MACHINE LEARNING AND ECG-BASED ARRHYTHMIA CLASSIFICATION EXPLOITING R-PEAK DETECTION

Thinh Pham Van, Ngoc Anh Phung, Trong Trung Anh Nguyen and Hai-Chau Le

Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

Abstract: The increasing incidence of heart-related diseases has prompted the development of efficient techniques to identify irregular heart problems. It has proven to be challenging to promptly and accurately diagnose many complicated and interferential symptom diseases including arrhythmia. Thanks to the recent evolution of artificial intelligence (AI) and the advances in signal processing, automated arrhythmia classification has become more effective and widely applied for physicians and practitioners with machine learning (ML) techniques and the use of electrocardiograms (ECG). In this work, we have investigated a machine learning-based arrhythmia classification problem based on ECGs and successfully proposed an efficient ECG-based machine learning solution employing R-peaks. In order to enhance the arrhythmia diagnosis performance, our developed approach exploits a Butterworth filter and utilizes the EEMD technique, Hilbert transformation, and a proper machine learning algorithm. The performance of the proposed method is evaluated with the most popular public dataset, MIT-BIH Arrhythmia. The numerical results imply that the developed method outperforms the notable algorithms given in the conventional works and obtains better performance with an accuracy of 93.4%, a sensitivity of 95.4%, and an F1-score of 96.3%. The attained high F1-score proves that the proposed method can effectively deal with the data imbalance while detecting arrhythmia, or in other words, it can be suitable and proper to deploy in practical clinical environments.

Keywords: ECG, EEMD, Hilbert transform, Machine learning, Arrhythmia classification.

I. INTRODUCTION

Nowadays, cardiovascular diseases have been known as one of the most dangerous health problems that caused a lot of deaths worldwide. As WHO mentioned on its website, CVD caused 17.9 million deaths for people in 2019, approximately 32% of the whole global deaths, and 85% of these deaths came from heart attacks and strokes [1]. Although with the improvement of medical and healthcare services and modern residents' healthy lifestyles in developed countries, cardiovascular problems have declined gradually, the death rate related to cardiovascular diseases is still high in low-income and middle-income countries. Therefore, developing efficient solutions to deal with these problems is extremely important and necessary to extend the life of patients. To reduce the negative effects of cardiovascular diseases, irregular rhythm identification soon is highly crucial. Fortunately, an electrocardiogram (ECG), which can capture the heart's electrical activity, is considered the most reliable method for identifying arrhythmias [2]. P-wave, T-wave, and QRS complex are essential elements of the ECG signal. The initial detectable movement on the ECG is the P-wave, which lasts for a range of 80-120 ms. The duration of the T-wave is 160-200 ms, and it indicates the repolarization of the ventricles, which is the basis for detecting rhythm irregularities. Lastly, the QRS complex consists of various deflections and bundle branch blocks, and it can be used to identify ventricular tachycardia. It can detect various forms of abnormal heartbeats and bring the foundation information for building healing strategies. Formerly, ECG comprehension for diagnosing and detecting diseases needs healthcare experts, but not whole medical facilities have enough professionals that achieve knowledge related to ECG. Moreover, ECG readings may be disrupted by interference, resulting in the loss of important data that could be used to identify arrhythmias. These challenges have motivated the development of ECG-based approaches that are simpler and more widely available to general users.

On the other hand, machine learning is recently adopted widely to solve a lot of problems in multiple fields such as healthcare, security, and telecommunications, … Many research used the outstanding ability of machine learning to identify irregular heartbeats existing in ECG signals. The work of [3] developed an automated arrhythmia identification system using different machine learning algorithms, such as Support Vector Machine (SVM), k-nearest neighbors (KNN), Random Forest (RF), and a hybrid model combining these algorithms. The objective was to improve the detection rate of irregular heartbeats with limited samples in the MIT-DB dataset without any specialized feature engineering. The results showed that SVM had the highest detection rate for abnormal heart rhythms with an accuracy of 83%. Similarly, in [4], the authors used regular feature techniques such as Principal Component Analysis (PCA) and Bag of Visual Words to

Contact author: Hai-Chau Le
Email: chaulh@ptit.edu.vn
Manuscript received: 2/2023, revised: 3/2023, accepted: 4/2023.

reduce input spatial and cluster data using 279 features of each recording from the UCI dataset. Multiple machine learning classifiers were applied to classify 16 classes, and SVM was the most effective classifier. In another notable study, the authors utilized SVM to identify abnormal rhythms in the MIT-BIH dataset [5]. They preprocessed the data by handling noise and baseline wander before segmenting it into sub-segments and removing irrelevant features. Their approach achieved approximately 91% accuracy for SVM, demonstrating its effectiveness in identifying arrhythmias. A binary classification scenario between two targets: the first class represented normal beat and the rest class included arrhythmias existing in California University at Irvine Machine Learning Data Repository was created by Pandey et al. [6]. This dataset overcame the cleaning phase and PCA was executed to analyze its properties to choose a set of features that included the most meaningful features. Afterward, each data division portion sequentially experimented with the ability of 8 classifiers after using feature selection techniques to determine the most suitable data splitting rate and the best model for abnormal heartbeat detection. The given results expressed that SVM and Naïve Bayes classifiers attained the highest accuracy of 89.74% with a division ratio of 90% original data for the training phase and 10% data for the testing phase. Moreover, the authors in [7] carried out a multiclass classification taking advantage of four machine learning algorithms including SVM, DT, RF, Naïve Bayes, and one deep learning algorithm, Artificial Neural Network (ANN), for 5 prediction classes that consist of one regular heartbeat class and the irregular heartbeat 4 classes of MIT-BIH Arrhythmia dataset. The amplitude, the area with a non-overlapped sliding window, and the area with overlapped sliding window were three different sets of features used to classify. Based on that, an accuracy of 99.59% was achieved.

In this article, an abnormal heartbeat identification method using machine learning algorithms and R-peak detection of ECG signals is proposed. By utilizing the capability of a Butterworth bandpass filter, EEMD technique and Hilbert transform, ECG signals from the MIT-BIH Arrhythmia dataset are processed to identify R-peak locations. With the detected R-peak set, multiple sub-segments are created and combined to achieve the final data. A binary classification between two classes including Normal and Abnormal is executed. The findings of this research can considerably improve the identification accuracy of irregular heartbeats.

The rest of this paper is organized as follows. Section II describes the dataset that is used in the work. Our method is explained in detail in Section III. Then, the simulation results and discussion are shown in Section IV. Finally, Section IV concludes the work.

II. DATA

In this work, MIT-BIH Arrhythmia dataset was utilized [8]. The BIH Arrhythmia Laboratory conducted a study between 1975 and 1979 and collected 48 half-hour excerpts of two-channel ambulatory ECG recordings from 47 subjects. Out of these recordings, 23 were randomly selected from a larger set of 4000 24-hour ambulatory ECG recordings, which were collected from a mixed population of inpatients (about 60%) and outpatients (about 40%) at Boston's Beth Israel Hospital. The remaining 25 recordings were specifically chosen to include less common but clinically significant arrhythmias, which would not have been well-represented in a small random sample.

The ECG recordings were converted into digital format at a sampling rate of 360 samples per second per channel, with a resolution of 11 bits and a range of 10 mV. For each record, multiple cardiologists provided annotations independently, and any discrepancies were resolved to generate computer-readable reference annotations for each beat. In total, there were approximately 110,000 annotations included in the database. Actually, we employed the benefits of stratified-lead II, which facilitates the analysis of the heart's electrical activity by producing visible waveforms. Real-time observation of the variations in heart rhythm in the electrical activity of the heart is also advantageous. We also consider two predictable arrhythmia classes which are *Normal* and *Abnormal* [9]. The Normal class includes Atrial Escape Beat, Nodal (Junctional) Escape Beat, and Normal Beat while the *Abnormal* class consists of Aberrated Atrial Premature Beat, Atrial Premature Beat, Fusion of Paced and Normal Beat, Fusion of Ventricular and Normal Beat, Left Bundle Branch Block Beat, Nodal (Junctional) Premature Beat, Paced Beat, Premature Ventricular Contraction, Right Bundle Branch Block Beat, Supraventricular Premature Beat, Ventricular Escape Beat and Unclassifiable Beats.

III. METHOD

Figure 1 shows a functional block diagram of our proposed method to detect arrhythmias. The developed solution includes two phases that are R-peak detection and machine learning classification. In the first phase, for detecting the R-peak positions, a Butterworth bandpass filter is applied to the ECG signal in order to eliminate the baseline wander and high-frequency noises then, the EEMD technique is used to decompose the filtered signal into an IMF set for combining the first three IMFs representative sufficient R-peaks information and the first derivative calculation is performed to figure out the minima or maxima points, Hilbert transformation is utilized to transform differentiated signals and identify the Hilbert envelope whose maximum value positions representing the R-peaks. In the second phase, R-peak data will be processed as input data and normalized for applying machine learning algorithms. To select a proper machine learning algorithm for creating the most effective solution for detecting arrhythmias, typical machine learning algorithms are verified and tested.

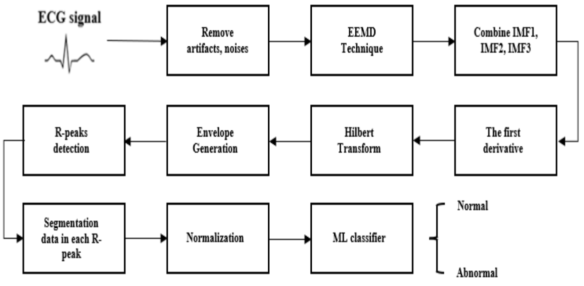

Fig. 1. Block diagram of proposed method

A. Machine learning algorithms

In this work, we investigate nine machine learning algorithms that are clustered into multiple individual algorithm groups: 1) Ensemble methods group, 2) Boosting algorithms group, 3) KNN, and 4) SVM. The reason why we selected different algorithm groups is that each group had different advantages and disadvantages, so we wanted to benchmark our data with multiple algorithms to maximize the classification result to the highest and determine the most suitable classifier.

Gradient Boosting (GB) [10]: In contrast to Bagging, Gradient Boosting is a machine learning algorithm used for both regression and classification requirements. Multiple weak learners are utilized to create a stronger learner. The algorithm works by sequentially adding weak components to an ensemble, with each ingredient attempting to correct the drawbacks made by the previous trees. The boosting procedure optimizes a loss function with respect to the ensemble of trees, by using gradient descent to minimize the loss.

Bagging (BG) [11]: Bagging is a type of ensemble learning algorithm that involves combining multiple models by training parallelly each subcomponent on individual random subsets of the training data. The main goal of bagging is to decrease the variance of the model by reducing overfitting. The final classification result was achieved based on the average value of multiple sub-elements in regression and voting results in classification problems. Bagging can be used with any type of model, but it is most commonly used with decision trees.

Random Forest (RF) [11]: A special extension of Bagging algorithms, especially, sub learners of them is multiple individual decision trees. The algorithm randomly selects a subset of the features from the training set of original data to build each subcomponent. The final result was achieved with the same method as Bagging. The purpose of using randomization in the construction of decision trees is to reduce overfitting and improve the generalization performance of the model. It is also relatively fast and can handle high-dimensional data with ease.

k-Nearest Neighbors (KNN) [11]: A machine learning method that uses the proximity of data points to make predictions. K nearest data points are identified in the feature space to a given query point and use their labels or values to predict the label or value of the query point. For classification tasks, the most common label among the k nearest points is assigned as the prediction, while for regression tasks, the average value of the k nearest points is used as the prediction. Unlike some other machine learning methods, KNN is non-parametric and makes no assumptions about the data distribution.

Support vector machine (SVM) [12]: A supervised machine learning algorithm that can be used for various tasks including classification, regression, and outlier detection. SVM aims to find the hyperplane that separates the data points of different classes in the feature space with maximum margin, which is the distance between the hyperplane and the closest data points of each class. When the data is linearly separable, the hyperplane is a line, while in non-linearly separable data, SVM uses kernel functions to map the data into a higher-dimensional feature space where a linear hyperplane can be used. The larger the margin, the better the generalization performance of the SVM model.

AdaBoost (AB) [13]: An iterative ensemble learning algorithm is used for classification tasks. The algorithm trains a sequence of base learners on a weighted version of the training data, where the weights of misclassified data points are increased in each iteration. The final prediction is a weighted combination of the predictions of all the base learners. Adaboost can be combined with any type of base learner and is known for its high accuracy and ability to handle complex datasets.

XGBoost (XGB) [14]: The algorithm is based on gradient boosting and is optimized to improve its speed and performance. It uses advanced regularization techniques to prevent overfitting and supports parallel processing to handle large datasets. XGBoost is known for its ability to achieve high accuracy and is commonly used in various machine-learning tasks.

LightGBM (LGBM) [15]: A machine learning algorithm that prioritizes training speed and memory efficiency. It uses gradient-based one-sided sampling (GOSS) to select a subset of data points for each iteration, which saves time by reducing computation. Additionally, LightGBM employs histogram-based algorithms for sorting and splitting data, further improving its efficiency.

Logistic Regression (LR) [16]: A supervised learning algorithm that estimates the parameters of a logistic function, also known as the sigmoid function. The sigmoid function maps the input variables to a value between 0 and 1, and a threshold is used to make the final classification prediction. Unlike non-parametric models, logistic regression makes assumptions about the underlying data distribution.

B. Data processing

a) Noises, artifacts and baseline wander elimination

To remove baseline and high-frequency noise from the original signal, a Butterworth filter is used. The frequency range of a normal ECG signal typically falls between 0.01 Hz and 100 Hz, but most of its energy, about 90%, is concentrated between 0.25 Hz and 35 Hz. Removing frequencies from 0 Hz to 0.5 Hz can also help reduce baseline drift. The selection of the Butterworth filter's cutoff frequency is crucial; if it is too low, the filter won't remove noise efficiently, and if it is too high, valuable information can be lost and the signal can be distorted. Therefore, a 5th-order Butterworth bandpass filter with a low cutoff frequency of 0.5 Hz and a high cutoff frequency of 35 Hz is used to smooth the original ECG signal. Figure 2 demonstrates the raw ECG and its filtered one.

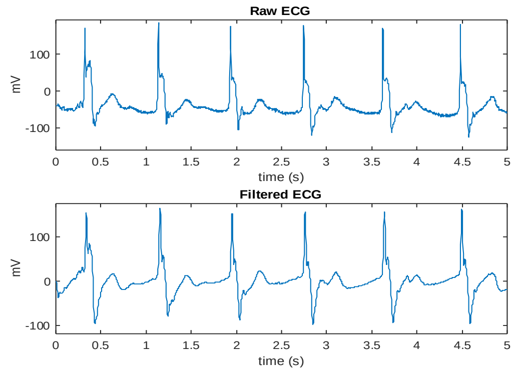

Fig. 2. Description of the baseline wander removal.

b) ECG Signal disintegration

The next step involved dividing the processed wave into several monocomponent signals, known as Intrinsic Mode Functions (IMFs), using Ensemble Empirical Mode Decomposition (EEMD) techniques. The EEMD approach is previously introduced in [17], and it offers the advantage of breaking down the original signal into its component IMFs, allowing for the examination of the properties of each component and the identification of ECG signal fiducial points such as the P wave, T wave, QRS complex, and others (see Figure 3).

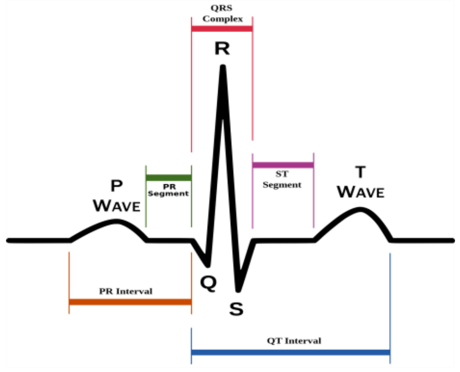

Fig. 3. Morphology of a normal ECG.

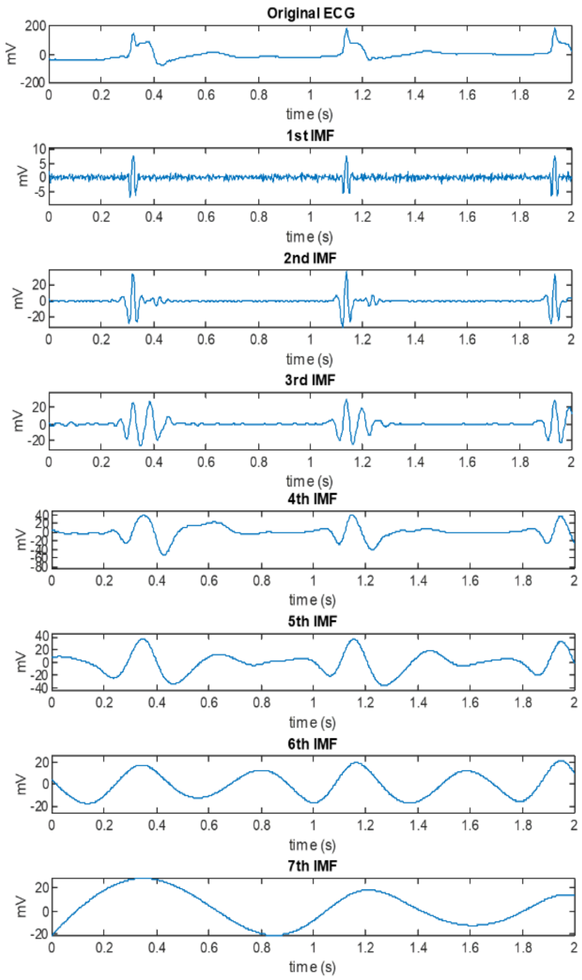

Fig. 4. The first seven IMFs of the filtered ECG signal are decomposed by using the EEMD.

Figure 4 displays the first seven IMFs in descending order of frequency, and it is evident that the majority of the valuable information is concentrated in IMF1, IMF2, and IMF3 because the waveform of these IMFs is extremely complex than the rest. These IMFs are combined into a new IMF using the function (1), and R-peak detection is performed using the multiplication of f1 and f2, where the equation of f2 was given in (2). Figure 5 illustrates the raw ECG signal, the Hilbert envelope of f1, and the Hilbert envelope of the multiplication of f1 and f2 to normalize the signal and reduce the R-peak detection error rate.

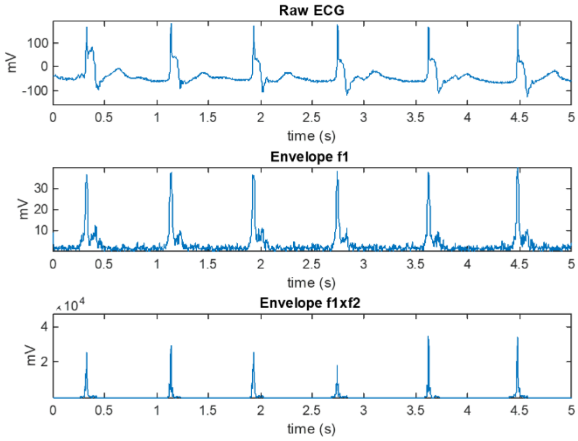

Fig. 5. (a) The raw ECG signal, (b) Hilbert envelope of the function f_1, (c) Hilbert envelope of the function f_1 x f_2.

$$f_1 = IMF_1 \times (IMF_2 + IMF_3) \qquad (1)$$

$$f_2 = IMF_1 + IMF_2 + IMF_3 \qquad (2)$$

c) Hilbert transformation and R-peak detection

The Hilbert transform is a method used for detecting R-peaks in ECG signals. As explained in a previous study [19], the transformed signal resulting from the Hilbert transform will intersect the x-axis at zero whenever a peak occurs in the differentiated signal. This allows each intersection at zero in the original waveform to be represented as a peak in the resulting HT-processed signal. Therefore, the Hilbert transform is a suitable method for detecting R-peaks in ECG signals.

After applying the first derivative to the combined IMF, the Hilbert transform is applied to calculate the Hilbert envelope using (3). The Hilbert envelope retains phase information and allows for accurate measurement of time-dependent characteristics of the signal. Additionally, it helps minimize signal distortion and improves the efficiency of the R-peaks detection stage.

$$B(t) = \left|\hat{x}(t)\right|^2 \qquad (3)$$

The location of the peak is identified by an adaptive time threshold based on the mean length of the R-R intervals between previous R-peaks (see Figure 6), which utilizes the most recent R-peak location.

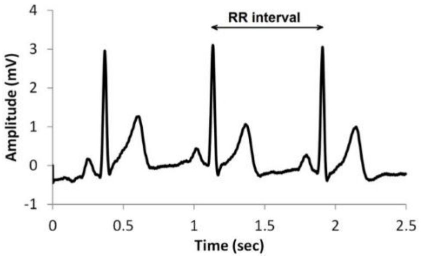

Fig. 6. Illustration for RR interval.

d) Final recording generation

After identifying the location of each R-peak in the ECG signal, the mean RR interval time is calculated by finding the average value of the time interval between consecutive R-peaks. The RR interval is an important metric for analyzing heart rate variability and can indicate irregularities in heart activity, such as arrhythmias. A segment of the signal with a width of 1.2 times the mean RR interval time is selected around each R-peak. The amplitude of each segment is then normalized to fall within the range of [0, 1]. The final ECG recordings are obtained by combining these segments.

After combining multiple signal segments of each R-peak, machine learning algorithms are used to detect heart rhythm disturbances. Two prediction classes are created, Normal and Abnormal, based on the annotations of the MIT-BIH dataset. The Normal label included usual beats, and the Abnormal label contains the remaining beats. Binary classification is performed using nine algorithms: GB, BG, RF, KNN, SVM, AB, XGB, LGBM, and LR for the two classes. Since the data used is imbalanced, the accuracy metric is not sufficient to evaluate the approach's performance. Therefore, the F1-Score is selected as the key metric to evaluate the model's performance.

IV. PERFORMANCE EVALUATION

The performance of the approach is evaluated using four standard metrics: Sensitivity (Sen), Positive Predictive Value (PPV), Accuracy (Acc), and F1-Score (F1). Sensitivity, also known as Recall or True Positive Rate, measures the ability of the model to correctly identify positive values. The Positive Predictive Value represents the rate of correctly identified true values among all predicted positive values. Accuracy is a metric commonly used to evaluate the performance of machine learning classifiers. Finally, the F1-Score is a good metric for evaluating the performance of a model when working with imbalanced data. Therefore, we use the F1-Score as the primary metric to evaluate our classifier. Moreover, a receiver operating characteristic (ROC) curve, a graphical representation of the performance of a binary classifier system, is also used. It plots the true positive rate (TPR) against the false positive rate (FPR) for different classification thresholds. The equations for these metrics are shown in (4), (5), (6), and (7).

$$Sen = TP/(TP+FN) \qquad (4)$$

$$PPV = TP/(TP+FP) \qquad (5)$$

$$Acc = (TP+TN)/(TP+TN+FP+FN) \qquad (6)$$

$$F1\text{-}Score = 2*(Sen*PPV)/(Sen+PPV) \qquad (7)$$

A. *R-peak detection*

The ECG signal's R-peaks are identified using a 5-order Butterworth bandpass filter with a low cut-off frequency of 0.5 Hz and high cut-off frequency of 35 Hz, as well as the EEMD technique and Hilbert transform. The R-peak detection rate is found to be excellent, with an error rate of only 0.028% [18]. Figures 7, 8, 9, 10, and 11 provide visual representations of the IMF combinations, the first derivative of the signal, the Hilbert transformation for each component, the Hilbert envelope, and the R-peak detection results.

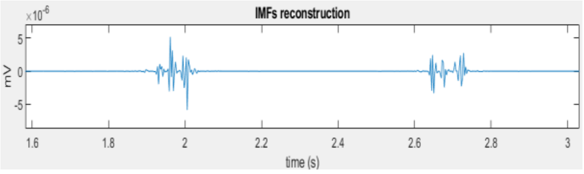

Fig. 7. The ensemble IMF combined from IMF1, IMF2, IMF3

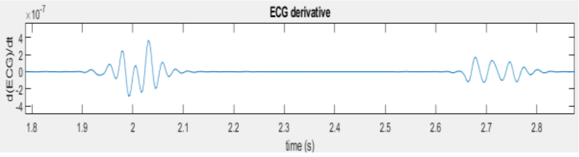

Fig. 8. The first derivative of combined signal

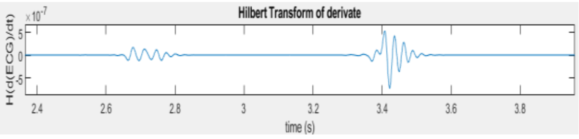

Fig. 9. Signal after using Hilbert transform

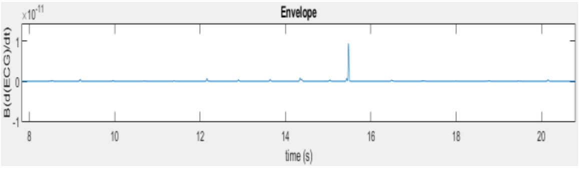

Fig. 10. Hilbert envelope

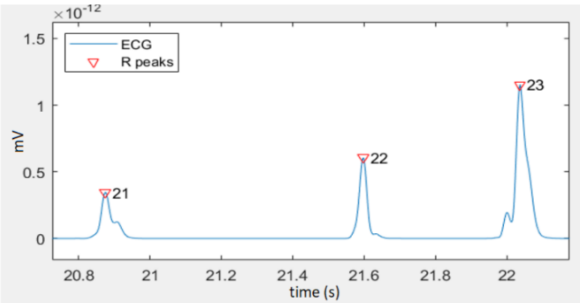

Fig. 11. R-peak detection

B. Machine learning classifiers

a) Gradient Boosting

For this algorithm, three most important hyperparameters are selected for the tuning stage, including: "n_estimators", "max_leaf_nodes", and "learning_rate". “n_estimators” controls the number of trees used in the boosting ensemble. Increasing this value can potentially improve the performance of the model, but it can also lead to overfitting. It is crucial to tune this parameter to find the right number of trees that balances bias and variance. "max leaf_nodes" limits the maximum number of leaf nodes in each tree. It controls the complexity of each tree, and therefore, the overall complexity of the boosting ensemble. Tuning this hyperparameter can prevent the model from overfitting by limiting the capacity of each tree. Finally, the “learning rate” hyperparameter is responsible for controlling the influence of each tree on the final prediction. A smaller learning_rate value implies that each tree has a smaller impact on the final prediction, which helps prevent overfitting. On the other hand, a smaller “learning_rate” also means that the boosting ensemble may need more trees to achieve the same level of performance. Consequently, it is essential to determine the best learning_rate value that balances the trade-off between bias and variance. Table 1 summarizes the best tunning values of the given hyper-parameters.

Table 1. Best tuning hyper-parameters for Gradient Boosting

Hyper-parameter	***n_estimators***	***max_leaf_nodes***	***learning_rate***
Best tuning value	500	10	1

b) Bagging

In this algorithm, “min_samples_leaf”, “min_samples_split” and “max_depth” hyper-parameters are optimized. A leaf node’s minimum number of samples is determined by “min_samples_leaf”. Besides, a minimum number of samples to split a node is also indicated by “min_samples_split” and “max_depth” which helps us know the maximum depth of the tree. This is essential to tune this hyper-parameter because if it is too low or too high, it can lead to underfitting or overfitting. Table 2 describes the best tunning values of the considered parameters.

Table 2. Best tuning hyper-parameters for Bagging

Hyper-parameter	***min_samples_leaf***	***min_samples_split***	***max_depth***
Best tuning value	1	2	None

c) Random Forest

Because Random Forest is an extension of Bagging, therefore we will use mentioned hyper-parameters in a) and b) for tuning: “max_depth”, “min_samples_leaf”, “min_samples_split”, “n_estimators”. The importance of these criteria was remarked on previously. The Random Forest’s best tuning hyper-parameters is shown in Table 3.

Table 3. Best tuning hyper-parameters for Random Forest

Hyper-parameter	***min_samples _leaf***	***min_samples _split***	***max _depth***	***n _estimators***
Best tuning value	1	2	70	200

d) K Nearest Neighbors

“n_neighbors” is the only hyper-parameter for optimizing the process for KNN. With this hyper-parameter, the number of nearest neighbors that takes part in the prediction phase is determined. Increasing this hyperparameter’s value can help to prevent overfitting by smoothing the decision boundary, but it may also increase bias by reducing the model's flexibility. The best tuning value of the KNN parameter is shown in Table 4.

Table 4. Best tuning hyper-parameters for KNN

Hyper-parameter	***n_neighbors***
Best tuning value	3

e) Support Vector Machine

The best tuning values of SVM parameters are summarized in Table 5. Four hyper-parameters are chosen for enhancing with this algorithm. “C” is a crucial value, if it is too low, underfitting can happen because simple decision will be used. In contrast, it can lead to overfitting problems. The boundary architecture is determined by “gamma”. Input data’s dimension is increased by a function and this function’s type is indicated by “kernel”. Lastly, our processed data was imbalanced between two target classes, so “class_weight” is tuned to balance the importance of training data’s classes.

Table 5. Best tuning hyper-parameters for SVM

Hyper-parameter	***C***	***class_weight***	***gamma***	***kernel***
Best tuning value	0.1	0: 1, 1: 5	0.01	‘rbf’

e) AdaBoost

Optimization hyper-parameters for AdaBoost, included: “n_estimator”, “learning_rate”, “max_depth”, and “class_weight” is mentioned specifically about their effect on the classification result of the model. The tuning hyper-parameters of this algorithm would be shown in Table 6.

Table 6. Best tuning hyper-parameters for Adaboost

Hyper-parameter	***n _estimator***	***learning _rate***	***max_depth***	***class_weight***
Best tuning value	200	1	3	0: 1, 1: 10

f) XGBoost

Besides the used hyper-parameters such as “max_depth”, “learning_rate”, and “n_estimators”, two extra hyper-parameters are bonused: “subsample”, “colsample_bytree”. The fraction of observations of each tree is randomly chosen by “subsample” and the fraction of features of each tree is randomly picked, both hyper-parameters usually in the range [0.5: 0.8]. The best tuning hyper-parameters for XGboost are described in Table 7.

Table 7. Best tuning hyper-parameters for XGboost

Hyper-parameter	***col sample _bytree***	***learning _rate***	***max _depth***	***n _estimators***	***sub sample***
Best tuning value	0.5	1	5	200	1

g) LightGBM

Five hyper-parameters are tuned and three of them are new hyper-parameters. “boosting type” indicates the using format of LGBM: Gradient Boosting or Random Forest, respectively suitable for small and large data. The number of leaves in each tree is adjusted by “num_leaves”. Finally, the minimum number of samples of a node in the decision tree is determined by “min_data_in_leaf”. Table 8 demonstrates the best tuning hyper-parameters for LGBM.

Table 8. Best tuning hyper-parameters for LGBM

Hyper-parameter	***boosting _type***	***num _leaves***	***learning _rate***	***min_data _in_leaf***	***Class _weight***
Best tuning value	‘gbdt’	15	0.05	20	0: 1, 1: 1

h) Logistic Regression

Table 9 summarizes the best tuning hyper-parameters for logistic regression algorithm. “penalty” is the extra hyper-parameter for this algorithm. It has two selections: ‘L1’ and ‘L2’ regularization. With L1, the penalty term is the absolute value of the coefficients and the square of the coefficients in L2.

Table 9. Best tuning hyper-parameters for LR

Hyper-parameter	***C***	***class _weight***	***penalty***
Best tuning value	100	0: 1, 1: 1	L2

C. Arrhythmia classification

Table 10. Arrhythmia classification performance

Classifier	***Acc(%)***	***Sen(%)***	***PPV(%)***	***F1(%)***
BG	93.1	94.9	97.3	96.1
BS	89.4	91.7	96.5	94.1
KNN	91.7	95.2	95.3	95.2
RF	**93.4**	**95.4**	**97.2**	**96.3**
SVM	88.79	89.35	98.94	93.9
AB	81.51	95.21	68.89	80.61
XGB	72.58	98.45	68.73	80.95
LGBM	88.76	88.7	99.82	93.94
LR	87.95	87.87	99.99	93.54

Table 10 presents the classification outcomes for abnormal heartbeats, revealing that the Random Forest classifier outperforms the other eight classifiers. Our proposed model achieves an F1-Score of 96.3%, which effectively classifies abnormal heartbeats even when the dataset is imbalanced. The other nine classifiers also performed well, with F1-Scores above 80%. While BG has a higher PPV than RF, its Sen and the primary metric F1-Score are inferior to those of RF, indicating that RF remained the superior model. Similarly, SVM, LGBM, and LR have higher PPV values and XGB has bigger Sen values, but the main metric F1-Score of these classifiers is worse than Random Forest. Table 10 also provides the results for detecting irregular rhythms using the nine chosen machine-learning classifiers. Additionally, the ROC curve and confusion matrix for the best-performing classifier (RF) are shown in Figures 12 and 13.

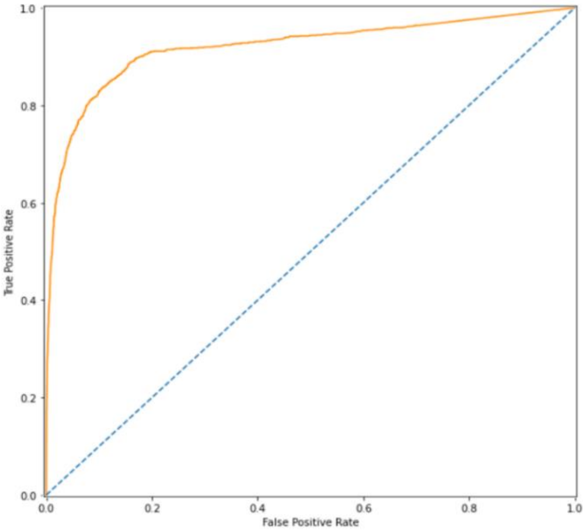

Fig. 12. ROC curve of random forest algorithm

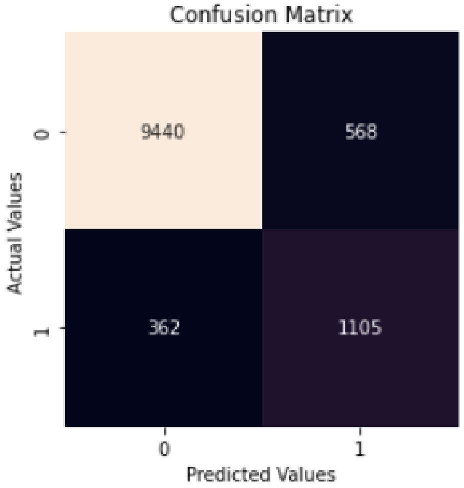

Fig. 13. Confusion matrix of Random Forest

To evaluate the effectiveness of our proposed model for arrhythmia detection, we compare it to other existing models in the field. Our comparison revealed that our model outperformed previous models with a high F1-Score, demonstrating the efficiency of our approach for handling the MIT-BIH Arrhythmia dataset. Table 11 summarizes the performance comparison between our proposed solution and conventional notable works. The proposed model achieves an F1-Score of 96.3%, indicating a high performance in identifying irregular rhythms. Nevertheless, our model had some limitations. It can only classify between two classes: Normal and Abnormal, which means it was only able to indicate the presence or absence of arrhythmia, but not the specific type of abnormal heartbeats.

In the future, multiple classifications with more complex machine learning and deep learning algorithms will be utilized to improve the overall classification result and widen the ability of the model to detect specific arrhythmias.

Table 11. Performance comparison

Model	*Acc(%)*	*Sen(%)*	*PPV(%)*	*F1(%)*
[20]	*N/A*	92.7	95.7	94.2
[21]	82.5	92.4	*N/A*	*N/A*
Proposed solution	**93.4**	**95.4**	**97.2**	**96.3**

N/A: Not applicable

V. CONCLUSIONS

The rise in heart-related diseases has led to a need for proper automatic diagnosis methods for detecting irregular heart problems which are challenging to promptly and accurately diagnosis. Thanks to the evolution of machine learning and the advance in signal processing, automated electrocardiogram-based arrhythmia detection has become more accurate and widely applied. We have studied machine learning and ECG-based arrhythmia detection and proposed an efficient solution that exploits R-peak detection and machine learning. Our proposed solution targeting a binary classification of heartbeats employs an efficient R-peak detection that uses a Butterworth bypass filter, Ensemble Empirical Mode Decomposition (EEMD), and Hilbert Transforms (HT) for processing ECG signals, and applies the most effective machine learning algorithm among typical ML algorithms to improve the performance of the arrhythmia diagnosis. In order to select the most suitable one with the highest achievable performance, typical nine ML algorithms were investigated. A popular public dataset, MIT-BIH Arrhythmia, is used for the numerical experiments. The attained results prove that our developed solution outperforms the notable traditional algorithms and it offers the best performance with an accuracy of 93.4%, a sensitivity of 95.4%, and an F1-score of 96.3%. The high obtained F1-score implies that our solution can overcome the data imbalance to detect arrhythmia correctly and be effective in practical clinical environments.

REFERENCES

[1]https://www.who.int/news-room/fact-sheets/detail/cardiovascul-ar -diseases-(cvds)

[2] Hoekema, Rudi, Gérard JH Uijen, and Adriaan Van Oosterom. "Geometrical aspects of the interindividual variability of multilead ECG recordings." *IEEE Transactions on Biomedical Engineering* 48.5 (2001): 551-559.

[3] Sraitih, Mohamed, Younes Jabrane, and Amir Hajjam El Hassani. "An automated system for ECG arrhythmia detection using machine learning techniques." Journal of Clinical Medicine 10.22 (2021): 5450.

[4] Shimpi, Prajwal, et al. "A machine learning approach for the classification of cardiac arrhythmia." *2017 international conference on computing methodologies and communication (ICCMC)*. IEEE, 2017.

[5] Subramanian, Kavya, and N. Krishna Prakash. "Machine learning based cardiac arrhythmia detection from ecg signal." *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2020.

[6] Pandey, Saroj Kumar, et al. "ECG arrhythmia detection with machine learning algorithms." *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19*. Springer Singapore, 2020.

[7] Singh, Vishavpreet, et al. "Arrhythmia detection-a machine learning based comparative analysis with mit-bih ecg data." *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*. IEEE, 2019.

[8] PhysioNet. "MIT-BIH Arrhythmia Database." PhysioNet, MIT, 1980, https://physionet.org/content/mitdb/1.0.0/.

[9] Ozdemir, Mehmet Akif, et al. "Abnormal ecg beat detection based on convolutional neural networks." *2020 medical technologies congress (TIPTEKNO)*. IEEE, 2020.

[10] C. M. Bishop, Patter recognition and Machine Learning. Book, NY, USA: Springer, 2006.

[11] Yaman, Emine, and Abdulhamit Subasi. ”Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification.” BioMed research international 2019 (2019).

[12] Noble, William S. "What is a support vector machine?." *Nature biotechnology* 24.12 (2006): 1565-1567.

[13] Schapire, Robert E. "Explaining adaboost." *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (2013): 37-52.

[14] Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." *R package version 0.4-2* 1.4 (2015): 1-4.

[15] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* 30 (2017).

[16] LaValley, Michael P. "Logistic regression." *Circulation* 117.18 (2008): 2395-2399.

[17] J. Li, G. Deng, W. Wei, H. Wang and Z. Ming, "Design of a Real-Time ECG Filter for Portable Mobile Medical Systems," *IEEE Access*, vol. 5, pp. 696-704, 2016

[18] Nguyen, Duc-Hieu, Minh-Tuan Nguyen, and Hai-Chau Le. "An Efficient Electrocardiogram R-peak Detection Exploiting Ensemble Empirical Mode Decomposition and Hilbert Transform." *2022 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2022.

[19] D. Benitez, P. Gaydecki, A. Zaidi and A. Fitzpatrick, "The use of the Hilbert transform in ECG signal analysis," *Computers in Biology and Medicine*, vol. 31, no. 5, pp. 399-406, 2001

[20] Alfaras, Miquel, Miguel C. Soriano, and Silvia Ortín. "A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection." *Frontiers in Physics* (2019): 103.

[21] Singh, Shraddha, et al. "Classification of ECG arrhythmia using recurrent neural networks." *Procedia computer science* 132 (2018): 1290-1297.

GIẢI PHÁP CHẨN ĐOÁN RỐI LOẠN NHỊP TIM DỰA VÀO TÍN HIỆU ĐIỆN TÂM ĐỒ VÀ KỸ THUẬT HỌC MÁY SỬ DỤNG PHÁT HIỆN ĐỈNH R

Tóm tắt: Tỷ lệ mắc các bệnh liên quan đến tim ngày càng tăng đã thúc đẩy sự phát triển của các giải pháp kỹ thuật hiệu quả để xác định các vấn đề bất thường về tim. Việc chẩn đoán kịp thời và chính xác nhiều bệnh về tim phức tạp và có triệu chứng giao thoa, bao gồm cả rối loạn nhịp tim, thực sự là một thách thức khó khăn. Gần đây, nhờ sự phát triển của công nghệ học máy và những tiến bộ trong xử lý tín hiệu, việc phân loại rối loạn nhịp tim tự động dựa trên tín hiệu điện tâm đồ đã trở nên hiệu quả hơn và được áp dụng rộng rãi. Trong bài báo này, chúng tôi nghiên cứu và đề xuất một giải pháp hiệu quả trong việc phát hiện rối loạn nhịp tim dựa trên kỹ thuật học máy và tín hiệu điện tâm đồ sử dụng các đỉnh R. Để nâng cao hiệu năng chẩn đoán rối loạn nhịp tim, phương pháp đề xuất của chúng tôi khai thác tính năng của bộ lọc Butterworth và sử dụng kỹ thuật EEMD, phép biến đổi Hilbert kết hợp cùng thuật toán học máy phù hợp. Hiệu năng của phương pháp đề xuất được đánh giá với bộ dữ liệu công khai phổ biến nhất, MIT-BIH Arrhythmia. Các kết quả mô phỏng số cho thấy phương pháp của chúng tôi đạt hiệu năng vượt trội so với các thuật toán đáng chú ý khác với độ chính xác 93,4%, độ nhạy 95,4% và F1-score là 96,3%. Giá trị F1-score cao chứng tỏ rằng phương pháp đề xuất có thể xử lý hiệu quả sự mất cân bằng dữ liệu trong khi phát hiện rối loạn nhịp tim, hay nói cách khác, nó có thể phù hợp và phù hợp để triển khai trong môi trường lâm sàng thực tế.

Từ khoá: Tín hiệu điện tâm đồ, EEMD, biến đổi Hilbert, học máy, phát hiện rối loạn nhịp tim.

Thinh Pham Van is currently a B.E. student in Electronics and Telecommunications Engineering Department of Posts and Telecommunications Institute of Technology (PTIT) of Vietnam. His research interests include machine learning, bioinformatics and network security.

Anh Phung Ngoc is a B.E. student in Information Technology Department of Posts and Telecommunications Institute of Technology (PTIT) of Vietnam. Her research interests include machine learning, bioinformatics.

Trong Trung Anh Nguyen received his bachelor's degree in Computer Science from Vietnam National University, Hanoi in 2013. From 2014 to 2019, he worked at Energy Research Institute, Nanyang Technological University, Singapore where he received his PhD. Currently, he is a lecturer at PTIT, Vietnam. His research interests include computational intelligence, software engineering, neural fuzzy system and future machine learning technology.

Hai-Chau Le received the B.E. degree in Electronics and Telecommunications Engineering from Posts and Telecommunications Institute of Technology (PTIT) of Vietnam in 2003, and the M.Eng. and D.Eng. degrees in Electrical Engineering and Computer Science from Nagoya University of Japan in 2009 and 2012, respectively. From 2012 to 2015, he was a researcher in Nagoya University of Japan and in University of California, Davis, USA. He is currently a lecturer in Telecommunications Faculty at PTIT. His research interests include optical technologies, network design and optimization and future network technologies. He is an IEEE member.