Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 6, No. 2, April 2024, pp: 107-115; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received February 10, 2024; revised February 26, 2024; accepted March 8, 2024; date of publication April 20, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i2.363

How to cite: Nuuruddin Hamid Arif, Mohammad Reza Faisal, Andi Farmadi, Dodon Turianto Nugrahadi, Friska Abadi, Umar Ali Ahmad, An Approach to ECG-based Gender Recognition Using Random Forest Algorithm, Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 2, pp. 107-115, April 2024.

An Approach to ECG-based Gender Recognition Using Random Forest Algorithm

Nuuruddin Hamid Arif[1], Mohammad Reza Faisal[1], Andi Farmadi[1], Dodon Turianto Nugrahadi[1], Friska Abadi[1], Umar Ali Ahmad[2,3]

[1] Computer Science Department, Lambung Mangkurat University, Banjarbaru, Indonesia
[2] Collaborative Researcher, Kanazawa University, Kanazawa, Ishikawa, Japan
[3] School of Electrical Engineering, Telkom University, Bandung, Indonesia
Corresponding author: Mohammad Reza Faisal (e-mail: reza.faisal@ulm.ac.id)

ABSTRACT Human-Computer Interaction (HCI) has witnessed rapid advancements in signal processing research within the health domain, particularly in signal analyses like electrocardiogram (ECG), electromyogram (EMG), and electroencephalogram (EEG). ECG, containing diverse information about medical history, identity, emotional state, age, and gender, has exhibited potential for biometric recognition. The Random Forest method proves essential to facilitate gender classification based on ECG. This research delves into applying the Random Forest method for gender classification, utilizing ECG data from the ECG ID Database. The primary aim is to assess the efficacy of the Random Forest algorithm in gender classification. The dataset employed in this study comprises 10,000 features, encompassing both raw and filtered datasets, evaluated through 10-fold cross-validation with Random Forest Classification. Results reveal the highest accuracy for raw data at 55.000%, with sensitivity at 46.452% and specificity at 63.548%. In contrast, the filtered data achieved the highest accuracy of 65.806%, with sensitivity and specificity at 67.097%. These findings conclude that the most significant impact on gender classification in this study lies in the low sensitivity value in raw data. The implications of this research contribute to knowledge by presenting the performance results of the Random Forest algorithm in ECG-based gender classification.

INDEX TERMS Random Forest, Gender Classification, ECG.

I. INTRODUCTION

In the development of Human-Computer Interaction (HCI), research on signal processing in the health field, such as Electrocardiogram (ECG), Electromyogram (EMG), and Electroencephalogram (EEG) signals, [1], [2], has developed rapidly. ECG contains various important information related to the medical history, identity, emotional state, age, and gender of an individual, which is reflected in the ECG signal. Researchers have demonstrated the potential of ECG as a biometric recognition tool [3] and for gender classification [4], [5]. ECG is a diagnostic tool that best represents the electro-physiological patterns of the depolarization and repolarization of the heart muscle with each heartbeat. ECG has been used extensively in the prognosis and diagnosis of various diseases and disorders [6], [7]. ECG records the heart's electrical activity, a voltage versus time graph through electrodes placed on the skin [8].

ECG reflects unique and easily measurable changes in heart potential, making it a specialized tool for human identification. [9], [10]. ECG depicts a graph of heart activity that is believed to have morphological or structural differences in each individual and remains stable over an extended period. [6], [9], [11], [12]. In diagnosing heart disease, ECG is the most crucial recording of heartbeats [13].

Gender identification is a fundamental attribute crucial in various fields such as facial recognition [14], soft biometrics, and (HCI). Gender is one of the key elements in security systems, video surveillance, online purchases, the judiciary, transportation, and demographic information collection [15], [16]. In the field of forensics, gender classification plays a role in assisting victims of criminal and civil cases, as well as the resolution of missing persons cases [17].

Gender classification is determining the male and female labels in biometric samples [18], [19]. Gender labelling can be achieved through facial recognition methods, gait analysis, dental X-rays, text data, and even using ECG signals [3], [13]. To facilitate the gender classification process based on ECG [20], a machine learning method is necessary. Machine learning can be defined as the process of extracting hidden patterns from a large dataset. Machine learning is widely applied in various fields [21]. Using machine learning methods, tasks such as prediction, classification, filtering, and data grouping can be performed [22]. One of the machine learning methods suitable for classification is Random Forest.

In previous research, a combination of Discrete Wavelet Transform (DWT), dimensionality reduction Independent Component Analysis (ICA), and Multilayer Perceptron (MLP) classification was employed for classifying arrhythmia diseases from the MIT-BIH ECG heartbeat signal data. The training and testing processes of the data utilized MLP-NN-based classification, yielding an accuracy of 96.50% for the classification method [6].

The study by [23] compared the decision tree algorithm and Support Vector Machine with multi-domain features from the MIT-BIH arrhythmia ECG signal data. The feature set consisted of eight features based on Empirical Mode Decomposition (EMD), three from Variable Mode Decomposition (VMD), and four from RR interval. The proposed method achieved the best results in decision tree classification with an accuracy of 98.89%, compared to SVM classification, which only reached an accuracy of 95.35%. The study by [24] combined DWT feature extraction with a 1-dimensional Convolutional Neural Network (CNN) algorithm for biometric authentication, using ECG-ID data with 90 subjects and only utilizing PQRST waves. The obtained results showed an accuracy of 92.2%.

The study by [25] utilized a method employing One-Dimensional Multi-Layer Co-Occurrence Matrices (1D-MLGLCM) to recognize individuals based on their ECG signals. These matrices were used to extract Haralick features for classification with algorithms such as Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), Bayes Net (BN), and K-Nearest Neighborhood (KNN). The proposed method achieved a success rate of 93.414% using SVM. The study by [6] compared the fine decision tree, medium tree, and ensemble RUSboosted tree algorithms with feature extraction from frequency and time domains for human identification. The accuracy results were Fine tree 95.2%, Medium tree 95.2%, and ensemble RUSboosted tree 95.5%.

This research aims to determine the prediction performance of gender classification based on ECG signals using the Random Forest method. The evaluation provides insights into the potential and effectiveness of the Random Forest algorithm in handling electrocardiogram (ECG) signal data. Furthermore, this research holds a significant potential impact on the broader field of biomedical signal processing, demonstrating that ECG-based gender classification is reliable and has relevant applications in developing non-invasive gender prediction methods within the healthcare domain. These findings contribute to advancing the understanding and application of Random Forest algorithms in the specialized area of gender classification using ECG signals, thereby enhancing the knowledge base in biomedical research.

II. MATERIAL AND METHODS

In general, the research process involves the classification results from the Random Forest method. These stages include data collection, partitioning the data into training and testing sets, testing the model, and evaluation. The proposed model can be seen in FIGURE 1.

A. DATASET

The dataset used in this study is a numeric signal dataset consisting of heartbeats, specifically known as the ECG ID database https://physionet.org/content/ecgiddb/1.0.0/. In this study, we utilized the entire dataset, which consists of two parts: raw data and filtered data. The data used in this research was obtained from the Physionet ECG ID, comprising 310 recording data from volunteers (44 males and 46 females aged between 13 and 75). The dataset distribution, the number of rows, and features can be seen in TABLE 1 and TABLE 2. The ECG data is continuous and consists of 10,000 features, labeled as X0 to X9999 with 310 records. It includes two classes, namely person_id and gender. The raw ECG signals are rather noisy and contain both high and low frequency noise components [26].

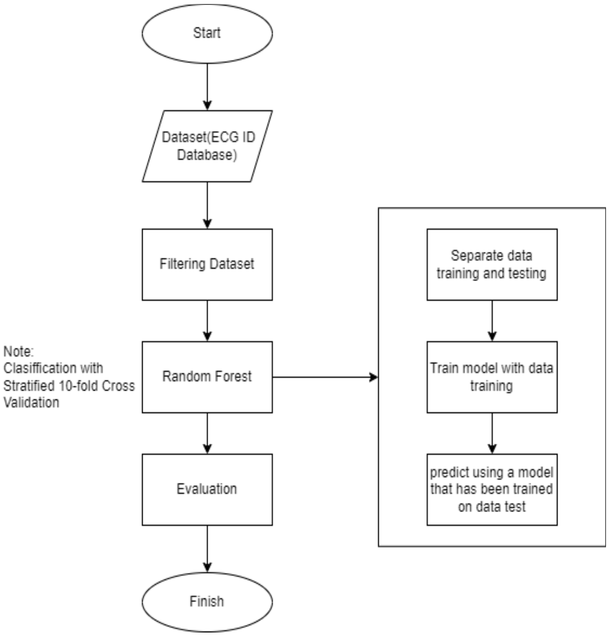

Start
Dataset(ECG ID Database)
Filtering Dataset
Note:
Clasiffication with Stratified 10-fold Cross Validation
Random Forest
Evaluation
Finish
Separate data training and testing
Train model with data training
predict using a model that has been trained on data test

FIGURE 1. The Research Flow of Random Forest Classification Models

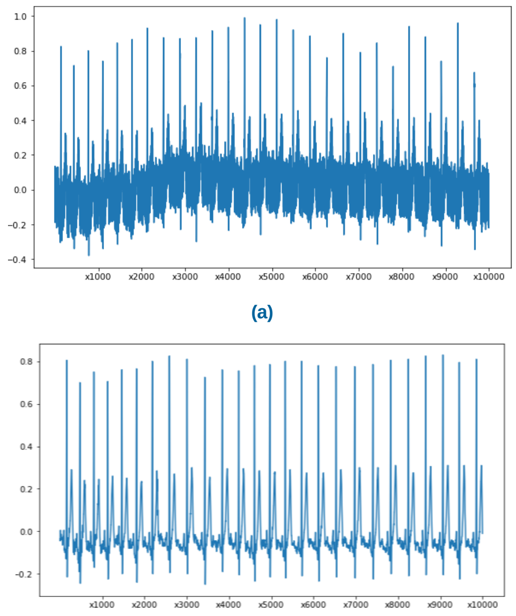
TABLE 1
Dataset distribution

No	Dataset	Man	Woman
1	Raw	44	46
2	Filtered	44	46

The dataset will be classified immediately after the filtering process, without involving additional feature selection and pre-processing steps, refer to FIGURE 1 for an overview of our research flow. This approach is implemented to directly evaluate the performance of the Random Forest model on ECG data.

TABLE 2
Data quantity

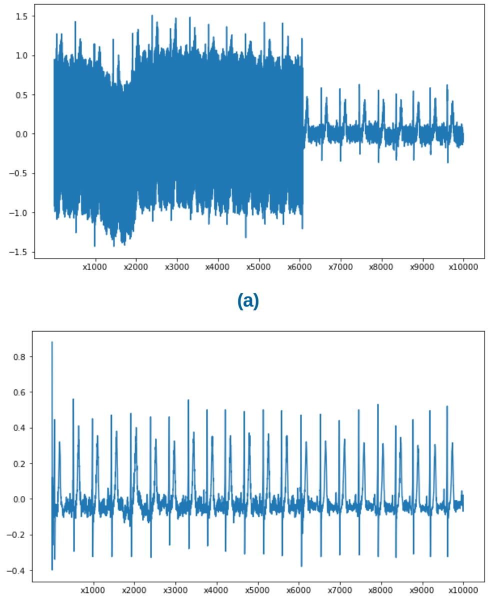
No	Dataset	Column	Row
1	Raw	10002	310
2	Filtered	10002	310



(a)

(b)

FIGURE 2. (a) Raw signal Form of Male Gender (b) Filtered signal Form of Male Gender



(a)

(b)

FIGURE 3. (a) Raw signal Form of Female Gender (b) Filtered signal Form of Female Gender

The signals in FIGURE 2 and FIGURE 3 represent raw data signals (a) and filtered data (b). We display the signals to understand the signal shapes of both datasets. FIGURE 2 is taken from TABLE III and TABLE IV, row 2 with the male gender, and FIGURE 3 is from row 307 with the female gender. The contents of both datasets can be seen in TABLE 3 and TABLE 4.

TABLE 3
Raw dataset

	X0	X1	X2	…	X9999	GENDER	PERSON_ID
0	-0.085	-0.080	-0.070	…	-0.080	male	Person_01
1	0.105	0.135	0.115	…	-0.050	male	Person_01
2	0.125	0.060	0.070	…	0.050	male	Person_01
3	0.045	0.105	0.080	…	-0.040	male	Person_01
4	-0.185	-0.240	-0.195	…	-0.060	male	Person_01
…	…	…	…	…	…	…	…
305	-0.820	-0.335	-0.180	…	-0.785	male	Person_88
306	-0.460	-0.250	-0.145	…	-0.615	female	Person_89
307	0.945	0.440	-0.260	…	-0.030	female	Person_89
308	-0.040	-0.145	0.100	…	0.270	female	Person_89
309	-0.080	-0.160	-0.040	…	0.165	female	Person_89

TABLE 4
Filtered dataset

	X0	X1	X2	…	X9999	GENDER	PERSON_ID
0	-0.115	-0.080	-0.120	…	-0.080	male	Person_01
1	0.060	0.135	0.020	…	-0.050	male	Person_01
2	-0.045	0.060	-0.045	…	0.050	male	Person_01
3	-0.170	0.105	-0.135	…	-0.040	male	Person_01
4	-0.070	-0.240	-0.050	…	-0.060	male	Person_01
…	...	…	...	…	…	…	…
305	-0.335	-0.335	-0.090	…	-0.785	male	Person_88
306	0.000	-0.250	0.130	…	-0.615	female	Person_89
307	0.335	0.440	-0.010	…	-0.030	female	Person_89
308	-0.090	-0.145	-0.065	…	0.270	female	Person_89
309	0.015	-0.160	0.010	…	0.165	female	Person_89

In the context of ECG signals, the 10,000 features refer to the representation of an ECG signal with 10,000 different values, represented by variables or features labelled X0 to X9999. ECG signal is recorded as a voltage-versus-time graph and measured using electrodes placed on the skin. The many features represent data points taken from ECG signals at specific time intervals. Each value (feature) in this representation can reflect a specific aspect of cardiac electrical activity at a particular time [8] [27].

B. RANDOM FOREST

Random Forest is an Ensemble Learning algorithm that utilizes the basic concept of Decision Trees. Random Forest consists of multiple Decision Trees built randomly and combined into one model. Random Forest combines many Decision Trees based on the Bagging technique. Bagging enhances the diversity of base learners by employing random sampling, thereby improving the algorithm's overall generalization performance [10]. To reduce the correlation between Decision Trees, Random Forest introduces random feature projection during the construction of each Decision Tree. This means that instead of applying all variables in one tree, each Decision Tree only selects a subset of features at each potential split in the Random Forest. Random feature projection can significantly reduce the correlation among trees because different trees grow on different feature sets, leading to smaller values [28].

Each tree is built using a random subset of data, and the final prediction is determined by a vote from all trees [29], [30]. This approach improves accuracy, reduces overfitting, and works well for classification and regression tasks. The formula for the decision tree algorithm is used in Eq. (1).

$$h\ (x) = 1 - \textstyle\sum_{i=1}^{T} f_i\ (x) \tag{1}$$

where $h\ (x)$ is the prediction, T is the number of trees, and $f_i\ (x)$ is the prediction of the i^{th} decision tree. Then, Bootstrap sampling as shown in Eq. (2).

$$D_i = RandomSample(D, size = N) \tag{2}$$

Randomly select N samples with replacements from the original dataset D_i for each tree (Eq. (3)).

$$m = \sqrt{p} \tag{3}$$

where p is the total number of features, m is the number of features considered for splitting at each node, typically set to the square root of the total number of features as shown in Eq. (4).

$$Gini\ (S) = 1 - \sum pi2\ k\ i = 1 \tag{4}$$

where pi is the probability of S belonging to class i, and k is the dataset's number of classes or categories. Pi represents the proportion of the dataset that belongs to class or category i (Eq. (5)).

$$(\hat{y}) = mode(h_1(x),\ h_1(x), \dots, h_T(x)) \tag{5}$$

Random Forest prediction equation combines individual tree prediction and takes the mode as the final prediction. The algorithm proceeds with the following steps [31], [32], [33].

The first step in the Random Forest algorithm is to select random samples from the database. Subsequently, a decision tree is constructed for each sample, and predictions are obtained from each decision tree. Afterwards, the frequency of each class result is counted. The most frequently occurring result is then selected as the final prediction for the Random Forest. Thus, this algorithm combines decisions from various trees to enhance overall prediction accuracy and reliability.

The study by [34] Random Forest algorithm involves adjusting several parameters, including two key parameters, to influence the model's performance. The two key parameters that impact the Random Forest model's performance are the number of trees (n_estimators) which represents the number of decision trees in the ensemble, specified as a positive integer. It signifies the quantity of decision trees to be constructed within the Random Forest model and the random number generator (random_state) parameter is utilized to set a seed value that controls the process of generating random numbers in trees. When a specific seed value is specified, each time the model is trained or predicted, the outcomes produced by functions utilizing random numbers will remain consistent.

TABLE 5
Research Random Forest parameter

Model	Parameter	Domain	
		min	max
Random Forest	n_estimator	50	500
Random Forest	random_state	42	42

The number of trees contribute a role in controlling the model's complexity, where an increase in the number of trees can enhance complexity but also potentially raise the risk of overfitting. Conversely, the random number generator, controlled by the random_state parameter, is responsible for generating random numbers used in building the trees in the ensemble. Proper configuration of both these parameters at TABLE 5 becomes crucial in the effort to achieve a balance between model complexity and overfitting control, An example of the implementation of the n_estimator parameter in the Random Forest model can be observed through the number of trees depicted in FIGURE 4 [35], [36].

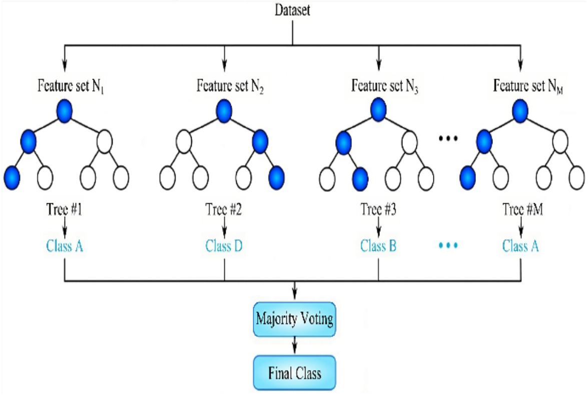


FIGURE 4. Random Forest Algorithm Structure

C. CONFUSION MATRIX

Confusion Matrix is one of the evaluation techniques in the form of a 2x2 matrix used to determine the success rate of a model by obtaining the number of correct classifications of the dataset into active and non-active classes using a classification algorithm [37]. The confusion matrix is depicted as a square matrix where rows represent the actual classes of instances, and columns represent the predicted classes (TABLE 6).

TABLE 6
Confusion matrix

		Actual	
		True	False
Prediction	True	True Positive(TP)	False Positive(FP)
	False	False Negative(FN)	True Negative(TN)

The Confusion Matrix generated by the model will be used to calculate accuracy. Accuracy is chosen for evaluating the model's performance because this research involves a classification case with balanced data [20] The accuracy is sufficient to determine the success rate of the model, and this rate is defined as the ratio of the correctly classified instances (Eq. (6)). The Confusion Matrix formula is used in the equation [35], [38], [39]

$$\text{Accurracy} = \frac{\text{TP+TN}}{\text{TP+TN+FN+TN}} \qquad (6)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \qquad (7)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (8)$$

Sensitivity is the ratio of true positive samples (TP) to the total number of samples that are actually positive (TP + FN) (Eq. (7)). Specificity is the ratio of true positive samples (TP) to the total number of samples classified as positive (TP + FP) [25] (Eq. (8)).

III. RESULT

This result calculates the accuracy produced by Random Forest, which is used to determine the success rate of the chosen method. However, the dataset used needs to separate the labels in the person_id column first because this study only requires the gender label. The results of the separation in the data can be seen in TABLE 7 and TABLE 8.

TABLE 7
Raw dataset after label separation

	X0	X1	X2	…	X9999	Gender
0	-0.085	-0.080	-0.070	…	-0.080	male
1	0.105	0.135	0.115	…	-0.050	male
2	0.125	0.060	0.070	…	0.050	male
3	0.045	0.105	0.080	…	-0.040	male
4	-0.185	-0.240	-0.195	…	-0.060	male
…	…	…	…	…	…	…
305	-0.820	-0.335	-0.180	…	-0.785	male
306	-0.460	-0.250	-0.145	…	-0.615	female
307	0.945	0.440	-0.260	…	-0.030	female
308	-0.040	-0.145	0.100	…	0.270	female
309	-0.080	-0.160	-0.040	…	0.165	female

TABLE 8
Filtered dataset after label separation

	X0	X1	X2	…	X9999	Gender
0	-0.115	-0.080	-0.120	…	-0.080	male
1	0.060	0.135	0.020	…	-0.050	male
2	-0.045	0.060	-0.045	…	0.050	male
3	-0.170	0.105	-0.135	…	-0.040	male
4	-0.070	-0.240	-0.050	…	-0.060	male
…	...	…	...	…	…	…
305	-0.335	-0.335	-0.090	…	-0.785	male
306	0.000	-0.250	0.130	…	-0.615	female
307	0.335	0.440	-0.010	…	-0.030	female
308	-0.090	-0.145	-0.065	…	0.270	female
309	0.015	-0.160	0.010	…	0.165	female

The results of label separation data will be divided into two parts: training data and testing data. Training data are used to train the model while testing data are used to make predictions based on the trained data. The data division from TABLE 7 and TABLE 8 use 10-fold cross-validation [40], and the results of the training and testing data split can be seen in TABLE 9.

TABLE 9
Training and Testing data split

Data set	Data Divided	Fold				
		1	2	…	9	10
Raw	Training	248	248	…	248	248
	Testing	62	62	…	62	62
Filtered	Training	248	248	…	248	248
	Testing	62	62	…	62	62

The divided data will be classified using the Random Forest model with parameters n_estimator and random_state. We set n_estimator from 50 to 500 of decision trees to be constructed within the Random Forest model and the random_state parameter value 42 to set a seed that controls the process of generating random numbers in trees. Each dataset will yield accuracy, sensitivity, and specificity values, and the model results can be observed in TABLE 10 and TABLE 11.

TABLE 10
Raw data result

N_estimator	Accuracy(%)	Sensitivity(%)	Specificity(%)
50	49.516%	40.968%	58.065%
100	52.581%	48.387%	56.774%
150	51.774%	45.484%	58.065%
200	54.194%	46.452%	61.935%
250	51.774%	45.161%	58.387%
300	**55.000%**	**46.452%**	**63.548%**
350	53.548%	46.129%	60.968%
400	50.806%	45.484%	56.129%
450	53.548%	47.419%	59.677%
500	54.032%	48.387%	59.677%

TABLE 11
Filtered data result

N_estimator	Accuracy(%)	Sensitivity(%)	Specificity(%)
50	60.000%	53.871%	53.871%
100	60.484%	59.677%	59.677%
150	64.839%	63.226%	63.226%
200	62.258%	59.677%	59.677%
250	61.452%	56.774%	56.774%
300	62.742%	63.226%	63.226%
350	63.226%	61.290%	61.290%
400	61.774%	61.290%	61.290%
450	64.839%	67.419%	67.419%
500	**65.806%**	**67.097%**	**67.097%**

IV. DISCUSSION

The evaluation results of the Random Forest model in this study align with previous research [25] which used the Random Forest model with the person_id label in TABLE 3 and TABLE 4. This study, however, employs the gender label to identify gender from TABLE 7 and TABLE 8. The evaluation results can be observed from the differences in accuracy, sensitivity, and specificity values between the two datasets. Raw data achieved the highest values at n_estimator 300 with accuracy, sensitivity, and specificity of 55.000%, 46.452%, and 63.548%, respectively. For n_estimator 50 to 500, raw data's accuracy and specificity values remain stable.

The visualization of raw data results from TABLE 10 can be seen in FIGURE 5.

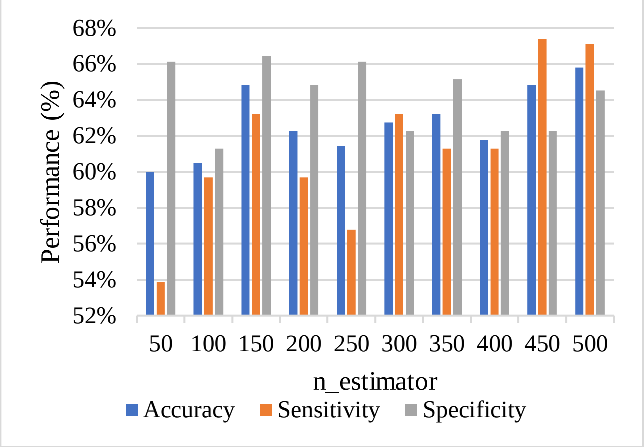


FIGURE 5. Visualization of Filtered Data Results

The filtered data achieved the highest values at n_estimator 500, with accuracy, sensitivity, and specificity reaching 65.806%, 67.097%, and 67.097%, respectively. The filtered data results in the range of n_estimator 50 to 500 show stable accuracy values, and sensitivity and specificity have better results in the range of n_estimator 300 to 500. The filtered data results from TABLE 10 can be visualised in FIGURE 6.

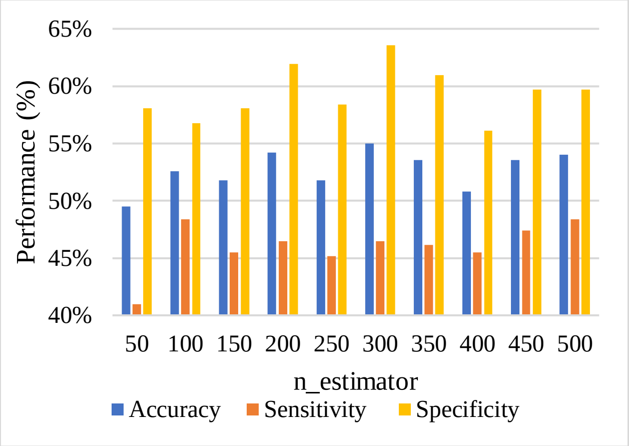


FIGURE 6. Visualization of Raw Data Results

From the two visualizations FIGURE 5 and FIGURE 6, raw data shows low sensitivity values for the n_estimator range of 50 to 500, while filtered data exhibits better sensitivity values in that range. One of the reasons for these differences can be observed in the visualization of the raw and filtered data signals in FIGURE 2 and FIGURE 3. Raw data tends to have more noise compared to filtered data.

From the noise level, it is evident that the most significant impact on the output results of this research is the decrease in sensitivity values. Therefore, noise has a considerable influence [41] on the outcomes of the n_estimator parameter in the random forest model. Further information regarding the highest values for both datasets can be found in TABLE 12 and FIGURE 7.

TABLE 12
Comparison between raw and filtered data

DATASET	N_ESTIMATOR	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
RAW	300	55.000%	46.452%	63.548%
FILTERED	500	65.806%	67.097%	64.516%

The weaknesses in this study lie in the absence of feature selection to support the model and the selection of parameter values because there was no optimization to find the best parameters. Although this research does not delve deeply into these issues, it can serve as a starting point for further investigation.

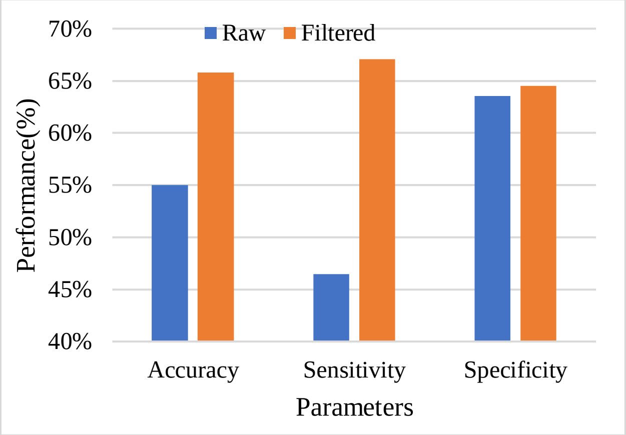


FIGURE 7. Comparison Visualization of Filtered and Raw Data Results

We have comparison of classification results between our research and the previous study conducted by [20] in TABLE 13. Our research and [20] also utilizes the ECG ID Database for gender classification.

TABLE 13
Comparison with existing work

DATASET	LSTM(%) [20]	BI-LSTM(%)[20]	OUR RESULT
RAW	57.58%	59.68%	55.00%
FILTERED	79.03%	74.19%	65.81%

In TABLE 13, our research results show lower values than the study [20]. This is attributed to the fact that the Random Forest algorithm, which we employed, performs less effectively than the LSTM and Bi-LSTM models. Random Forest, as a representation of conventional machine learning, has not been able to compete optimally with deep learning algorithms such as LSTM and Bi-LSTM, which are more effective in processing sequential data.

The strength of LSTM in processing sequential data makes it superior in this context. Conversely, the advantage of Random Forest lies in feature selection, but in this research, its capability has not been fully utilized. For future research, several steps can be taken to improve the model's performance. One approach is to use a Hybrid CNN LSTM, where CNN can extract the best features while LSTM can process sequential data. Additionally, an additional proposal could involve adding feature selection to both Random Forest and Deep Forest to obtain better features and enhance their performance.

The implications of this study contribute to knowledge by presenting the performance results of the Random Forest algorithm in gender classification. The comparison between raw and filtered data indicates that filtered data outperforms raw data when using the Random Forest model. Specifically, this research randomly assigns parameter values without prior testing to identify the optimal parameters. This approach could potentially lead to inaccuracies in classification results.

V. CONCLUSION

The study's evaluation results of the Random Forest model show that raw data performed with the highest values at n_estimator 300, achieving an accuracy of 55.000%, sensitivity of 46.452%, and specificity of 63.548%. On the other hand, filtered data achieved better results with the highest values at n_estimator 500, reaching an accuracy of 65.806%, sensitivity of 67.097%, and specificity of 67.097%.

The research uses raw and filtered datasets from this analysis, each exhibiting different performance characteristics. The sensitivity values in the raw data are notably lower across the range of n_estimators, indicating the impact of noise, especially in the sensitivity parameter. The visualizations of raw and filtered data signals further highlight the noise disparity. The evaluation reveals that the filtered data outperforms the raw data, achieving higher accuracy, sensitivity, and specificity values. The most significant drawback identified is the low sensitivity in the raw data, primarily attributed to the higher noise levels. The implications of this research suggest the need for noise reduction, feature selection, and parameter adjustments in future studies to enhance model performance.

The study provides insights into the challenges and outcomes of applying the Random Forest algorithm to gender classification based on ECG data. The study acknowledges a limitation in parameter selection, as there is no optimization for finding the best values, and emphasises the importance of addressing noise and optimizing parameters for better accuracy. Despite not delving deep into this issue, it is recognized as a potential area for further investigation. This research shows that the random forest model can determine an individual's gender information from ECG heart rate signal data. Considering the highly personal nature of medical information and the societal impact of this technology, it is crucial to be mindful of preventing the misuse of this technology on patients.

Therefore, the privacy of medical data must be carefully safeguarded. The findings contribute to knowledge by presenting the performance results of the Random Forest algorithm in ECG-based gender classification and contribute to the advancement of biomedical informatics regarding gender classification using ECG data. This is intended to facilitate experts in accurately identifying an individual's gender on a larger and broader scale through ECG signals.

ACKNOWLEDGMENT

We sincerely thank all the individuals from the Computer Science Department at Lambung Mangkurat University who have contributed to successfully completing this computer project. Their valuable inputs and suggestions have significantly improved the quality of our work. We are also indebted to our project team members for their collaboration and hard work, which have been vital to the research's accomplishments.

REFERENCES

[1] P. Wang and J. Hu, "A hybrid model for EEG-based gender recognition," *Cogn Neurodyn*, vol. 13, no. 6, pp. 541–554, Dec. 2019, doi: 10.1007/s11571-019-09543-y.

[2] N. K. Al-Qazzaz, M. K. Sabir, S. H. Bin Mohd Ali, S. A. Ahmad, and K. Grammer, "Complexity and Entropy Analysis to Improve Gender Identification from Emotional-Based EEGs," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/8537000.

[3] E. Al Alkeem *et al.*, "Robust Deep Identification using ECG and Multimodal Biometrics for Industrial Internet of Things," *Ad Hoc Networks*, vol. 121, Oct. 2021, doi: 10.1016/j.adhoc.2021.102581.

[4] J. L. Cabra, D. Mendez, and L. C. Trujillo, "Wide machine learning algorithms evaluation applied to ECG authentication and gender recognition," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, May 2018, pp. 6–12. doi: 10.1145/3230820.3230830.

[5] R. Ku. Tripathy, A. Acharya, and S. Kumar Choudhary, "Gender Classification from ECG Signal Analysis using Least Square Support Vector Machine," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 145–149, Dec. 2012, doi: 10.5923/j.ajsp.20120205.08.

[6] Z. Ahmad, A. Tabassum, L. Guan, and N. M. Khan, "ECG heartbeat classification using multimodal fusion," *IEEE Access*, vol. 9, pp. 100615–100626, 2021, doi: 10.1109/ACCESS.2021.3097614.

[7] T. V. Janahiraman and P. Subramaniam, "Gender classification based on asian faces using deep learning," *2019 IEEE 9th International Conference on System Engineering and Technology, ICSET 2019 - Proceeding*, pp. 84–89, Oct. 2019, doi: 10.1109/ICSENGT.2019.8906399.

[8] C. Y. Chen *et al.*, "Automated ECG classification based on 1D deep learning network," *Methods*, vol. 202, pp. 127–135, Jun. 2022, doi: 10.1016/J.YMETH.2021.04.021.

[9] S. R. Bajare and V. V. Ingale, "ECG Based Biometric for Human Identification using Convolutional Neural Network," *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, Jul. 2019, doi: 10.1109/ICCCNT45670.2019.8944895.

[10] C. Bohan and H. Yang, "ECG Signal Processing and Human State Detection Based on Wearable Electrodes," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Jun. 2021. doi: 10.1088/1742-6596/1952/3/032055.

[11] N. Feng, S. Xu, Y. Liang, and K. Liu, "A Probabilistic Process Neural Network and Its Application in ECG Classification," *IEEE Access*, vol. 7, pp. 50431–50439, 2019, doi: 10.1109/ACCESS.2019.2910880.

[12] B. Fatimah, G. Priyanka, R. Sultana, and N. Rekha, "Analysis of ECG for biometric identification," *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, Jul. 2020, doi: 10.1109/ICCCNT49239.2020.9225361.

[13] F. Murat, O. Yildirim, M. Talo, U. B. Baloglu, Y. Demir, and U. R. Acharya, "Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review," *Comput Biol Med*, vol. 120, p. 103726, May 2020, doi: 10.1016/J.COMPBIOMED.2020.103726.

[14] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Age and gender classification from speech and face images by jointly fine-tuned deep

neural networks," *Expert Syst Appl*, vol. 85, pp. 76–86, Nov. 2017, doi: 10.1016/j.eswa.2017.05.037.

[15] J. Rwigema, J. Mfitumukiza, and K. Tae-Yong, "A hybrid approach of neural networks for age and gender classification through decision fusion," *Biomed Signal Process Control*, vol. 66, p. 102459, Apr. 2021, doi: 10.1016/J.BSPC.2021.102459.

[16] Y. Zhou, H. Ni, F. Ren, and X. Kang, "Face and gender recognition system based on convolutional neural networks," *Proceedings of 2019 IEEE International Conference on Mechatronics and Automation, ICMA 2019*, pp. 1091–1095, Aug. 2019, doi: 10.1109/ICMA.2019.8816192.

[17] M. V. Rajee and C. Mythili, "Gender classification on digital dental x-ray images using deep convolutional neural network," *Biomed Signal Process Control*, vol. 69, p. 102939, Aug. 2021, doi: 10.1016/J.BSPC.2021.102939.

[18] A. Krishnan, A. Almadan, and A. Rattani, "Understanding Fairness of Gender Classification Algorithms across Gender-Race Groups," *Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020*, pp. 1028–1035, Dec. 2020, doi: 10.1109/ICMLA51294.2020.00167.

[19] A. Venugopal, Y. O. Yadukrishnan, and R. N. Nair, "A SVM based Gender Classification from Children Facial Images using Local Binary and Non-Binary Descriptors," *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, pp. 631–634, Mar. 2020, doi: 10.1109/ICCMC48092.2020.ICCMC-000117.

[20] K. Yudhaprawira Halim, D. Turianto Nugrahadi, M. Reza Faisal, R. Herteno, and I. Budiman, "Gender Classification Based on Electrocardiogram Signals Using Long Short Term Memory and Bidirectional Long Short Term Memory," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 606–618, 2023, doi: 10.26555/jiteki.v9i3.26354.

[21] E. H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Inform Med Unlocked*, vol. 15, p. 100178, Jan. 2019, doi: 10.1016/J.IMU.2019.100178.

[22] S. Baloch and M. S. Muhammad, "An Intelligent Data Mining-Based Fault Detection and Classification Strategy for Microgrid," *IEEE Access*, vol. 9, pp. 22470–22479, 2021, doi: 10.1109/ACCESS.2021.3056534.

[23] S. Sahoo, A. Subudhi, M. Dash, and S. Sabut, "Automatic Classification of Cardiac Arrhythmias Based on Hybrid Features and Decision Tree Algorithm," *International Journal of Automation and Computing*, vol. 17, no. 4, pp. 551–561, Aug. 2020, doi: 10.1007/s11633-019-1219-2.

[24] F. F. TALININGSIH, Y. N. FU'ADAH, S. RIZAL, A. RIZAL, and M. A. PRAMUDITO, "Sistem Otentikasi Biometrik Berbasis Sinyal EKG Menggunakan Convolutional Neural Network 1 Dimensi," *MIND Journal*, vol. 7, no. 1, pp. 1–10, Jun. 2022, doi: 10.26760/mindjournal.v7i1.1-10.

[25] N. Demir, M. Kuncan, Y. Kaya, and F. Kuncan, "Multi-Layer Co-Occurrence Matrices for Person Identification from ECG Signals," *Traitement du Signal*, vol. 39, no. 2, pp. 431–440, Apr. 2022, doi: 10.18280/ts.390204.

[26] T. S. Lugovaya, "Biometric human identification based on electrocardiogram," *Master's thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University 'LETI', Saint-Petersburg, Russian Federation*, 2005.

[27] U. Satija, B. Ramkumar, and M. S. Manikandan, "A review of signal processing techniques for electrocardiogram signal quality assessment," *IEEE Rev Biomed Eng*, vol. 11, pp. 36–52, 2018.

[28] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[29] I. Södergren, M. P. Nodeh, P. C. Chhipa, K. Nikolaidou, and G. Kovács, "Detecting COVID-19 from Audio Recording of Coughs Using Random Forests and Support Vector Machines.," in *Interspeech*, 2021, pp. 916–920.

[30] R. T. Yunardi, R. Apsari, and M. Yasin, "Comparison of Machine Learning Algorithm For Urine Glucose Level Classification Using Side-Polished Fiber Sensor," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 2, no. 2, pp. 33–39, 2020.

[31] P. Kulkarni, S. Umarani, V. Diwan, V. Korde, and P. P. Rege, "Child cry classification-an analysis of features and models," in *2021 6th International Conference for Convergence in Technology (I2CT)*, IEEE, 2021, pp. 1–7.

[32] P. A. Riadi, M. R. Faisal, D. Kartini, R. A. Nugroho, D. T. Nugrahadi, and D. B. Magfira, "A Comparative Study of Machine Learning Methods for Baby Cry Detection Using MFCC Features," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 1, pp. 73–83, 2024.

[33] A. H. Primandari, "Implementasi Metode Random Forest dan Xgboost pada Klasifikasi Customer Churn," 2020.

[34] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, and A. Fernández-Delgado, "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?," 2014. [Online]. Available: http://www.mathworks.es/products/neural-network.

[35] T. Li and M. Zhou, "ECG classification usingwavelet packet entropy and random forests," *Entropy*, vol. 18, no. 8, 2016, doi: 10.3390/e18080285.

[36] F. Khozeimeh *et al.*, "RF-CNN-F: random forest with convolutional neural network features for coronary artery disease diagnosis based on cardiac magnetic resonance," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-15374-5.

[37] G. Zeng, "On the confusion matrix in credit scoring and its analytical properties," *Communications in Statistics-Theory and Methods*, vol. 49, no. 9, pp. 2080–2093, 2020.

[38] G. Bortolan, I. Christov, and I. Simova, "Potential of rule-based methods and deep learning architectures for ecg diagnostics," *Diagnostics*, vol. 11, no. 9, Sep. 2021, doi: 10.3390/diagnostics11091678.

[39] V. M. Patro and M. Ranjan Patra, "Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy," *Transactions on Machine Learning and Artificial Intelligence*, vol. 2, no. 4, Aug. 2014, doi: 10.14738/tmlai.24.328.

[40] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction," *Clean Eng Technol*, vol. 15, p. 100664, 2023.

[41] S. Chatterjee, R. S. Thakur, R. N. Yadav, L. Gupta, and D. K. Raghuvanshi, "Review of noise removal techniques in ECG signals," *IET Signal Processing*, vol. 14, no. 9, pp. 569–590, 2020.

AUTHORS BIBLIOGRAPHY

Nuuruddin Hamid Arif originated from Banjarbaru, South Kalimantan. Since 2018, he has been involved in the academic world as a student in the Department of Computer Science, Lambung Mangkurat University. His current field of research lies in the field of data science. In addition, his final assignment includes research centered around gender classification based on ECG signals. The goal of this research effort was to determine gender from ECG signals.

Mohammad Reza Faisal was born in Banjarmasin. Following his graduation from high school, he pursued his undergraduate studies in the Informatics department at Pasundan University in 1995, and later majored in Physics at Bandung Institute of Technology in 1997. After completing his bachelor's program, he gained experience as a training trainer in the field of information technology and software development. Since 2008, he has been a lecturer in computer science at Universitas Lambung Mangkurat, while also pursuing his master's program in Informatics at Bandung Institute of Technology in 2010. In 2015, he furthered his education by pursuing a doctoral degree in Bioinformatics at Kanazawa University, Japan. To this day, he continues his work as a lecturer in Computer Science at Universitas Lambung Mangakurat. His research interests encompass Data Science, Software Engineering, and Bioinformatics.

Andi Farmadi is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science.

Dodon Turianto Nugrahadi is a lecturer in Department of Computer Science, Lambung Mangkurat University. His research interest is centered on Data Science and Computer Networking. He completed his bachelor's degree in Informatics Engineering in the UK. Petra, Surabaya in 2004. After that, he pursued a master's degree in Information Engineering at Gajah Mada University, Yogyakarta in 2009. His current area of research revolves around Network, Data Science, Internet of Things (IoT), and network Quality of service (QoS).

Friska Abadi finished her bachelor's degree in Computer Science from Universitas Lambung Mangakurat in 2011. Subsequently, in 2016, she obtained her master's degree from the Department of Informatics at STIMIK Amikom, Yogyakarta. Following that, she joined Universitas Lambung Mangakurat as a lecturer in Computer Science. Currently, she holds the position of head of the software engineering laboratory. Her current area of research revolves around software engineering.

Irwan Budiman successfully finished his bachelor's degree in the informatics department at the Islamic University of Indonesia. Subsequently, he assumed the role of a lecturer in Computer Science at Universitas Lambung Mangkurt starting from 2008. Additionally, in 2010, he pursued a master's degree in Information Systems at Diponegoro University. Currently, Irwan Budiman holds the position of chair for the computer science study program at Universitas Lambung Mangkurat. His area of research expertise lies in Data Science.

Umar Ali Ahmad completed his bachelor's and master's degrees at Telkom University in 2007 and 2012. Then, he continued his doctoral studies at Kanazawa University. He is a lecturer in the School of Electrical Engineering at Telkom University and a Collaborative Researcher at Kanazawa University, Japan. The research fields currently being carried out are computer vision, remote sensing, telecommunication and information technology.