Computer Engineering and Applications Vol. 14, No. 2, June 2025

Implementation of Feature Selection for Optimizing Voice Detection Based on Gender using Random Forest

Abdurahman, Marsella Vindriani, Aditya Putra Perdana Prasetyo*, Sukemi, M. Ali Buchari, Sarmayanta Sembiring, Ricy Firnando, Rahmat Fadli Isnanto, Kemahyanto Exaudi, Aldi Dudifa, Rafki Sahasika Riyuda

Department of Computer Systems, Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia
**aditrecca@gmail.com*

ABSTRACT

Gender-based voice detection is one of the machine learning applications that has various benefits in technology and services, such as virtual assistants, human-machine interaction systems, and voice data analysis. However, the use of too many features, including irrelevant features, can cause a decrease in accuracy and model performance. This research aims to optimize voice-based gender detection by applying a feature selection method to select significant features based on their correlation value to the target. Experimental results show that by using only the significant features selected through correlation analysis, the accuracy of the model is significantly improved compared to using all available features. This research confirms the importance of feature optimization to support the development of more efficient and accurate gender-based speech detection models.

Keywords: Selection, Machine Learning, Audio Recognition, Random Forest, Gender

1. INTRODUCTION

Differentiating the gender of a voice may sound easy if done manually by a human [1]. However, if there are thousands or even millions of voices that need to be classified, it becomes clear that a system capable of automatically classifying human voices is needed. The current case is that voice-based chatbots and virtual assistants still cannot tell the gender of the person they are talking to. However, if it were possible, we would be able to learn the habits or trends that certain genders often talk about. Pattern recognition systems are components that play a crucial role in replicating human sensory abilities, especially in terms of vision and hearing. For example, computers must have a logical mechanism to recognize patterns in the sound being processed to mimic the human sense of hearing [2]. This prompted the author to try a simple idea to recognize sound patterns that computers can recognize well.

To recognize a particular pattern, the main challenge lies in how the data collection process is carried out in order to generate a representative amount of numerical data that is aligned with the available samples. This research seeks to utilize simple techniques to recognize voices and classify them based on gender so that computers can identify them by applying various feature extraction theories to audio data. The main objective of this research is to analyze which features can affect speech

recognition and prove how Random Forest parameters are used to analyze voices based on gender.

2.1 Human Voice

All creatures are created with different traits and characteristics. This difference applies to physical and genetic characteristics as well as the voice produced by each individual. Each individual has a different voice because their vocal characteristics are unique. The most striking difference lies in the size of the vocal anatomy which affects the frequency of the sound produced by each individual. For example, the size difference between men and women affects the pitch of their voices. Male voices tend to have a lower pitch with a frequency of around 120 Hz, while female voices have a higher pitch with a frequency of around 210 Hz [3].

Then, Random Forest is suitable for the case of large and complex datasets. Liaw and Wiener established Random Forest as a learning process and model building in the R programming language. Random Forest belongs to algorithms that function as efficient classification and regression techniques. They highlighted the flexibility of Random Forest that handles datasets that have many variables, and can make calculations about predicting the importance of data types [10].

2.2 Audio Recognition

Audio recognition, refers to technology that enables computers to identify and process human speech, and convert it into text or commands [4]. In gender classification, audio recognition focuses on distinguishing male and female voices by examining specific acoustic characteristics such as pitch, tone and frequency variations. These differences naturally occur between the sexes due to physiological differences in vocal anatomy. In addition, audio recognition has a crucial role in various fields, such as human-robot interaction, surveillance, and industrial automation [5].

In a more specific context, human-robot interaction relies heavily on audio recognition technology to facilitate communication. In industrial environments, robots can be programmed to recognize the gender of workers and adjust the communication style accordingly. In addition, audio recognition technology is currently widely applied to digital assistants, smart home devices, as well as AI-based search portals such as Siri and Alexa that provide support in the form of voice or text commands. Although the technology is still evolving, audio recognition is becoming increasingly accurate and is considered a worthwhile investment for various applications [6].

2.3 Feature Selection

Feature selection is an important process in data analysis and machine learning that aims to select the most relevant subset of features from a set of data. In this context, features refer to the variables used to predict the target. This process not only helps improve model accuracy, but also reduces complexity, speeds up training time, and prevents overfitting. To implement feature selection on a dataset, the first step is to analyze the relationship between the features and the target. One commonly used method is to calculate the correlation coefficient, which gives an idea of how strong

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

the relationship is between two variables [7]. The acoustic features in the data are listed in Table 1 below.

TABLE 1.
Acoustic Features of Gender Dataset [8].

PROPERTY	DESCRIPTION
duration	panjang sinyal
meanfreq	frekuensi rata-rata (dalam kHz)
sd	deviasi standar frekuensi
median	frekuensi median (dalam kHz)
Q25	kuantil pertama (dalam kHz)
Q75	kuantil ketiga (dalam kHz)
IQR	rentang antar kuantil (dalam kHz)
skew	kemiringan
kurt	kurtosis
sp.ent	entropi spektral
sfm	kerataan spektral
mode	frekuensi mode
centroid	pusat frekuensi
peakf	frekuensi puncak
meanfun	rata-rata frekuensi fundamental yang diukur di seluruh sinyal akustik
minfun	frekuensi fundamental minimum yang diukur di seluruh sinyal akustik
maxfun	frekuensi fundamental maksimum yang diukur di seluruh sinyal akustik
meandom	rata-rata frekuensi dominan yang diukur di seluruh sinyal akustik
mindom	minimum frekuensi dominan yang diukur di seluruh sinyal akustik
maxdom	maksimum frekuensi dominan yang diukur di seluruh sinyal akustik
dfrange	rentang frekuensi dominan yang diukur di seluruh sinyal akustik
modindx	indeks modulasi

2.4 Random Forest

Random Forest is an ensemble learning algorithm introduced by Breiman. Random Forest consists of a collection of decision trees created using different subsets of data. Each tree in the random forest model is trained using the bootstrap technique. In the bootstrap technique, the model is generated using a random subset of the training data. To make predictions, Random Forest uses a voting process so that the class that has the largest number of trees in the forest is the final prediction or classification model and the average of the regression models. The advantage of Random forest is that it is low to overfitting, as each tree on the task acts independently and the error on the data is the same due to the random subset [9].

2.5 Evaluation Metrics

Accuracy is the ratio of the number of correct predictions to the total number of predictions made by the model. This method is used as a model performance evaluator, especially on binary classification problems can give a biased picture on unbalanced datasets because high results can be achieved by predicting only the most classes [11]. The Accuracy metric can be calculated using the following formula:

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

Recall, or Sensitivity, measures the ability of the model to correctly detect all positive samples. Recall is important because the model is expected to minimize errors by detecting as many positive cases as possible [11]. The Recall metric can be calculated using the following formula:

$$Recall = \frac{TP}{TP + FN}$$

Precision, used to measure how accurate a model is in classifying data [11]. Precision can be defined as the ratio between the number of True Positives divided by the total number of True Positives plus False Positives made by the model, both correct and incorrect. In a mathematical formula, precision can be expressed as follows

$$Precision = \frac{TP}{TP + FP}$$

The precision value indicates that the model has a good level of accuracy in identifying the positive class, making it reliable in applications that require critical decision-making. For example, in the medical field, high precision is essential to ensure that patients identified as positive actually need treatment, thereby reducing the risk of incorrect diagnosis [11]. Therefore, a deep understanding of precision evaluation metrics is essential for professionals in the field of data science and statistical analysis, especially in the context of developing and evaluating predictive models.

F1 Score, an evaluation metric used to measure the performance of classification models, especially in the context of imbalanced data. This metric is a harmonized average between precision and recall, where precision measures how many of the positive predictions are correct, while recall measures how many of the total positives are successfully identified by the model [11]. F1 Score is calculated by the formula:

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

In other words, F1 Score provides a more comprehensive picture of the balance between precision and recall, making it very useful in situations where one of the two is more important than the other.

2. MATERIAL AND METHOD

This research uses the feature selection method to optimize gender detection using random forest in recognizing human voices based on gender. This research is divided into several stages, starting from feature selection to determine 5 features based on correlation values with labels close to the value of 1, split training and testing data, gender detection with Random Forest, comparing gender detection between 22 features and 5 features to find out whether the results are not much different in terms of time or accuracy. The framework of this research can be seen in Figure 1 below.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

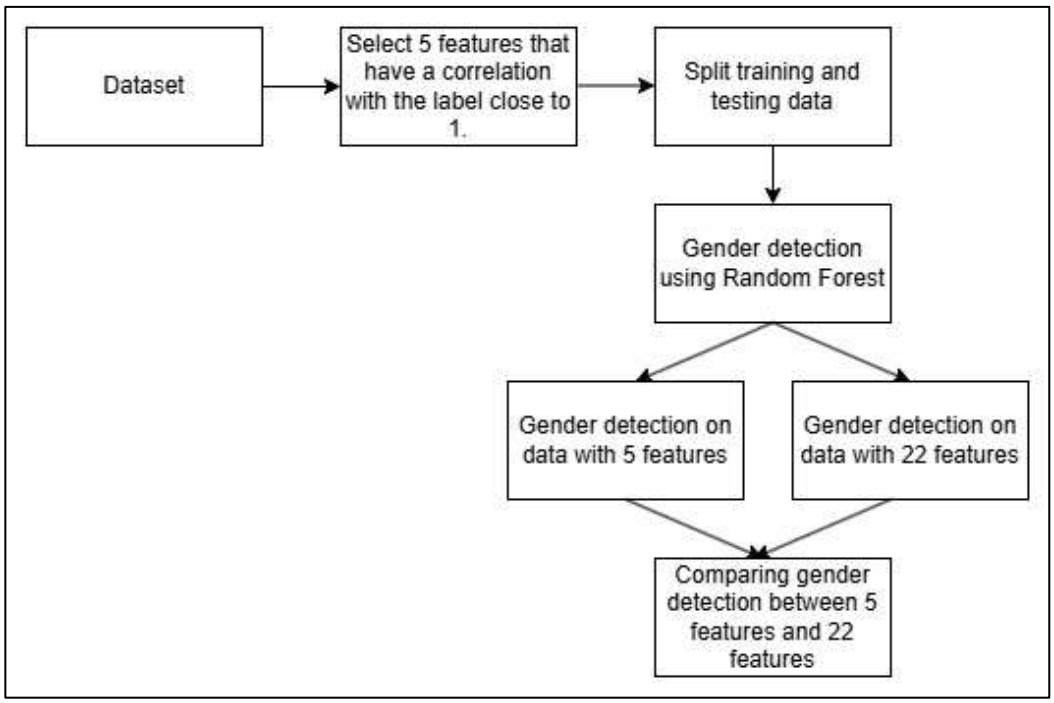

FIGURE 1. Research Block Diagram

The dataset used in this research is taken from the Kaggle website with the file name voice.csv. The dataset has 23 features with 3168 recordings consisting of 1584 male voices and 1584 female voices that have been labeled and ready to use. The target in this research is a feature called "label", which contains the gender identity of each human voice consisting of male and female.

After feature selection, the next crucial step in data processing is data splitting, which is often referred to as data splitting. This process aims to divide the entire dataset into two main parts, namely 80% of the data for training and 20% of the data for testing. This split is very important as it allows the model to learn from the training data and then be tested using data that it has never seen before. This can evaluate how well the model can generalize the information it has learned. Data splitting is done randomly to ensure that the two data sets are representative of the entire dataset. By performing a proper split, it can be more confident that the results obtained from model testing will give an accurate picture of the model's performance in real situations.

In this study, feature selection was conducted to identify and retain the most significant features that contribute to gender classification by finding five features that have correlations close to 1 with the target, indicating a very strong positive relationship. By using Pearson correlation analysis technique that can identify these features systematically. After obtaining the five selected features, it can proceed to the next stage of model development which will be more efficient and effective thanks to the selection of the right features. Through this approach, it not only improves the performance of the model, but also ensures easier and more accurate interpretation of the results, thus providing added value in data-driven decision making. The detection process was performed twice to compare the evaluation metric values as well as the effectiveness of the model in detecting voice gender between the dataset that has

undergone feature selection and the dataset that has not undergone feature selection. Evaluation metrics such as accuracy, precision, recall, and F1-score were measured to assess the performance of the model in detecting gender. The comparative analysis between these two experiments not only provides insight into the importance of feature selection in the development of machine learning models, but also highlights the potential of Random Forest as a reliable method in speech signal processing. Thus, the results of this study are expected to make significant contributions in the field of speech recognition and related applications, as well as pave the way for further research in the development of more efficient and accurate algorithms.

3. RESULT AND DISCUSSION

After a targeted feature selection process that aims to reduce features that have no effect from a total of 22 features, 5 important features were obtained, including: Standard Deviation, Interquartile Range, Quartile25, Mean Frequency, and Spectral Entrophy. The correlation value of each feature against the target can be seen in Table 2 below.The U-Net model with EfficientNetB7 CNN backbone achieves a dice coefficient value of 77.64%. This high value of the Dice coefficient shows its capacity to measure the degree of resemblance between the two areas, termed the predicted area and the actual value area, with equal importance to both.

TABLE 2.
Correlation Value of each Feature.

Feature	Correlation Value
meanfun	0.83
IQR	0.61
Q25	0.51
sp.ent *(Spectral Entrophy)*	0.49
sd *(Standard Deviation)*	0.47

Comparison of detection between models using 5 features and 22 features shows that the implementation of feature selection plays a significant role in improving the evaluation metrics and the performance of the model in detecting the dataset. In this context, the use of Random Forest algorithm as a classification method proved to be effective, due to its ability to manage high-dimensional data and provide more accurate results. By using only the 5 most relevant features, the model can reduce complexity and increase efficiency, resulting in faster processing time without compromising accuracy. In contrast, models using 22 features tend to experience overfitting, where the model is overtrained on the training data and is unable to generalize well to new data. This can be seen in the following table, which shows that evaluation metrics such as accuracy, precision, and recall improve significantly when using 5 features compared to 22 features. Thus, the application of feature selection not only helps in simplifying the model, but also improves the overall performance in detecting patterns present in the dataset, making it a very valuable approach in the development of efficient and effective Machine Learning models.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

TABLE 3.
Comparison between 22 Features and 5 Features on each Criterion.

Kriteria Pembanding	22 Fitur	5 Fitur
Akurasi (%)	98	99
Presisi (%)	98	99
Recall (%)	97	99
F1 Score (%)	98	99
Waktu (detik)	0.3	0.1

4. CONCLUSION

After performing the feature selection process, 5 main features were obtained that proved to have a significant contribution to model performance. These features are Standard Deviation, Interquartile Range, Quartile25, Mean Frequency, and Spectral Entropy. The selection of these features is based on analysis that shows that they have high relevance in helping the model understand the data better. When used in the Random Forest algorithm-based classification model, these features had a positive impact on the model's performance, as seen by the significant improvement in evaluation metrics such as accuracy, precision, and recall. These results were significantly better compared to the model that used all 22 features without selection.

By reducing the number of features used, the complexity of the model is lowered, allowing for faster training and inference. This contributes to processing efficiency, which is crucial in large-scale or real-time Machine Learning applications. Feature selection also helps reduce the risk of overfitting, which is a condition where the model overfits itself to the training data so that its performance degrades on new data.

These results underscore the importance of the feature selection process in the development of more effective, efficient and reliable Machine Learning models. This not only reduces the number of features, but also ensures that each feature used adds value to the model performance. Thus, the application of feature selection becomes one of the strategic approaches to create models that are not only accurate but also fast and reliable in various operational situations.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Dr. Ir. Sukemi, M.T and Abdurahman,S.Kom., M.Han the lecturers of the Signals and Systems course, for their guidance and support throughout the completion of this work.

REFERENCES

[1] H. Setiawan and A. Setiawan, “ANALISIS KLASIFIKASI SUARA BERDASARKAN GENDER DENGAN FORMAT WAV MENGGUNAKAN ALGORITMA K-MEANS,” 2012.

[2] S. Kushwah, S. Singh, K. Vats, and M. V. Nemade, “Gender Identification Via Voice Analysis,” International Journal of Scientific Research in Computer

Science, Engineering and Information Technology, pp. 746–753, Mar. 2019, doi: 10.32628/cseit1952188.

[3] A. A. Shafhah, P. P. Adikara, and S. Adinugroho, "Klasifikasi Jenis Kelamin Berdasarkan Suara Menggunakan Metode Learning Vector Quantization," 2020. [Online]. Available: http://j-ptiik.ub.ac.id

[4] A. K. Jha, N. Singhal, and A. Chhabra, "Voice Recognition Techniques: A Review Paper," Educational Administration Theory and Practices, Mar. 2024, doi: 10.53555/kuey.v30i3.5944.

[5] T. Sandhan, S. Sonowal, and J. Y. Choi, "Audio Bank: A high-level acoustic signal representation for audio event recognition," in International Conference on Control, Automation and Systems, IEEE Computer Society, Dec. 2014, pp. 82–87. doi: 10.1109/ICCAS.2014.6987963.

[6] A. Bodepudi, M. Reddy, S. S. Gutlapalli, and M. Mandapuram, "Voice Recognition Systems in the Cloud Networks: Has It Reached Its Full Potential?," 2019.

[7] R. Ishak, "Volume 4 Nomor 2 Juli 2022 Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix With Heatmap," Jambura Journal of Electrical and Electronics Engineering, vol. 169, [Online]. Available: https://siakun.unisan.ac.id/

[8] K. Zvarevashe and O. O. Olugbara, Gender Voice Recognition Using Random Forest Recursive Feature Elimination with Gradient Boosting Machines.

[9] L. Breiman, "Random Forests," Berkeley, 2001.

[10] A. Liaw and M. Wiener, "Classification and Regression by randomForest," vol. 2, no. 3, pp. 18–22, 2002.

[11] S. A. Hicks et al., "On evaluation metrics for medical applications of artificial intelligence," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-09954-8.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)