai.v33i01.33015589.
- [23] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker identification using a hybrid CNN-MFCC approach," Mar. 2020, doi: 10.1109/icetst49965.2020.9080730.
- [24] R. Upadhyay and S. Lui, "Foreign English accent classification using deep belief networks," in *2018 IEEE 12th International Conference on Semantic Computing ICSC*, Jan. 2018, pp. 290–293, doi: 10.1109/icsc.2018.00053.
- [25] K. Azizah, M. Adriani, and W. Jatmiko, "Hierarchical transfer learning for multilingual, multi-speaker, and style transfer DNN-based TTS on low-resource languages," *IEEE Access*, vol. 8, pp. 179798–179812, 2020, doi: 10.1109/access.2020.3027619.

ACKNOWLEDGEMENTS





We would like to thank to the Universitas Stikubank for granted funding through scientific publication incentives.

BIOGRAPHIES OF AUTHORS







Kristiawan Nugroho     works as a lecturer at the faculty of information technology and industry, Universitas Stikubank. He obtained a bachelor's degree in 2001 in the information systems department, Faculty of Computer Science, Universitas Dian Nuswantoro Semarang, then in 2007 He graduated from Universitas Dian Nuswantoro with a master's degree in informatics engineering. He also obtained Doctoral degree in computer science with a concentration in Machine Learning and Artificial Intelligence in 2022 at Dian Nuswantoro University Semarang. He has conducted various researches in machine learning, speech recognition and sentiment analysis. He can be contacted via email kristiawan@edu.unisbank.ac.id.







Isworo Nugroho     works as a lecturer at the faculty of information technology and industry, Universitas Stikubank. He obtained a bachelor's degree in 2001 in the management department, Faculty of Ekonomis, Stikubank University, then in 2003 he obtained a Master's degree in computer science, Gajah Mada University. He has conducted various researches in text processing, data mining and statistical science. He can be contacted via email isworo@edu.unisbank.ac.id.



De Rosal Ignatius Moses Setiadi     a Bachelor of Science in Informatics Engineering from Universitas Soegijaprana, Indonesia, and a Master of Science in Informatics Engineering from Universitas Dian Nuswantoro Semarang, both in 2012. He is presently a lecturer and researcher at the Faculty of Computer Science at Universitas Dian Nuswantoro in Semarang, Indonesia. He has written more than 138 peer-reviewed journal and conference publications that Scopus has indexed. His areas of interest in study include machine learning, cryptography, image steganography, and watermarking. His email address is moses@dsn.dinus.ac.id.



Prof. Omar Farooq     Omar Farooq joined the Department of Electronics Engineering, AMU Aligarh as Lecturer in 1992 and is currently working as a professor. He was awarded Commonwealth Scholar from 1999-2002 towards PhD at Loughborough University, UK, and a one-year postdoctoral fellowship with the UKIERI in 2007-2008. With a focus on speech recognition, signal processing is his broad area of study interest. He has approximately 250 publications published by him or with him in reputable academic journals and conference proceedings, and he has helped 9 scholars complete their PhDs. He is a Senior Member, Institute of Electrical and Electronics Engineers, (IEEE, USA). He can be contacted via email omar.farooq@amu.ac.in.