

# Enhanced multi-ethnic speech recognition using pitch shifting generative adversarial networks

Kristiawan Nugroho<sup>1</sup>, Kristophorus Hadiono<sup>1</sup>, Felix Andreas Sutanto<sup>1</sup>, Dhendra Marutho<sup>2</sup>,  
Omar Farooq<sup>3</sup>

<sup>1</sup>Faculty of Information and Industrial Technology, Universitas Stikubank, Semarang, Indonesia

<sup>2</sup>Faculty of Engineering and Computer Science, Universitas Muhammadiyah Semarang, Semarang, Indonesia

<sup>3</sup>Department of Computer Science, Alagarh Muslim University, Aligarh, India

## Article Info

### Article history:

Received Oct 29, 2023

Revised Feb 18, 2024

Accepted Feb 28, 2024

### Keywords:

Data augmentation

Generative adversarial network

Multi-ethnic

Pitch shifting

Speech recognition

## ABSTRACT

Research in the field of speech recognition is a challenging research area. Various approaches have been applied to build robust models. A problem faced in speech recognition research is overfitting, especially if there is insufficient data to train the model. A large enough amount of data can train the model well, resulting in high accuracy. Data augmentation is an approach often used to increase the quantity of dataset. This research uses a data augmentation approach, namely pitch shifting, to increase the quantity of speech dataset, which is then processed into spectrogram data and then classified using a generative adversarial network (GAN). Using the pitch shifting-generative adversarial network (PS-GAN) model, this research produces high accuracy performance in multi-ethnic speech recognition, namely 98.43%, better than several similar studies.

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



## Corresponding Author:

Kristiawan Nugroho

Faculty of Information Technology and Industry, Universitas Stikubank

Jl. Tri Lomba Juang, Kota Semarang, Indonesia

Email: kristiawan@edu.unisbank.ac.id

## 1. INTRODUCTION

Research in the field of speech recognition continues to develop, producing various new algorithms with increasingly better performance. The use of deep learning, especially convolutional neural networks (CNN) and recurrent neural networks (RNN), has resulted in significant improvements in speech recognition performance. Further research into network architecture and more efficient training methods will continue to be needed and further explored in the field of speech recognition, and developments in this technology continue to open up new opportunities. Speech recognition, also known as automatic speech recognition (ASR), is a research area that is still active [1] and challenging today. The results of ASR research have been implemented in various areas of life, including in the health sector through medical record systems [2] and rural healthcare systems [3], in business and industry, in the language sector [4], and in signal processing [5].

Research in the field of ASR has produced various new methods that have increasingly better performance. However, the problem of limited dataset quantity, especially in deep learning architectures, is an interesting challenge to solve [6]. The large quantity of data in the deep learning architecture is important, especially in training data, for producing robust performance models. A large quantity of data results in a better generalization process, preventing overfitting, overcoming data imbalances, and improving the performance of the resulting model. Several approaches to increasing the quantity of research data have also been carried out by researchers, including the data augmentation method, namely, the way to increase the amount of data is by augmenting the data. This involves rotating, cropping, shifting, or changing colors on images, text, or other data.

Several studies involving a speech-based data augmentation approach have been carried out by González *et al.* [7] to identify speakers in stressful conditions using the voice activity detector (VAD) module algorithm to achieve an accuracy level of 99.45%. In another study, data augmentation was also used by Kathania *et al.* [8] on children's voice recognition in noisy environments. Using the deep neural network-hidden Markov model (DNN-HMM) approach, this research reduced the word error rate (WER) to 12.09%. Data augmentation techniques were also used by Praseetha and Joby [9] in speech-based emotion recognition using the gated recurrent unit (GRU) on the Toronto emotional speech set (TESS) dataset, which achieved a model accuracy level of 93%. Still, in the speech emotion recognition (SER) Atmaja and Sasou [10], the Japanese Twitter-based emotional speech and IEMOCAP datasets using the wav2vec 2.0 feature and data augmentation achieved a model accuracy level of 97.95%.

One method of sound-based data augmentation is pitch shifting, which is a way of changing the tone or pitch of a speech in music or audio. Several studies using the pitch shifting technique, including those conducted by Ning [11], on individual speakers' pitch disturbances. The pitch shifting approach was also used by Rosenzweig *et al.* [12] who proposed an adaptive pitch shifting approach to adjust the intonation of a cappella singer. The research results showed the success of this approach in regulating voice intonation. In other research, Ye *et al.* [13] used pitch shifting combined with a CNN architecture using the TIMIT and UME datasets to produce a performance that showed better generalization capabilities than the gaussian mixture model (GMM) method. The pitch-shifting approach is also often used in data augmentation, such as in [14], [15], which is supported by using CNN to produce a promising level of model performance. Several studies that have been conducted show that pitch shifting is a suitable method to use to increase the quantity of research data.

Several researchers have used the generative adversarial network (GAN) approach in some of the research they have carried out. GAN is a method that can produce new, realistic data, such as images, text, or sound, which can be used in various creative applications, such as generative art, facial synthesis, or music creation. In several studies, GANs have been proven to provide superior results in image segmentation [16] and other research in computer vision [17]. Augmentation can be done traditionally, as in Adityawan *et al.* [18], but using GANs can provide more optimal results in training models [19]. Several image-based studies using GANs have been carried out by Dewi *et al.* [20] to detect very complex traffic sign images using least squares generative adversarial networks (LSGAN), deep convolutional generative adversarial networks (DCGAN), and wasserstein generative adversarial networks (WGAN). Li *et al.* [21] also used GANs to improve methods for image recognition at low light levels, where the research results showed that the proposed method had good performance in recognizing images in low light conditions. Research in the field of ASR also uses the GAN approach, such as research conducted by Pan and Zheng [22] with CNN-GAN on electroencephalogram (EEG) signal detection, which has succeeded in providing model performance of 90.26%. Baek [23] also used DCGANs to improve SER performance using the RAVDESS and EmoDB datasets, successfully achieving a model performance level of unweight accuracy (UA) of up to 91.3%. In other research, Jia and Zheng [24] also used multi-channel time–frequency domain generative adversarial networks (MC-TFD GANs) and Mixup on SER, showing that the mean opinion score (MOS) and UA scores of the utterances produced were the synthesis method improved. The increases were 4% and 2.7%, respectively. GAN has positively contributed to improving the performance of speech recognition models.

In the field of research regarding ethnic speech recognition, several studies have been carried out, including by Mouaz *et al.* [25], using the Moroccan dialect with the hidden Markov model method, which achieved a model accuracy of 90%. This research proposes a method for improving ASR detection performance using pitch-shifting data augmentation and GAN through an approach abbreviated as pitch shifting generative adversarial network (PS-GAN), which aims to increase the number of sound datasets, which are then processed into spectrograms and then continue with the PS-GAN model creation process by training the spectrogram data ending with evaluation of the performance of the model that has been produced.

This paper consists of several parts, namely the Introduction, the first part of the paper that discusses the background of the problem, and the development of similar research that has been carried out along with the proposed approach. Section 2 contains methods explaining the research stages and the proposed model. Section 3 contains results and discussion, which discusses the results of research and analysis that have been carried out compared with the results of another research. The final section of this paper is the conclusion, which contains conclusions and future research the researcher will conduct.

## 2. METHOD

Various research on speech recognition has been carried out using various approaches, including classical machine learning and deep learning. However, the various studies that have been carried out still have various limitations. Some of the gaps that occur include problems with the available sound dataset, data quality is still a concern. Unbalanced, incomplete, or unrepresentative data can produce biased or less accurate models. Apart from that, the lack of integration of information from various modalities such as text, images and social

context can increase the accuracy and reliability of the speech recognition system so that a robust model is needed as a solution to the problems faced.

### 2.1. Proposed method

This research uses the PS-GAN approach to process multi-ethnic speech recognition. This work also processes speech data into a spectrogram, enabling more profound analysis, better interpretation, and further signal processing applications for various purposes, including audio analysis. Figure 1 illustrates several phases of the research project. The research stages carried out were divided into 3 stages, namely dataset processing, speech conversion to spectrogram and processing using the GAN approach.

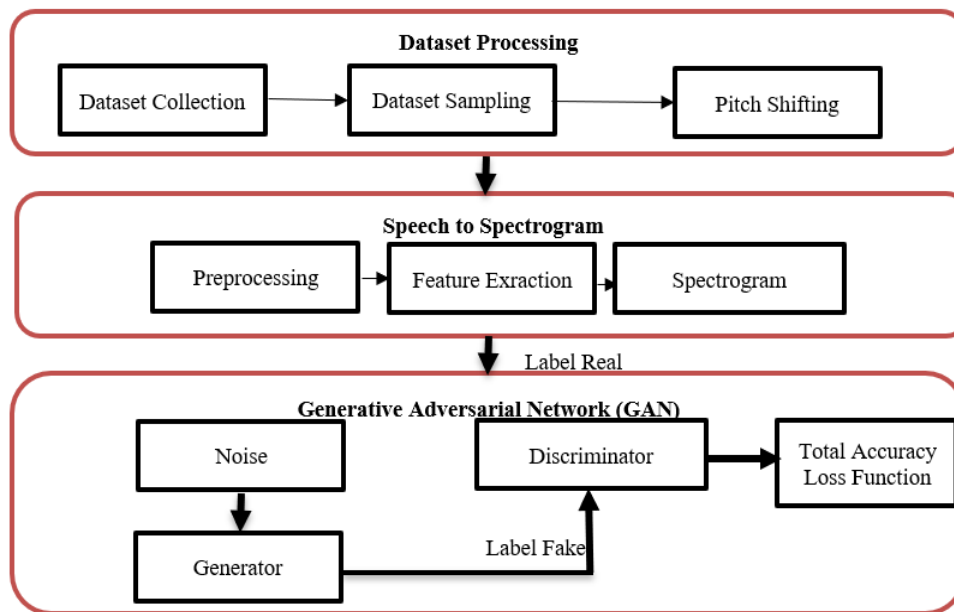


Figure 1. Proposed method

This research uses the pitch shifting method, which is used to increase the quantity of research data as well as GAN, which is one of the most phenomenal approaches currently, is a type of artificial neural network architecture that was developed to produce new, very realistic data in various fields, such as images, text or sound. GANs consist of two opposing neural networks, namely a generator and a discriminator, which compete with each other in the learning process. The GAN approach is used in image creation and style transfer and in text, sound, video processing, and others. Goodfellow first introduced GANs. They consist of two main components: a generator and a discriminator, and they work together in a competitive process to produce realistic data. The generator is an artificial neural network that aims to produce fake data similar to real data. At the same time, the discriminator is an artificial neural network whose job is to differentiate between real and fake data. The research stages in Figure 1 can be described as:

- Data collection: the first step is to collect speech data, which will be used as a training dataset. This research uses the 301 languages in Indonesia dataset, which can be downloaded via the web address <https://www.youtube.com/watch?v=FkwXbCY1rWg&t=1s>.
- Dataset sampling: in the research carried out using the Adobe Audition application, where we took speech samples from 10 ethnic groups, namely Banda Aceh, Padang, Ambon, Bali Tabanan, Banten, Banyumas, Betawi, Jawa Timur, Yogyakarta, and Madura. We sampled each sound for 2 seconds.
- Pitch shifting: this is a stage to increase the quantity of research data. Speech data is processed using the Pitch Shifting approach by changing the pitch of the voice without changing the duration or tempo for each ethnic speech.
- Preprocessing of speech data: speech data must be preprocessed before being used in GAN. This stage includes noise removal, amplitude normalization, and data slicing into smaller speech segments (frames).
- Feature extraction: before creating a spectrogram, steps need to be taken to extract features from speech data. One of the commonly used features is mel-frequency cepstral coefficients (MFCC). In this study, MFCC was used to extract multi-ethnic speech features.

- Spectrogram creation: after the feature extraction process is carried out, a spectrogram can be created. A spectrogram is a visual representation of the frequency spectrum of a sound signal over time. In this research, a Python script was used to build several spectrograms from the speech data features that have been produced.
- GAN modeling: the next stage is to create a GAN model that can produce spectrograms. GANs consist of two main networks: generator and discriminator. The generator will produce fake spectrograms, while the discriminator will try to differentiate between real spectrograms (from the training dataset) and fake spectrograms produced by the generator.
- GAN training: the GAN model must be trained using the original spectrogram dataset. The generator aims to learn the actual data distribution and produce a realistic spectrogram.
- Evaluation and tuning: after training, it is necessary to evaluate the results. Tuning hyperparameters and refining the GAN model to get better results is necessary.
- Spectrogram generation: once the GAN model is trained properly, it can generate fake spectrograms. It can be used in various applications, such as sound synthesis or signal processing.

## 2.2. Evaluation

The loss function measures the extent to which the GAN model has succeeded in producing realistic data. The purpose of the loss function is to measure how far the model predictions are from the actual target value. The (1) to calculate the loss function.

$$\min G \max D V(D, G) = E_{x \sim p_{\text{data}}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

Where  $G$  is the generator, and  $D$  is the discriminator in this formula.  $P(z)$  is the generating distribution;  $x$  is a sample from  $p_{\text{data}}(x)$ ;  $z$  is a sample from  $P_z$ .  $p_{\text{data}}(x)$  is the distribution of real data. Then, the generator network is  $G(z)$ , and the discriminator network is  $D(x)$ .

The accuracy of multi-ethnic speech recognition was also measured. Accuracy measurements aim to evaluate how well the model can make predictions or classifications. it was formulated in (2).

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

In (2), it can be explained that TP is true positive, While TN is true negative. FP is also called false positive, and FN is called false negative. Measuring accuracy is an important step in developing a machine-learning model to ensure the quality, consistency, and reliability of the model being built.

## 3. RESULTS AND DISCUSSION

Various studies on speech recognition have been carried out, but none has yet explicitly explained that the use of representative and diverse training data is crucial in developing a reliable speech recognition system. Previous studies still lacked training data, including various accents, dialects, and languages. This research proposes PS-GAN as a method with better model performance than other methods in training multi-ethnic speech. Several stages of the research carried out can be explained.

### 3.1. Dataset processing

This study used a speech dataset for research on the recognition of multi-ethnic speakers in Indonesia [24], a dataset from dozens of ethnic speeches in Indonesia. Still, this study will take the dataset from 10 ethnicities representing tribes in Java and outside Java. Then, the multi-ethnic speech signal must be carried out in a speech sampling process for 2 seconds to select speakers consisting of the tribes of Banda Aceh, Padang, Ambon, Bali Tabanan, Banten, Banyumas, Betawi, Jawa Timur, Yogyakarta, and Madura. After collecting the speech sampling results, it turns out it has a varying number of votes, so it is determined that each ethnic group will be filled with ten vote samples so that the shortfall in the number of votes will be added using the pitch shifting data augmentation approach.

### 3.2. Pitch shifting data augmentation

Pitch shifting is a technique commonly used in audio signal processing and machine learning to change the pitch (frequency) of an audio recording while maintaining other characteristics of the original sound, such as duration and timbre. This technique is frequently used in various applications, including speech recognition, music analysis, and audio synthesis. In this work, the pitch shifting method was used because it has the advantage of increasing the accuracy of speech recognition. This is because the pitch shifting process can help equalize the high and low levels of sound so that speech recognition machines can more easily identify

vocal patterns and characteristics needed for recognition. Apart from that, pitch shifting can also be used to improve the sound quality recognized by the system. By adjusting the pitch of the voice, the sounds produced by the system can sound more natural and more easily recognized by humans. In this research, pitch shifting is used to increase the quantity of ethnic speech to 10 way with pseudocode in Algorithm 1.

#### Algorithm 1. Pitch shifting data augmentation pseudocode

```

1. Import the necessary audio processing library
2. shifted_audio = audio_processing_library.pitch_shift(audio_sample, semitone_shift)
3. Return shifted_audio
4. min_semitone_shift = -2
5. max_semitone_shift = 2
6. audio_samples = load_audio_dataset()
7. augmented_data = []
8. For each audio_sample in audio_samples:
9.   random_semitone_shift = random_choice(min_semitone_shift, max_semitone_shift)
10.  shifted_audio = pitch_shift(audio_sample, random_semitone_shift)
11.  augmented_data.append(shifted_audio)
12. Save_or_use_augmented_data(augmented_data)

```

### 3.3. Speech to spectrogram

After forming 100 multi-ethnic sound WAV files where ten ethnicities each contain ten sounds, the next step is to convert the sound files into a spectrogram, a signal representation in time [26]. Spectrograms make it possible to analyze sound frequencies over time. This can help in understanding the acoustic characteristics of a particular sound source. An example of the visualization results of spectrogram processing on multi-ethnic speech as seen in Figure 2. Each ethnicity will produce a spectrogram, as in Figure 2, which shows the changing frequency energy in an audio signal over time. Spectrograms are an important tool in audio signal analysis in various applications, including speech processing and recognition. In the field of speech signal processing, spectrograms are also used to detect and classify speech.

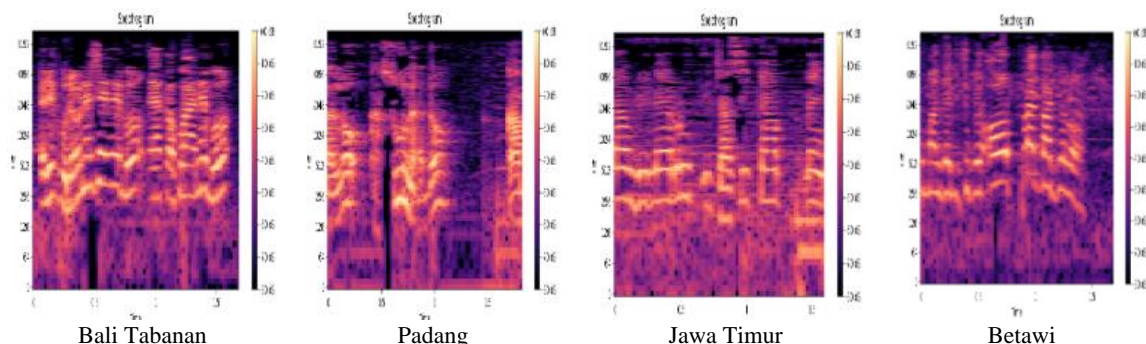


Figure 2. Multi-ethnic spectrogram

### 3.4. Generative adversarial network

In this work, the GAN approach was used by creating a generator and discriminator models using the Keras library in Python. Then, training data was collected using 20 epochs to calculate real accuracy, fake accuracy, and total accuracy. In this research, GAN was used to process multi-ethnic ethnic spectrogram data in Indonesia with maximum accuracy results in the 14<sup>th</sup> epoch with real accuracy 1.0, fake accuracy=0.96875 and total accuracy 0.984375, which means the model produced the most optimal performance in this epoch. Visualization of the performance results of the PS-GAN model, as shown in Figure 3.

It can be seen in Figure 3 that the generator loss and discriminator loss measurements are close to 0, which means the generator produces a better model. Likewise, in terms of total accuracy, it shows that the model has very good performance with an accuracy close to 100%. In the research that has been carried out, it was found that the proposed method (PS-GAN) produces a high level of accuracy in recognizing speech spectrograms of multi-ethnic speakers. These results are better than those of other studies in the field of speech recognition that also used the GAN approach. The results of measuring accuracy compared to other studies are shown in Table 1.

The results of the observations we have made through this research show that using the pitch shifting approach to increase the quantity of speech data for several ethnic groups in Indonesia produces new quality

data. This research also uses a deep learning approach: a GAN. PS-GAN has succeeded in producing a speech recognition model with superior performance.

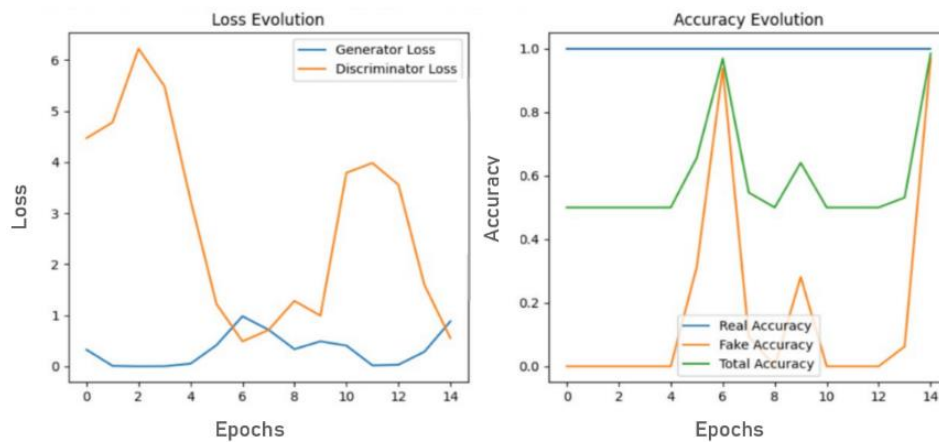


Figure 3. PS-GAN performance chart

Table 1. Comparison with other methods

Methods	Dataset	Accuracy (%)
Power spectral density-generative adversarial network (PSD-GAN) [22]	MAHNOB-HCI	70.34
Convolutional neural networks spectrogram image features (CNN-SIFs) [27]	Traffic sound datasets (TSD)	97.18
Spectrogram deep convolutional generative adversarial network (S-DCGAN) [28]	Power Spectral Density	91.25
Proposed Method (PS-GAN)	Indonesian Multi-ethnics Speech	98.43

#### 4 CONCLUSION

Investigations on multi-ethnic speech recognition face the problem of a lack of speech datasets, which will result in a reduced quantity of training data and cause the resulting model to perform poorly. This research succeeded in developing the PS-GAN method to process multi-ethnic speech data into a spectrogram and then classify it using GAN, whose performance has been proven to produce a high level of total accuracy and is proven to perform well in multi-ethnic speech recognition in Indonesia. Our findings prove that speech data was preprocessed, then converted to a spectrogram and processed using the PS-GAN approaches, providing the best accuracy performance results of 98.43%. In future research, the PS-GAN method will compare its performance with the CNN method in recognizing multi-ethnic speech spectrograms in Indonesia to obtain the best strong method for recognizing multi-ethnic speech.

#### ACKNOWLEDGEMENTS

Authors would like to thank the Universitas Stikubank for granted funding through scientific publication incentives and research facilities that have been provided.




#### REFERENCES

- [1] H. F. Pardede, P. Adhi, V. Zilvan, A. Ramdan, and D. Krisnandi, "Deep convolutional neural networks-based features for Indonesian large vocabulary speech recognition," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 610–617, 2023, doi: 10.11591/ijai.v12.i2.pp610-617.
- [2] X. Xia, Y. Ma, Y. Luo, and J. Lu, "An online intelligent electronic medical record system via speech recognition," *International Journal of Distributed Sensor Networks*, vol. 18, no. 11, 2022, doi: 10.1177/15501329221134479.
- [3] A. A. Onitilo, A. R. Shour, D. S. Puthoff, Y. Tanimu, A. Joseph, and M. T. Sheehan, "Evaluating the adoption of voice recognition technology for real-time dictation in a rural healthcare system: A retrospective analysis of dragon medical one," *PLoS One*, vol. 18, no. 3, pp. 1–17, 2023, doi: 10.1371/journal.pone.0272545.
- [4] L. Hu and J. Jia, "Smart speech recognition system for chinese lanuage learning enhancement," *Scientific Programming*, vol. 2022, 2022, doi: 10.1155/2022/1701474.
- [5] G. Liu, S. Cai, and C. Wang, "Speech emotion recognition based on emotion perception," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, 2023, doi: 10.1186/s13636-023-00289-4.




- [6] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, pp. 1–27, 2021, doi: 10.3390/s21041249.
- [7] E. R. -González, A. M. -Sánchez, A. G. -Antolín, and C. P. -Moreno, "Data augmentation for speaker identification under stress conditions to combat gender-based violence," *Applied Sciences*, vol. 9, no. 11, pp. 1–14, 2019, doi: 10.3390/app9112298.
- [8] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "Using data augmentation and time-scale modification to improve asr of children's speech in noisy environments," *Applied Sciences*, vol. 11, no. 18, 2021, doi: 10.3390/app11188420.
- [9] V. M. Praseetha and P. P. Joby, "Speech emotion recognition using data augmentation," *International Journal of Speech Technology*, vol. 25, no. 4, pp. 783–792, 2022, doi: 10.1007/s10772-021-09883-3.
- [10] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, pp. 1–14, 2022, doi: 10.3390/s22165941.
- [11] L. H. Ning, "Identifying distinct latent classes of pitch-shift response consistency: Evidence from manipulating the predictability of shift direction," *Frontiers in Psychology*, vol. 13, pp. 1–17, 2022, doi: 10.3389/fpsyg.2022.1058080.
- [12] S. Rosenzweig, S. Schwar, J. Driedger, and M. Muller, "Adaptive pitch-shifting with applications to intonation adjustment in a cappella recordings," *2021 24th International Conference on Digital Audio Effects (DAFx)*, pp. 121–128, 2021, doi: 10.23919/DAFx51585.2021.9768268.
- [13] 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, 2020, doi: 10.1155/2020/8927031.
- [14] M. F. M. Esa, N. H. Mustafa, H. Omar, N. H. M Radzi, and R. Sallehuddin, "Learning convolution neural network with shift pitching based data augmentation for vibration analysis," *IOP Conference Series: Materials Science and Engineering*, vol. 864, no. 1, 2020, doi: 10.1088/1757-899X/864/1/012086.
- [15] H. C. Chu, Y. L. Zhang, and H. C. Chiang, "A CNN sound classification mechanism using data augmentation," *Sensors*, vol. 23, no. 15, 2023, doi: 10.3390/s23156972.
- [16] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, no. 1, 2021, doi: 10.1016/j.jjime.2020.100004.
- [17] L. Jin, F. Tan, and S. Jiang, "Generative adversarial network technologies and applications in computer vision," *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, 2020, doi: 10.1155/2020/1459107.
- [18] H. T. Adityawan, O. Farroq, S. Santosa, H. M. M. Islam, M. K. Sarker, and D. R. I. M. Setiadi, "Butterflies recognition using enhanced transfer learning and data augmentation," *Journal of Computing Theories and Applications*, vol. 1, no. 2, pp. 115–128, Nov. 2023, doi: 10.33633/jcta.v1i2.9443.
- [19] A. R. -Esparza, M. I. C. -Murguia, J. A. R. -Quintana, and C. A. -Quintana, "Leukocyte recognition using a modified AlexNet and image to image GAN data augmentation," *Mexican Conference on Pattern Recognition*, 2023, pp. 139–148. doi: 10.1007/978-3-031-33783-3\_13.
- [20] C. Dewi, R. C. Chen, Y. T. Liu, and H. Yu, "Various generative adversarial networks model for synthetic prohibitory sign image generation," *Applied Sciences*, vol. 11, no. 7, 2021, doi: 10.3390/app11072913.
- [21] F. Li, J. Zheng, and Y. fang Zhang, "Generative adversarial network for low-light image enhancement," *IET Image Processing*, vol. 15, no. 7, pp. 1542–1552, 2021, doi: 10.1049/ipr2.12124.
- [22] B. Pan and W. Zheng, "Emotion recognition based on EEG using generative adversarial nets and convolutional neural network," *Computational and Mathematical Methods in Medicine*, vol. 2021, pp. 1–11, 2021, doi: 10.1155/2021/2520394.
- [23] J.-Y. Baek, "Enhanced speech emotion recognition using DCGAN-based data augmentation," *Electronics*, vol. 12, no. 18, 2023, doi: 10.3390/electronics12183966.
- [24] N. Jia and C. Zheng, "Emotion speech synthesis method based on multi-channel time–frequency domain generative adversarial networks (MC-TFD GANS) and mixup," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 1749–1762, 2022, doi: 10.1007/s13369-021-06090-9.
- [25] B. Mouaz, B. H. Abderrahim, and E. Abdelmajid, "Speech recognition of Moroccan dialect using hidden Markov models," *Procedia Computer Science*, vol. 151, pp. 985–991, 2019, doi: 10.1016/j.procs.2019.04.138.
- [26] E. Gelvez-Almeida, A. Vásquez-Coronel, R. Guatelli, V. Aubin, and M. Mora, "Classification of Parkinson's disease patients based on spectrogram using local binary pattern descriptors," *Journal of Physics: Conference Series*, vol. 2153, no. 1, 2022, doi: 10.1088/1742-6596/2153/1/012014.
- [27] Y. L. Xu K. Yao J, "Improved convolutional neural network and spectrogram image feature for traffic sound event classification," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2023, doi: 10.1177/09544070231189910.
- [28] Z. J. Xu, R. F. Wang, J. Wang, and D. H. Yu, "Parkinson's Disease Detection Based on Spectrogram-Deep Convolutional Generative Adversarial Network Sample Augmentation," *IEEE Access*, vol. 8, pp. 206888–206900, 2020, doi: 10.1109/ACCESS.2020.3037775.

## BIOGRAPHIES OF AUTHORS






**Kristiawan Nugroho**    works as a lecturer at the Faculty of Information Technology and industry, Stikubank University. He obtained a bachelor's degree in 2001 in the Department of Information Systems, Faculty of Computer Science, Dian Nuswantoro University, then in 2007 he obtained a master's degree in informatics engineering, Dian Nuswantoro University. He also obtained a 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 research in data mining, machine learning, speech recognition, and sentiment analysis. He can be contacted at email: kristiawan@edu.unisbank.ac.id.






**Kristophorus Hadiono**    is employed as a lecturer at Universitas Stikubank's Faculty of Information Technology and Industry. In 2001, he graduated from Universitas Stikubank's Faculty of Information Technology with a bachelor's degree in information systems. Then, in 2010, he graduated from Gadjah Mada University with a master's degree in computer science. In 2011, he enrolled in Assumption University of Thailand to pursue a doctorate in information technology, which he completed in 2016. In addition, he carried out a number of studies in statistical science, data mining, and text processing. He can be contacted at email: kristophorus.hadiono@edu.unisbank.ac.id.






**Felix Andreas Sutanto**    is employed as a lecturer at Universitas Stikubank's Faculty of Information Technology and Industry. In 2005, he graduated from Universitas Stikubank's Faculty of Information Technology with a bachelor's degree in information systems. Then, in 2010, he graduated from Gadjah Mada University with a master's degree in computer science. In 2014, he has conducted various research in computer network and cloud computing. He can be contacted at email: felix@edu.unisbank.ac.id.



**Dhendra Marutho**    was born in Semarang, Indonesia on 27 March 1981, He is received his Diploma in Mechanical Engineering form Diponegoro Universtiy in 2003. He then went on to earn a bachelor's degree in computer engineering from STMIK Provisi, Semarang, in 2010, followed by a master's degree in computer science from Dian Nuswantoro University, Semarang, in 2019. He is currently working towards his Ph.D. in Computer Science at Dian Nuswantoro University. Concurrently, he is a lecturer at Muhammadiyah University, Semarang. His research interest is artificial intelligence, machine learning, data mining, computer vision, and image processing. He can be contacted at email: dhendra@valudata.net.



**Prof. 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 Ph.D. at Loughborough University, UK, and a one-year UKIERI postdoctoral fellowship in 2007-08. His broad area of research interest is signal processing with a specialization in speech recognition. He has authored/coauthored over 250 papers in refereed academic journals & conference proceedings and steered 9 researchers to PhD graduation. He is a Senior Member, Institute of Electrical and Electronics Engineers, (IEEE, USA). He can be contacted at email: omar.farooq@amu.ac.in.