A Robust Gender Recognition System using Convolutional Neural Network on Indonesian Speaker

I Nyoman Switrayana*, Sirojul Hadi, Neny Sulistianingsih
Faculty of Engineering, Bumigora University
Ismail Marzuki street no. 22, Mataram, Nusa Tenggara Barat, Indonesia
*e-mail: nyoman.switrayana@universitasbumigora.ac.id

(***received***: 4 December 2023, ***revised***: 22 April 2024, ***accepted***: 22 April 2024)

Abstract

Voice is one of the biometrics that humans have. Humans can be recognized by the sounds produced by their vocal cords and vocal tracts. One of the uses of voice is to recognize gender. Despite extensive research, gender recognition using machine learning remains unsatisfactory due to the complexity of voice features and the limitations of conventional algorithms. In this research, voice-based gender recognition is performed by applying deep learning. The deep learning model used is the Convolutional Neural Network (CNN). The input of CNN is the result of feature extraction from the Mel-Frequency Cepstral Coefficients (MFCC) method. MFCC produces Mel-Spectograms which are important features of sound. The dataset used is Indonesian speech. In the research, there are imbalanced and balanced dataset scenarios to see the performance of the model. To produce a balanced dataset, random undersampling is performed on the majority class. In addition, the effect of dividing training and testing data with a composition of 70:30, 80:20, and 90:10 was observed. The results show that the model has 100% accuracy for all imbalanced dataset scenarios. Then the highest accuracy is 99.65% for the balanced dataset scenario with 70:30 splitting. In summary, it can be concluded that CNN performs very well in identifying gender from voice features overall, although its performance decreases when random undersampling is applied to the dataset.

Keywords: Biometric, convolutional neural network (CNN), deep learning, gender recognition, mel-frequency cepstral coefficients (MFCC), voice

1 Introduction

Voice recognition is one of the topical issues in Natural Language Processing (NLP) [1] and Signal Processing [2]. Recognition is done through voice input. The human voice contains a lot of data and information that can be used to determine habits, age, gender, and emotion [3]. From the voice, it can also be recognized who is speaking, language, accents, and even diseases suffered. Therefore, voice is one of the biometrics that humans have apart from fingerprints, iris, face, and palm [4]. The topic of the problem solved in this research is gender identification.

Human gender can be identified from voice signal processing [2]. Gender recognition is to recognize the voice signal produced by a man or woman. Voice signals are the result of a combination of voiced and unvoiced signals [5]. The periodic production of voice signals is highly dependent on the vocal cord and vocal tract of each human being. It is because of the uniqueness of the vocal cord and vocal tract that voice signals can be distinguished from one another. Voice signals produced by men or women can generally be distinguished easily by human hearing. Technology development is oriented towards the concept of user personalization so voice recognition systems are needed [6]. With gender identification, the solution space can be minimized for speaker recognition and personalizing technology services. Another utilization is in the criminal search sector, whether the perpetrator is male or female. Generally, these crimes are committed through voice messages or phone calls. Other applications of gender identification can be the classification of treatment in the medical sector based on gender, advertising and marketing strategies, call centers, security systems, and

Customer Relationship and Management Systems [3], [7]. Gender recognition during game play also facilitates or helps players to communicate well with fellow players [8].

In order for the computer to recognize gender, the voice signals need to be represented into a form or format that is easily understood by the computer. The transformation of speech signals is done in signal processing called feature extraction. Feature extraction is the stage of transforming the voice signal into a feature vector. This feature vector is then used to train the model. The learning model is tasked with extracting the most important information patterns in the feature vector so that male or female voice characters can be distinguished. A better representation of voice features causes the learning results of a model to be maximized [9]. Based on the analysis of the comparison of temporal and spectral features by [10] to recognize gender, temporal features are not suitable for recognizing gender. Spectral features are more recommended for use in recognition tasks. Then, the results of spectral feature analysis conducted by [11] for accent recognition show Mel Frequency Cepstral Coefficients (MFCC) produces better features than spectogram, chromagram, and spectral centroid. CNN here gives good results for recognizing accent. Besides being used to recognize gender, voice is also used to recognize age by [12]. It is also explained that MFCC coefficients are the most contributing features in recognizing gender and age. Therefore, MFCC is used as a voice feature extraction technique in this study.

Gender recognition in Bengali language [1] using 20 processed voice features from MFCC and 6 other features (chroma feature, Root Mean Square Error, Spectral Centroid, spectral bandwidth, roll off, and zero crossing rate) resulted in the highest accuracy of 99.13% by gradient boosting algorithm. Observation and analysis of classifiers to perform gender identification from speech signals [5], [7], [13], [14] show the performance of Support Vector Machine (SVM) outperforms other machine learning models. However, gender recognition conducted by [2] on Indonesian language shows that Artificial Neural Network has the highest performance of KNN and SVM with 93.07% accuracy. Javanese gender recognition [15] using Deep Learning (deep neural network) gave the highest accuracy than logistic regression and SVM. Another javanese language gender recognition was done by [16]. In that study, Backpropagation Neural Network (BPNN) was proposed as the learning model. BPNN gives 95% accuracy and for Indonesian word speech the performance of BPNN is still not good enough.

Gender and emotion recognition [17] resulted in accuracy below 85%, the proposed Convolutional Neural Network (CNN) consists of a convolution layer, batch normalization layer, and max-pooling layer. Then with the reshaping technique of 1D dimensional signal data into 3D as CNN input on Arabic gender identification [18], it is able to give 98.91% accuracy. This reshaping technique gives 0.4% improvement with 1D to 3D signal preprocessing as CNN input. Another deep learning architecture is Bidirectional Long Short-Term Memory (BLSTM) with a division of training and testing datasets of 80:20 resulting in the highest accuracy of 90.5% [19].

Based on previous research, the use of deep learning is more effective and promising for recognizing gender with a high level of performance. Therefore, it is necessary to further investigate how the performance of deep learning architecture, especially CNN. In previous studies, there were still few researchers who conducted research using speech data or voice signals whose speakers were speaking Indonesian. Most studies use speech in English, Arabic, France, Spanish, Mandarin [11]. In this study, it is investigated how the performance of the proposed CNN model if the recorded voice used is in Indonesian. The dataset used is obtained from common voice (Mozilla repository) which provides voice recordings of native speakers of various languages. In previous studies, no one has tested the reliability of the system in a dataset with an imbalanced number of labels (imbalanced dataset). So here it will be tested against the state of the number of datasets. CNN architecture is proposed in this study because of CNN's excellent image processing capabilities and speed in classification with high accuracy [20]. The image in question is the MFCC coefficients represented in the mel-spectrogram.

2 Literature Review

The investigated research by [21] focused on utilizing machine learning algorithms for gender identification based on historical speech datasets. Through the application of feature selection algorithms like the Fisher Score Algorithm and tree-based methods, the study assessed various acoustic features, consistently identifying the fundamental frequency (F0) as the most influential factor in gender identification. The paper explored dataset classification using Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree algorithms. The study compared the accuracy of all models and Decision Tree demonstrating superior accuracy. The Extra Tree classifier corroborated the significance of F0 and F4 as the most predictive parameters for gender identification. The study by [22] discussed machine learning algorithms, specifically the KNN algorithm and SVM. KNN estimates class probabilities using various distance metrics. The comparison highlights the importance of distance metrics of KNN. The Euclidean and City Block distance functions achieved the highest accuracy in KNN. In this case, KNN demonstrated better accuracy than SVM.

In the work of [23] explored the use of Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN) for gender detection and speaker identification tasks. Using Mel-frequency cepstral coefficients (MFCC) features and different normalization techniques, the models utilizing z-score and Gramian matrix transformation were noted to exhibit better performance compared to those that employed max-min normalization of MFCC. MLP demonstrated better generalization errors compared to CNN, but it took more time to converge. The proposed age and gender recognition [24] method utilizes a CNN with a Multi-Attention Module (MAM) for efficient analysis of speech spectrograms. The MAM selectively focuses on crucial information, improving accuracy in determining speakers' age and gender. The framework includes three CNN models for age, gender, and age-gender classification. The model addresses challenges in automatic gender and age identification, employing a lightweight attention mechanism for relevant feature capture. It demonstrates high precision, recall, and F1-score values, showcasing proficiency in age and gender classification.

The research conducted by [25] presents a comprehensive exploration of gender identification in speech processing using statistical features of pitch. By proposing a novel feature set, PFG (Pitch Feature for Gender), and evaluating its effectiveness across three diverse speech corpora, the study demonstrates the significance of statistical pitch features in speaker-dependent speech processing tasks. Several methods, including CNN, MLP, SVM, and LR, demonstrated high accuracy in various experiments conducted with the TIMIT and CHAINS datasets. The SLR-63 and Malayalam corpus achieved the maximum accuracy of 99.01%, particularly with the CNN model. Overall, the paper underscores the importance of statistical pitch features and their potential in achieving high accuracy in gender recognition from human voice.

The examined research [26] presented a comprehensive methodology for gender identification in speech samples, leveraging machine learning and CNN. The study introduced a system designed to improve gender estimation precision by extracting fundamental frequency and MFCC features. Various machine learning methods and CNN were employed. The CNN architecture, encompassing a 1D convolutional layer, pooling layer, fully connected layer, and activation functions, was thoroughly explained. The effectiveness and superiority of CNN in gender identification from speech have been proven to be exceptionally high. The application of the ResNet50 model in deep neural networks for gender recognition was conducted by[27]. Comparative analysis with traditional approaches highlighted ResNet50's superior performance. The study emphasized speech classifier effectiveness across various datasets. Addressing model generalization, ResNet50 demonstrated state-of-the-art performance on the Mozilla dataset and satisfactory results on additional datasets. The research underscored CNN models' superiority, achieving an impressive 98.57% accuracy in gender recognition without manual feature extraction. The success of the CNN model served as the primary foundation for its application in this research. The superior capability of CNN in gender recognition significantly contributed to the findings of this study.

3 Research Method

The research methodology used outlines four stages. The first stage is the initial processing of the dataset that has been collected. The second stage is one of the most important stages, namely performing feature extraction. Feature extraction aims to represent sound signal features into feature vectors. Then the third stage is the design or modeling stage. The model used is Convolutional Neural Network (CNN). Model training on the dataset is carried out at this stage. And the last is the testing or evaluation stage of the model training results. These stages can be shown in Figure 1.

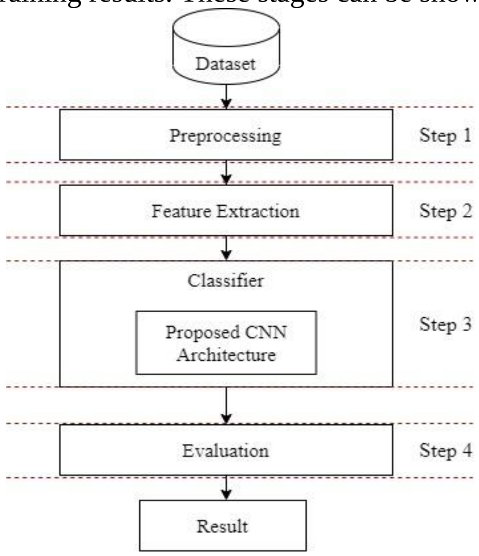

Figure 1. Stages of the proposed research methodology for gender identification

3.1. Dataset

The dataset used to identify gender is an open-source dataset that can be obtained from common voice. The available datasets are intended to train machine learning models based on voice technology. With the collection of datasets that are not only in one language (multi-language), it is expected to develop technology that utilizes voice input in various countries. The data collected is the result of individuals or communities volunteering to record their voice and add sentences to the recording. Voice recording files are saved in .mp3 format. Voice recording and sentence input contributions are made on the website. Therefore, the recording environment can vary from one speaker to another. In this case, the recording environment includes the device or tool used for recording, the situation, and conditions at the time of recording.

The dataset used is the Common Voice Delta Segment 12.0 version. There are 6 hours of recording of which 2 hours of recording data has been validated. The number of speakers is 37 speakers. In addition to the recording files, in the dataset there are several files that describe the dataset. Each recording identity has information on the sentence spoken, accents, gender, age, and others. There are 5514 recording files. And after checking again on the dataset description, not all data has a description.

3.2. Preprocessing

Preprocessing at this stage is done to improve the quality of the dataset. The quality of the dataset is very influential in the performance of the model. There are some data in the dataset in the recording file and the description is still not synchronized. Some data also does not have a description. So that preprocessing needs to be done. To solve this problem, the data taken is only data that has a description and has been validated. The process carried out is filtering by matching the name of the record file with the description data and taking data that is not null or has been validated for gender. From 5514 record files, 4340 records were selected for use. The 4340 records data consists of 78% male (3385 record files with male gender) and the rest are female gender records (955 records).

This uneven number of records causes a case of imbalanced dataset. Therefore, there are two scenarios of data usage in the research. The first scenario is to use the 4340 datasets with the imbalanced dataset problem. The second scenario is to use balanced data. Where the female data totaling 955 becomes a reference in taking the dataset portion of the male record. A random undersampling technique of the majority class (male class) is performed to obtain a balanced dataset. ecords is randomly taken as many as 955. So that from here the data no longer experiences imbalanced dataset cases. These two scenarios are proposed to see the reliability of the model. In the preprocessing stage, the recording files are also transformed into sound signals so that they can be processed at a later stage.

3.3. Feature Extraction

The feature extraction stage is one of the most crucial stages. Because in this stage feature extraction determines how the performance of the learning results of a model. Feature extraction is the stage of representing previous speech signals in feature vectors. This stage involves signal processing techniques. There are several feature extraction techniques including Gammatone Cepstral Coefficients (GTCC), Spectral Entropy, Harmonic Ratio, Mel-Frequency Cepstral Coefficients (MFCC), and others [28]. Features that can also be extracted in signals are acoustic features [7, 13]. Acoustic features of sound such as pitch, median, frequency, etc. [29].

Features are divided into time domain features (amplitude, energy, variance, average, etc.) and frequency domain features (power spectral density, mean power, asymmetric features). The feature extraction method adopted in this research is MFCC. MFCC is one of the methods for extracting spectral features[6]. MFCC is one of the state of the art methods for extracting information on signals. The performance of MFCC with other extraction methods has been tested[11] and the results show MFCC outperforms other methods. From[1, 5, 19] and several other studies also use MFCC as a feature extraction method. The stages of MFCC[2, 6] can be seen in Figure 2.

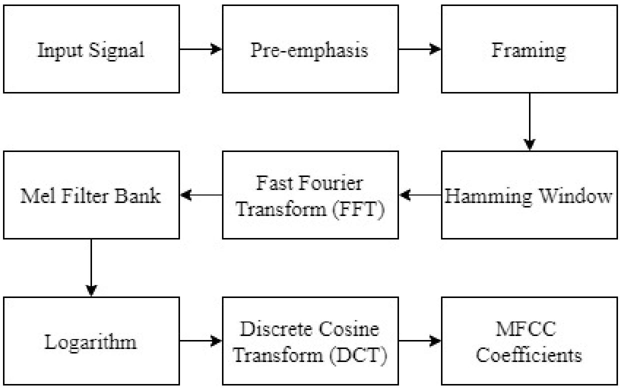

Figure 2. The flow of feature extraction by MFCC

MFCC extracts features by considering human perception and sensitivity to frequency. The following are the stages of MFCC [14, 19]:

1) Pre-emphasis: the process of signal amplification by which the signal is applied using a high pass filter. The result is to pick up the high frequencies.
2) Framing: the input signal will be segmented into signal frames with a standard frame size of 25 ms. The segmentation process also pays attention to the overlap between frames, which is generally 15 ms. The standard frame size and overlap here can be one of the parameters

according to the state of the input signal.

3) Hamming window: the process of windowing all the frames of the framing result. What is done in this process is to apply the signal with the function in equation 1 (multiplication). This aims to capture all the signal features and reduce signal leakage. W(n) for hamming window. The signal in the frame is denoted by S(n), n = 0 to n = N-1.

$$W(n,a) = (1-a) - a\cos\left(\frac{2\pi n}{N-1}\right), \quad where\ 0 \le n \le N-1 \tag{1}$$

4) Fast Fourier Transform (FFT): the process of converting a signal that is in the time domain into the frequency domain. The mathematical formula for the transformation used is in

$$z_i(k) = \sum_{n=1}^{N} S_i(n)h(n)e^{\frac{-2\pi}{N}} \tag{2}$$

Where $S_i(n)$ is the signal in the time domain, $z_i(k)$ is the signal in the frequency domain, $h(n)$ is the window with N samples long, and k is the length of the FFT.

5) Mel-Filter Bank: the process of converting frequency into a Mel-scale then takes energy (features) by applying a triangular filter band. The process of converting the frequency into a Mel-scale has been completed using equation 3.

$$f_{mel} = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

The formation of the filter bank is constructed from the lowest frequency to the highest frequency with N filters. The Mel-frequencies are then mapped to their respective filter banks based on their frequency.

6) Logarithm: the Mel-frequency values in the filter bank are applied with the natural logarithm function. This is in reference to the human auditory system, which operates on a logarithmic frequency scale.

7) Discrete Cosine Transform (DCT): The process of transforming signal from the frequency domain to the time domain. DCT is performed using equation 4.

$$C_m = \sum_{k=1}^{N} \cos\left(m(k-0.5)\frac{\pi}{N}\right) E_k \tag{4}$$

where m = 1, 2, ..., 12. N represents the number of triangular filter banks (20). E_k denotes the logarithmic result.

8) MFCC coefficients: the output of MFCC is used as a feature in model training. This output is represented in a spectrogram form, which is commonly referred to as a Mel-spectrogram

3.4. Classifier

The proposed deep learning model in the research is a Convolutional Neural Network (CNN). The convolutional layer is a crucial component in the CNN architecture [18]. The proposed CNN architecture in the study consists of two convolutional layers and two max-pooling layers, followed by fully connected layers. The layers are arranged sequentially, starting with the input layer, followed by the convolutional layer and the max-pooling layer. After the max-pooling layer, a dropout layer (0.5) is added. The convolutional layers are used to learn the input and generate feature maps. The max-pooling layers extract the most important features from the convolution results, reducing the size of the feature maps.

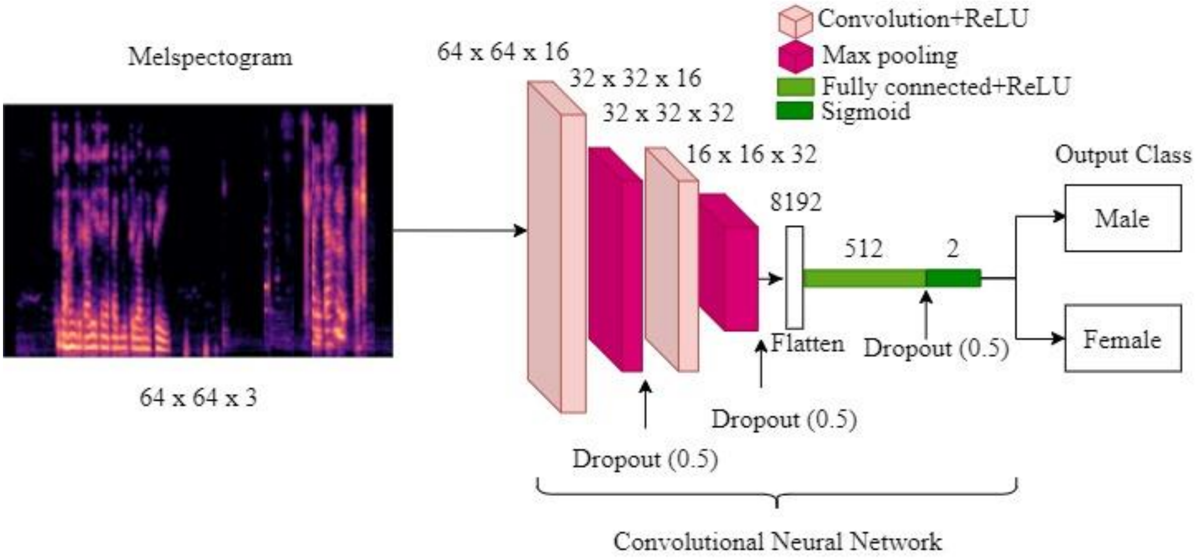

Figure 3. The proposed CNN architecture

Dropout is employed as a regularization technique to prevent overfitting and accelerate the learning process. Dropout randomly deactivates/drops out a certain number of neurons during training, thereby enhancing the model's generalization ability. It has been proven that the addition of dropout layers is highly beneficial in preventing overfitting in the proposed Deep Neural Network (Multilayer Perceptron) [30].

To adjust the output shape of the CNN's learning results to the next layer, a flattened layer is required. The subsequent layers consist of two fully connected layers with relu and sigmoid activation functions. The target class is binary (male/female). The input data trained on the CNN is a spectrogram with a size of 64x64x3. The excellent capabilities of CNN in working with image data are utilized in this study. The proposed CNN architecture is trained for 50 epochs using the Adam optimizer. The loss function for binary classification is binary cross entropy. The architecture of the proposed CNN can be seen in Figure 3. The input to the CNN consists of augmented Mel-spectrogram images. Data augmentation techniques used include rescaling, random flipping, and random rotation. According to a study [31] that utilized Long Short-Term Memory (LSTM) for age and gender recognition, overfitting issues were encountered. Data augmentation applied at the signal level, such as noise adding, time stretch, shifting, and pitch shift, proved effective in reducing the overfitting problem. In contrast, in this study, data augmentation is performed at the output level (Mel-spectrogram images).

3.5. Evaluation

The training and testing process of the CNN model receives input data in both imbalanced and balanced datasets. Then, from these two data conditions, the data is split into training and testing sets using proportions of 70:30, 80:20, and 90:20, respectively. The experimental scenarios are shown in Table 1. CNN_1 represents the Convolutional Neural Network trained and tested with a balanced dataset (955 recordings for both male and female data). CNN_2 represents the Convolutional Neural Network trained and tested with an imbalanced dataset (3385 male recordings and 955 female recordings).

Table 1. Experimental Scenarios

CNN_1	CNN_2
Balanced Dataset	Imbalanced dataset

Data Splitting (70:30)	Data Splitting (80:20)	Data Splitting (90:10)	Data Splitting (70:30)	Data Splitting (80:20)	Data Splitting (90:10)

The calculation of each evaluation metric is calculated using the following mathematical formula:

$$\text{Precision } = \frac{TP}{TP+FP} \tag{5}$$

$$\text{Recall } = \frac{TP}{TP+FN} \tag{6}$$

$$\text{F1} - \text{Score } = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \tag{7}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{8}$$

Precision is used to measure the proportion of classes that are predicted to be true positive with positive prediction results. The recall represents the proportion of positive class that is correctly predicted for all positive class data sample prediction results. Accuracy calculates the number of correct prediction results for all classes against all data. Then F1-Score takes the harmonic mean of precision and recall. The use of the F1-Score is one of them for imbalanced datasets to produce an evaluation measure that represents model performance.

4 Results and Analysis

Gender identification can be conducted due to the differences in the vocal signals produced between males and females. Figure 4 and 5 illustrates the vocal signals of a female and a male. These vocal signals are the result of uttering the sentence "Aku baru saja digigit nyamuk" (I was just bitten by a mosquito). From the figures, it can be observed that the vocal signal produced by females has different amplitudes and signal lengths. When compared, the amplitude of the female voice is higher, and the signal is longer. These differences can result in the extraction of different features from the vocal signal. Figure 6 and 7 displays the Mel spectrograms of each signal. The variations in the Mel spectrograms demonstrate distinct distributions of MFCC coefficients. This is due to the different vocal characteristics of each gender, even when uttering the same sentence.

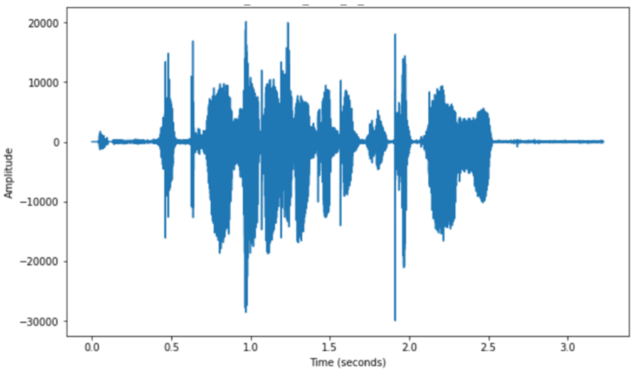

Figure 4. Female voice signal in time domain

ISSN:2302-8149
e-ISSN:2540-9719

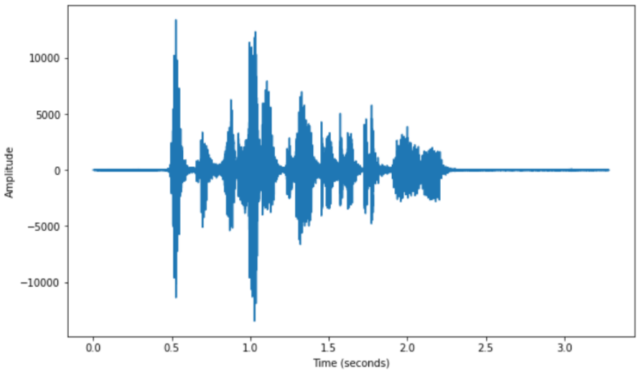

Figure 5. Male voice signal in time domain

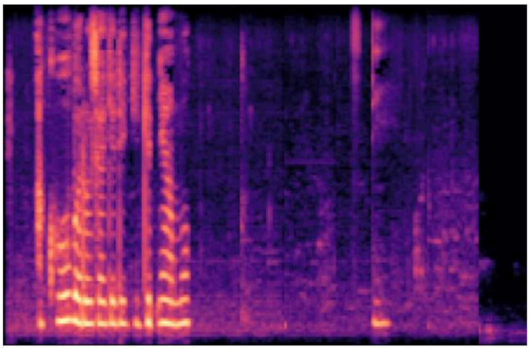

Figure 6. The mel-spectrogram of female voice signal

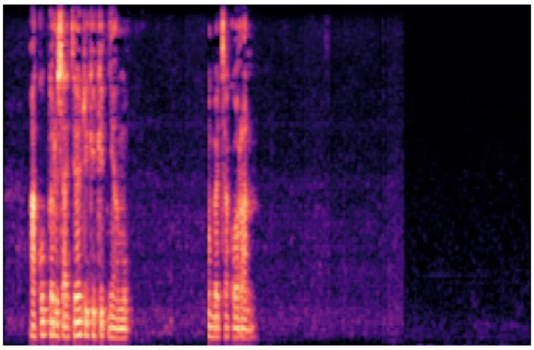

Figure 7. The mel-spectrogram of male voice signal

The model was trained and tested based on the previously described experimental scenario, interesting evaluation results were obtained. Table 2, 3, and 4 shows the evaluation results of all experimental scenarios. In the imbalanced dataset scenario, where the majority class was male, CNN achieved 100% precision, recall, F1-Score, and accuracy. The proportion of dataset separation (70:30, 80:20, and 90:20) did not affect the model's performance. This demonstrated that CNN was highly effective in identifying gender when one class had a much larger sample size than the other class. However, in the balanced dataset scenario where random undersampling was performed on the majority class, a decrease in performance was observed.

Table 2. CNN Evaluation Results Balanced And Imbalanced Dataset In 70:30 Data Split

			Data Splitting 70:30		
	Class	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)

CNN_1	Female	99.32	100	99.66	**99.65**
	Male	100	99.32	99.66	
CNN_2	Female	100	100	100	**100**
	male	100	100	100	

Table 3. **CNN Evaluation Results Balanced and Imbalanced Dataset In 80:20 Data Split**

		Data Splitting 80:20			
	Class	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
CNN_1	Female	99.03	100	99.51	**99.48**
	Male	100	98.87	99.43	
CNN_2	Female	100	100	100	**100**
	male	100	100	100	

ion Results Balanced and Imbalanced Dataset In 90:10 Data Split

		Data Splitting 90:10			
	Class	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
CNN_1	Female	89.11	100	99.05	**98.95**
	Male	100	97.73	98.43	
CNN_2	Female	100	100	100	**100**
	male	100	100	100	

This indicates that the reduced number of male samples in the dataset poses a challenge for the model in accurately identifying gender, resulting in decreased performance. Misclassifications of the male class in the balanced dataset are the primary cause of the model's performance decline. The model with a 70:30 composition in the balanced dataset achieves the highest accuracy, specifically 99.65%. Figure 8, 9, and 10 illustrates the confusion matrix derived from the CNN model's results on the balanced dataset, with a 70:30, 80:20, and 90:10 training and testing data split. This phenomenon could be attributed to the model's limited ability to recognize patterns associated with male gender due to the loss of valuable information contained in male samples during the random undersampling process. Additionally, it could also be influenced by the selection of insufficiently representative data within the male class during both the training and testing phases.

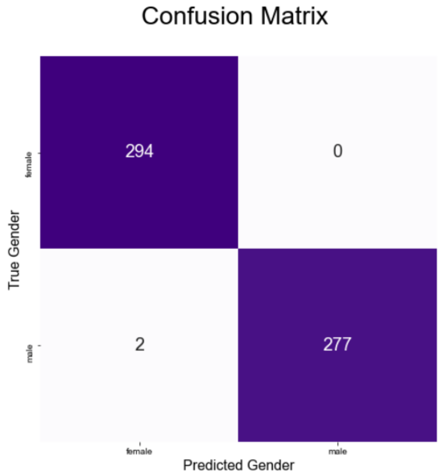

Figure 8 . Confusion matrix balanced dataset with 70:30 data split

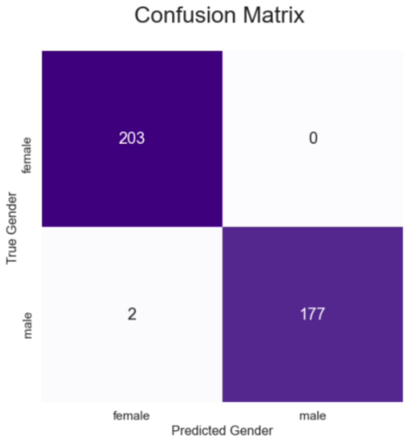

Figure 9. Confusion matrix balanced dataset with 80:20 data split

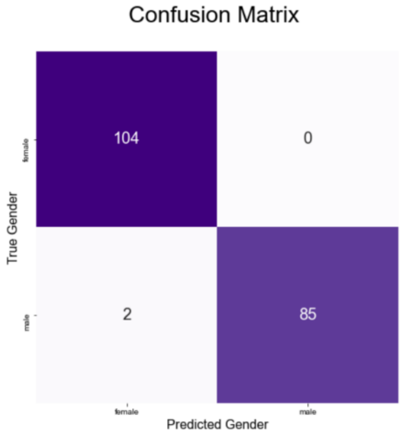

Figure 10. Confusion matrix balanced dataset with 70:30 data split

When observed in more detail, the number of gender prediction errors in the case of a balanced dataset in each data split is two. These are instances where the label should have been "male" but was classified as "female." This same number of prediction errors is what leads to a non-significant decrease in accuracy. Consequently, the largest testing data split benefits the most by achieving a higher accuracy. However, overall, the CNN model demonstrates excellent performance in gender identification tasks using both imbalanced and balanced datasets. It is important to note that when balancing the dataset, there may be a decrease in performance.

In this study, a performance decrease was observed due to the undersampling of the majority class. The accuracy dropped in the balanced dataset scenarios based on data splitting proportions is depicted in Figure 11. In Figure 11, the observed performance metric is accuracy. The decrease in accuracy caused by the dataset-splitting proportions is not significantly substantial. This indicates that different dataset-splitting scenarios also have an impact on the model's performance in the balanced dataset resulting from random undersampling. However, upon closer examination, the accuracy decrease does not exceed 1% when compared to all accuracy results in the balanced dataset. Thus, the proposed CNN architecture is highly robust in this study.

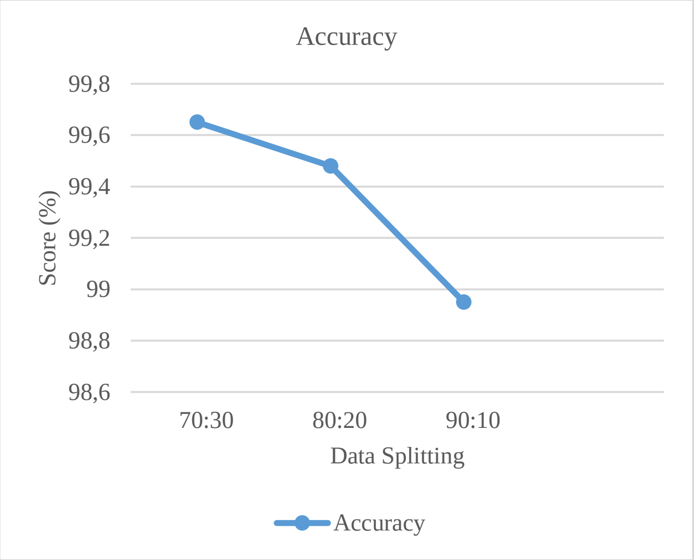

Figure 11. Model accuracy on the balanced dataset

This research corroborates the findings of prior studies [15] which have consistently affirmed the superiority of deep learning over conventional machine learning approaches. The research conducted by [18] implemented an efficient CNN and achieved remarkably high accuracy. In this work, this is substantiated by demonstrating a higher level of accuracy and the more complex learning capabilities of CNNs. Table 5 shows the comparative results of related studies.

Table 5. C rison results of this work with some previous related works

Research by	Research methods	Language	Accuracy
K. Nugroho, et.al [15]	Deep Learning - Artificial Neural Network (ANN)	Java (local language)	97.78%
A. M. Jasim, et.al [18]	CNN	Arabic	98.91%
This work	**CNN**	**Indonesian**	**100%**

5 Conclusion

Based on the experimental results, CNN performs exceptionally well in recognizing male or female voice features. CNN can identify gender with high accuracy even in datasets with class imbalances. The model can recognize patterns or characteristics specific to male and female voices. However, when the dataset is balanced using random undersampling, there is a decrease in performance in the majority class (males). Therefore, it is necessary to explore dataset-balancing methods that can provide more optimal performance while maintaining accuracy in both categories. In future research, it is essential to explore dataset-balancing techniques such as oversampling, generative-based balancing, and others that can preserve important information in the majority class while maintaining accuracy in the minority class. Furthermore, to enhance the information on male or female voice characteristics, it may be worth considering combining the features of MFCC with other available features in the audio signal. Given the large feature dimensions, it is advisable to combine learning models as suggested in [9, 13]. Additionally, training the model with multiple languages [28] can further enrich the information and broaden its applicability.

Reference

[1] S. M. S. I. Badhon, M. H. Rahaman, and F. R. Rupon, “A Machine Learning Approach to Automating Bengali Voice based Gender Classification,” Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2019, pp. 55–61, 2020, doi: 10.1109/SMART46866.2019.9117385.

[2] E. Tanuar, E. Abdurachman, F. L. Gaol, and Lukas, “Analysis of Gender Identification in Bahasa Indonesia using Supervised Machine Learning Algorithm,” 2020 3rd Int. Conf. Inf. Commun. Technol. ICOIACT 2020, pp. 421–424, 2020, doi: 10.1109/ICOIACT50329.2020.9332145.

[3] M. A. Uddin, M. Biswas, and R. K. Pathan, “Gender Recognition from Human Voice using Multi-Layer Architecture,” 2020.

[4] S. Katiyar, S. Kumar, and H. Walia, “A Novel Approach to Identify Age and Gender using Deep Learning,” 2021 9th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir. ICRITO 2021, pp. 1–5, 2021, doi: 10.1109/ICRITO51393.2021.9596153.

[5] A. Singhal and D. K. Sharma, “Analysis of Classifiers for Gender Identification using Voice Signals,” 2021 5th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2021, pp. 2021–2024, 2021, doi: 10.1109/ISCON52037.2021.9702469.

[6] M. La Mura and P. Lamberti, “Human-Machine Interaction Personalization : a Review on Gender and Emotion Recognition Through Speech Analysis,” pp. 319–323, 2020.

[7] T. A. Topu, S. Siddique, A. K. M. Masum, S. A. Khushbu, S. M. S. I. Badhon, and S. Abujar, “Bengali Continuous Speech Voice-based Gender Classification,” 2021 12th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2021, 2021, doi: 10.1109/ICCCNT51525.2021.9579838.

[8] B. Fatima, A. Raheel, A. Arsalan, M. Majid, M. Ehatisham-ul-haq, and S. M. Anwar, “Gender Recognition using EEG during Mobile Game Play,” pp. 634–639, 2021.

[9] A. I. Ahmed, D. I. Ndzi, J. Chiverton, and M. Al-faris, “Machine Learning based Speaker
sing Transformed Features,” pp. 13–18, 2021.

[10] E. Priya, J. Priyadharshini .S, P. Satya Reshma, and S. .S, “Temporal and Spectral Features based Gender Recognition from Audio Signals,” 2022.

[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] S. Goyal, V. V. Patage, and S. Tiwari, “Gender and Age Group Predictions from Speech Features using Multi-Layer Perceptron Model,” 2020 IEEE 17th India Counc. Int. Conf. INDICON 2020, pp. 3–8, 2020, doi: 10.1109/INDICON49873.2020.9342434.

[13] G. Sharma and S. Mala, “Framework for Gender Recognition using Voice,” Proc. Conflu. 2020 - 10th Int. Conf. Cloud Comput. Data Sci. Eng., pp. 32–37, 2020, doi: 10.1109/Confluence47617.2020.9058146.

[14] S. Chaudhary and D. Kumar Sharma, “Gender Identification based on Voice Signal Characteristics,” pp. 869–874, 2018.

[15] K. Nugroho, E. Noersasongko, and H. A. Santoso, “Javanese Gender Speech Recognition using Deep Learning and Singular Value Decomposition,” pp. 251–254, 2019.

[16] L L. M. Liztio and C. A. Sari, “Gender Identification based on Speech Recognition using Backpropagation Neural Network,” pp. 88–92, 2020.

[17] P. Sachin, N. Correa, A. H. Shenoy, A. C. Ballal, and P. Mittal, “Gender and Emotion Classification by Hierarchical Modelling using Convolutional Neural Network,” 2022 2nd Asian Conf. Innov. Technol. ASIANCON 2022, pp. 1–6, 2022, doi: 10.1109/ASIANCON55314.2022.9908796.

[18] A. M. Jasim, S. R. Awad, F. L. Malallah, and J. M. Abdul-jabbar, “Efficient Gender Classifier for Arabic Speech using CNN with Dimensional Reshaping,” pp. 1–5, 2021.

[19] R. D. Alamsyah and S. Suyanto, “Speech Gender Classification using Bidirectional Long Short Term Memory,” pp. 646–649, 2023.

[20] K. V. Balaji and R. Sugumar, “A Comprehensive Review of Diabetes Mellitus Exposure and Prediction using Deep Learning Techniques,” 2022 Int. Conf. Data Sci. Agents Artif. Intell. ICDSAAI 2022, no. Ml, 2022, doi: 10.1109/ICDSAAI55433.2022.10028832.

[21] R. Rehman, K. Bordoloi, K. Dutta, N. Borah, and P. Mahanta, “Feature Selection and Classification of Speech Dataset for Gender Identification: A machine Learning Approach,” *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 22, pp. 3449–3459, 2020.

[22] B. Jena, A. Mohanty, and S. K. Mohanty, “Gender Recognition of Speech Signal using KNN and SVM,” *SSRN Electron. J.*, no. Icicnis, pp. 548–557, 2021, doi: 10.2139/ssrn.3769786.

[23] O. Mamyrbayev, A. Toleu, G. Tolegen, and N. Mekebayev, “Neural Architectures for Gender Detection and Speaker Identification,” *Cogent Eng.*, vol. 7, no. 1, 2020, doi: 10.1080/23311916.2020.1727168.

[24] A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, “Age and Gender Recognition using a Convolutional Neural Network with a Specially Designed Multi-attention Module Through Speech Spectrograms,” *Sensors*, vol. 21, no. 17, 2021, doi: 10.3390/s21175892.

[25] G. U. Shagi and S. Aji, “A machine Learning Approach for Gender Identification using Statistical Features of Pitch in Speeches,” *Appl. Acoust.*, vol. 185, p. 108392, 2022, doi: 10.1016/j.apacoust.2021.108392.

[26] H. Q. Jaleel, J. J. Stephan, and S. A. Naji, “Gender Identification from Speech Recognition using Machine Learning Techniques and Convolutional Neural Networks,” *Webology*, vol. 19, no. 1, pp. 1666–1688, 2022, doi: 10.14704/web/v19i1/web19112.

[27] A. A. Alnuaim *et al.*, “Speaker Gender Recognition based on Deep Neural Networks and ResNet50,” *Wirel. Commun. Mob. Comput.*, vol. 2022, 2022, doi: 10.1155/2022/4444388.

[28] A. A. Alashban and Y. A. Alotaibi, “Speaker Gender Classification in Mono-Language and Cross-Language using BLSTM Network,” pp. 66–71, 2021.

[29] M. D. Prasetio, “Single Speaker Recognition using Deep Belief Network Gender Classification Voices,” pp. 253–258, 2019.

and G. Hajela, “Voice Gender Recognizer Recognition of Gender from Voice using Deep Neural Networks,”, pp. 319–324, 2020.

[31] G. R. Nitisara, S. Suyanto, and K. N. Ramadhani, “Speech Age-Gender Classification using Long Short-Term Memory,” pp. 8–11, 2023.