Classification of cardiovascular diseases using ECG signals and a Genetic Algorithm

[1]CRISTINEL-GABRIEL RUSU, [2]SIMONA MOLDOVANU, [3]LUMINITA MORARU

[1]The Modelling & Simulation Laboratory, Iorgu Iordan Secondary School Tecuci
Dunarea de Jos University of Galati, Galati, ROMANIA
[2]Department of Computer Science and Information Technology,
The Modelling & Simulation Laboratory, Dunarea de Jos University of Galati, Galati, ROMANIA
[3]Department of Chemistry, Physics & Environment,
The Modelling & Simulation Laboratory, Dunarea de Jos University of Galati, Galati, ROMANIA

Abstract: To assist physicians in diagnosing cardiovascular diseases (CVD), a variety of meaningful features such as QRS complex, T and JJ waves, QRS and QRST areas of ECG signal were extracted and further, they fed seven machine learning algorithms for CVD classification. The initially selected features are optimized using a Genetic Algorithm (GA) optimization algorithm and the new features are fed the machine learning classifiers. Thus, traditional machine learning Decision Tree (DT), Random Forest (RF), Ada Boost (AB), Quadratic Discriminant Analysis (QDA), Gaussian NB (GNB), K Nearest Neighbors (KNN), and Gradient Boosting (GB) techniques were implemented to the ECG signals. The same classification task is performed without a GA optimization algorithm. The Kaggle database of electrocardiograms (ECG) containing patients with and without cardiovascular diseases was used to classify ECG signals. The best classification performance results were collected when the machine learning algorithms were optimized using GA.

Keywords: ECG signal classification, genetic algorithm, machine learning classifiers, optimization

Received: July 16, 2023. Revised: October 25, 2023. Accepted: December 27, 2023. Published: February 21, 2024.

1. Introduction

Electrocardiogram is a fast and cheap tool used to diagnose cardiac abnormalities, [1]. ECGs provide an early screening of CVDs to diagnose patients with cardiac diseases, hence resulting in disabilities or activity limitations which are important factors in reducing the quality of life. CVDs still are not adequately examined with an objective and adapted tool even if the scientific literature shows the efforts that researchers put into solving this problem.

The CVDs produce a lot of medical data and alternative solutions could be produced to detect diseases at an early stage. The ECG signal depicts the alteration of the electrical potential generated by the heart muscle. Usually, the ECG signals occur as some critical waves and associated complexes. A beat of ECG consists of P and T-waves and QRS and QRS-T complexes and lots of algorithms were devoted to the automatic ECG classification. They allow for capturing clinical information and are prone to viable feature-extraction strategies with the declared objective of improving the classification results, [2], [3], [4], [5], [6], [7], [8], [9], [10], [11].

In the literature, there are many machine learning and deep learning methods used in the diagnosis of CVDs. In traditional machine learning, support vector machines (SVM) and decision rules were used to classify heartbeats, [12]. A hybrid classification method based on Bayesian and Extreme Learning Machine technology for arrhythmia detection was reported in, [13]. To surpass the drawbacks of machine-learning approaches, [14], proposed a deep-learning framework that integrates various networks to efficiently diagnose arrhythmia diseases. The study, [15], implemented a low-dimensional ECG beat feature vectors (12 features) for automatic arrhythmia beat classification. The study, [16], extracted various features from ECG scans and used them to train an SVM model for cardiac arrhythmia prediction. The study, [17], extracted ECG features using a Discrete Wavelet Transform method for denoising and classifying them using an SVM classifier. The study, [3], conducted an ECG classification strategy using the Extreme Learning Machine and GA-Wavelet Kernel Extreme Learning Machine (WK-ELM) technique. The GA optimization tool provided the optimal values of the WK-ELM parameters for classification. An algorithm based on GA combined with a back propagation neural network along with a wavelet decomposition for feature extraction is proposed for classifying ECG signals, [18]. GA is used to solve the dimensionality issue and to optimize the weights and biases of the network, as well. The study, [19], used GA and deep learning algorithms for the prediction of some types of anemia. To improve the performance of the prediction, the hyperparameters were optimized using GA.

Few studies have been conducted to get the optimal classification model of CVD based on the tuning of useful parameters. In most cases, optimization is imperative along with the best prediction of the dependent variable based on the independent variables.

Also, finding the optimal input parameters without more computational costs is another major requirement. GA is one of the commonly used optimization tools and the feature tuning for various machine learning classifiers using GA increases the accuracy and reduces the computational complexity, [20], [21].

The main objective of this study is to explore the GA's ability to optimize the classification prediction for the detection of CVDs by using the ECG signals. By combining GA with different classifiers to form a new approach we intended to produce better utility values of inputted features and to get the best classification output results. The advantages of GA are used in the solution of complex optimization problems as it utilizes a fitness function for optimization. For this purpose, an optimization approach based on GA is proposed to optimize the prediction of seven traditional machine learning algorithms. The CVD diagnosis is accomplished based on different features of ECG signals. The most relevant features optimized by GA fed the following classifiers: DT, RF, AB, QDA, GNB, KNN, and GB. Also, the initially selected features inputted the same classifiers. The best classification performance has been achieved when GA optimizes the classification rules.

2. Materials and Methods

2.1 Dataset

In this study, CVD data of 421 patients taken from the, [22], were used. The data set consists of 176 patients with CVD and 245 cases without CVD, [23]. As shown in Table I, seven selected features and two classes of CVD/healthy were used in the study.

TABLE I. LIST OF FEATURES IN THE DATASET

Features	Type	Min	Max	Average
Q wave	Numeric	-2.7	19.9	-0.19264
R wave	Numeric	0	0	6.087411
S wave	Numeric	-13.8	3.7	-0.9848
T wave	Numeric	-8.7	155.2	0.996675
QRS area	Numeric	-33.3	74.3	14.16936
QRST area	Numeric	-18	1	21.26295
JJ wave	Numeric	-9	0	-0.21045

The P wave appears during a trial depolarization. The QRS complex consists of Q, R, and S waves generated during ventricular depolarization. The T wave is generated during the repolarization action of the ventricles preparing the heart for the next heartbeat. The QRS area is generated by summing the absolute values of Q, R, and S wave areas. The QRS-T area is obtained when the T-wave area is subtracted from the QRS area. The dataset is organized into 75% for training and 25% for testing.

2.2 Genetic Algorithm optimization method & classifiers

GAs are evolutionary algorithms mostly used for relevant feature selection, [24]. Usually, GA replicates the most adequate individual and extends this procedure at a colony called population (individualized by chromosomes) for best-generation reproduction. GA uses the fitness function to generate new solutions based on the initially selected solutions. To establish the best chromosome mutation, GA randomly interchanges the genes, and the determined rate of mutation depends on the chromosome's fitness. Three basic operations are followed in a GA: selection, crossover, and mutation, [5]. The main task of the fitness function is to optimize the rules during the selection of the solution. Only the solutions fully satisfying the fitness function are further selected to participate in the reproduction. For the new generation reproduction either crossover (i.e., a new chromosome is generated between two parent chromosomes and contains characteristics of both parents' chromosomes) or mutation (i.e., different genes are changed randomly from a single parent) are utilized.

Classification represents a supervised learning method to extract patterns able to describe the importance of analyzed classes or to predict future tendencies. Feature selection helps identify and remove irrelevant and/or redundant features. GA optimization algorithm helps in selecting those values of the parameters or features inputted so that the percentage error rate of the classification is minimized. Figure 1 displays the self-explanatory block diagram of the proposed algorithm for ECG signal classification.

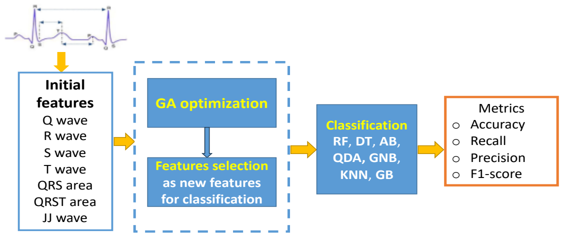

Fig. 1. The self-explanatory block diagram of the proposed algorithm for ECG signals classification. The initially selected features are optimized using a GA optimization algorithm and the new features are fed seven machine learning classifiers.

3. Results and Discussion

Table II shows the models' parameters produced during training. They were continuously updated during the learning process until the mapping between the input features and the targets became optimal. Figure 2 shows the curve of the fitness function in terms of generations for the algorithms selected as classifiers. It evaluates how close the provided solution is to the optimal solution. In the case of the RF classifier, the characteristic error computed for each generation converges to the least value.

TABLE II. THE OPTIMAL MODELS' PARAMETERS FROM THE DATA DURING TRAINING

Classifiers and GA	Parameters
GA	cross-validation splitting strategy (cv=3), population size (ps=20), generations (g=20), evolutionary algorithm was Mu Plus Lambda
RF	The number of trees in the forest (nt=100),
DT	Gini was the function for measuring the quality of a splitting data
AB	The maximum number of estimators (ne=100), controls the random seed (cr=40)
QDA	Absolute threshold (at=1.0e^-5)
GNB	Portion of the largest variance of all features (pv=1.0e^-10)
KNN	Number of neighbors (nn=2), Ball Tree was the logarithm used to compute the nn, minkowski was the metric to use for distance computation.
GB	Learning rate (lr=0.1), Friedman was the function to measure the quality of a split

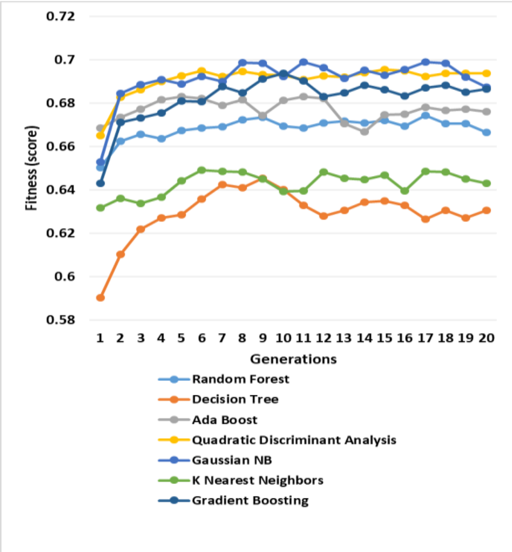

Fig. 2. The fitness function convergence in terms of generations for the selected classifiers. For DT, QDA, GNB, and KNN classifiers convergence is obtained but the dispersion of the values denotes the instability of the solution

Table III contains the relevant features selected by GA for chosen classifiers. GA introduces mutations that reduce overfitting. The higher number of newly selected features belongs to the RF algorithm and the lower number is for the DT algorithm. The performance statistics of both GA-machine learning classifiers and machine learning classifiers individually are shown in Table IV and Table V.

TABLE III. THE NUMBER OF SELECTED FEATURES BY GA

Method	Features	Classifier	Accuracy (%)
Kernel difference weighted KNN [25]	ECG data UCI	KNN	70.66
Classification and Regression Trees [26]	ECG data UCI	RF	84.58
		Subspace KNN	71.39
Kernel Extreme learning machine (KELM) with GA [27]	PhysioNet	KELM-GA	86.67
T-wave integral and total integral [28]	PhysioNet	Naive Bayes NB	94.7
Proposed method	ECG data UCI	RF-FA	98.1

TABLE IV. THE COMPARISON RESULTS OF GA-MACHINE LEARNING CLASSIFIERS. CONFUSION MATRIX AND PERFORMANCE METRICS INDICATE THE EXCELLENT RESULTS PROVIDED BY RANDOM FOREST (RF) ALGORITHM

Classifiers	TP	TN	FP	FN	Accuracy	Precision	Recall	F1_score
RF	68	36	0	2	0.981	1.000	0.971	0.986
DT	67	17	21	1	0.792	0.761	0.985	0.859
AB	66	28	10	2	0.887	0.868	0.971	0.917
QDA	63	16	22	5	0.745	0.741	0.926	0.824
GNB	62	20	18	7	0.766	0.775	0.899	0.832
KNN	58	29	9	10	0.821	0.866	0.853	0.859
GB	62	28	10	6	0.849	0.861	0.912	0.886

TABLE V. THE COMPARISON RESULTS OF MACHINE LEARNING CLASSIFIERS, INDIVIDUALLY

Classifiers	TP	TN	FP	FN	Accuracy	Precision	Recall	F1_score
RF	61	41	16	9	0.803	0.792	0.871	0.830
DT	46	38	22	21	0.661	0.676	0.687	0.681
AB	61	34	22	14	0.725	0.735	0.813	0.772
QDA	67	24	18	18	0.717	0.788	0.788	0.788
GNB	62	28	28	9	0.709	0.689	0.873	0.770
KNN	58	35	21	13	0.732	0.734	0.817	0.773
GB	52	45	16	14	0.764	0.765	0.788	0.776

The classifiers had various performance scores in the classification, as shown in Table V. We can notice improved performance statistics when the selected features are optimized using GA. Overall, the RF classifier produced a superior performance score over other classifiers when the same data of ECG signals was used for both experimental conditions. AB and GB classifiers performed well with an accuracy of 88.7% and 84.9%, respectively. Moreover, QDA and GNB had poor classification performance. The reported results indicate that the GA can reduce the number of features used improving the performance of the original algorithms. However, this affirmation is not true for all analyzed classifiers. In the case of the DT algorithm, a massive feature reduction has been provided by GA with the cost of the performance of classification. In this case, the DT classifier has a compromised accuracy as the system GA-DT has predicted a few dependent variables based on the independent variables and GA-DT experienced collinearity issues when using many independent variables. On the other hand, the RF algorithm performs consistently before and after optimization. It has the best performance in both experimental cases. Classification using DT, QDA, and GNB perform poorly compared to other methods. The obtained results are compared with the existing state-of-art approaches. It is discussed in Table VI,[25],[26],[27],[28].

TABLE VI. COMPARISON OF THE PROPOSED APPROACH WITH OTHER METHODS

Classifiers	Features						
	JJ	Q	R	S	T	QRS area	QRST area
RF	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
DT	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
AB	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
QDA	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
GNB	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE
KNN	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
GB	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE

Some limitations of this study include the limited dataset. We have discussed two different conditions namely patients with CVD and without CVD. In future work, we will discuss particular conditions such as ECG with Epilepsy and EEG signal with tumor, etc.

4. Conclusion

A method based on the Genetic Algorithm and machine learning classifiers was developed to optimize the inputted features for the ECG signals classification task. When the genetic algorithm worked with classifiers for ECG classification an optimization of the inputted features was achieved. Comparison with initial classification results, that were provided by classifiers solely, indicated that the proposed method is more effective. We conclude that GA is a valuable tool for simplifying the classification task by reducing the number of inputted features and improving the algorithm's performance. As a future study direction, the proposed technique will be tested on other ECG databases.

Acknowledgment

This work was supported by the project "DINAMIC", financed by the Romanian Ministry of Research and Innovation, Contract no. 12PFE/2021.

References

[1] U.R. Acharya, H. Fujita, V.K. Sudarshan, S.L. Oh, M. Adam, J.E. Koh, J.H. Tan, D.N. Ghista, R.J. Martis, C.K. Chua, K. P. Chua, S. T. Ru, “Automated detection and localization of myocardial infarction using electrocardiogram: a comparative study of different leads,” Knowl Based System, vol. 99, 2016, pp. 146-156, doi:10.1016/j.knosys.2016.01.040.

[2] F.G. Silva Teodoro, S.M. Peres, C.A.M. Lima, “Feature Selection for Biometric Recognition Based on Electrocardiogram Signals,” presented at the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2911-2920, doi: 10.1109/IJCNN.2017.7966216.

[3] A. Diker, E. Avcı, M. Gedikpinar, “Determination of R-peaks in ECG signal using Hilbert transform and Pan-Tompkins algorithms,” presented at the 25th Signal Processing and Communications Applications Conference (SIU) in 2017, Antalya, Turkey, 2017, pp. 1-4, doi: 10.1109/SIU.2017.7960235.

[4] S. Chen, W. Hua, Z. Li, J. Li, X. Gao, “Heartbeat classification using projected and dynamic features of ECG signal,” Biomedical Signal Processing Control, vol. 31, 2017, pp. 165-173, doi: https://doi.org/10. 1016/j.bspc.2016.07.010.

[5] A. Diker, D. Avci, E. Avci, M. Gedikpinar, “A new technique for ECG signal classification genetic algorithm Wavelet Kernel extreme learning machine,” Optik, vol. 180, 2019, pp. 46-55, doi: https://doi.org/10.1016/j.ijleo.2018.11.065.

[6] H. Dai, H.G. Hwang, V.S. Tseng, “Convolutional neural network based automatic screening tool for cardiovascular diseases using different intervals of ECG signals,” Computer Methods and Programs in Biomedicine, Volume 203, 2021, doi: https://doi.org/10.1016/j.cmpb.2021.106035.

[7] X. Wang, R. Li, S. Wang, S. Shen, W. Zhang, B. Zhou, Z. Wang, “Automatic diagnosis of ECG disease based on intelligent simulation modeling,” Biomedical Signal Processing and Control, vol. 67, 2021, doi: https://doi.org/10.1016/j.bspc.2021.102528.

[8] S. Moldovanu, L. Moraru, D. Bibicu, “Characterization of myocardium muscle biostructure using first order features,” Digest Journal of Nanomaterials and Biostructures, vol. 6, No 3, 2011, PP. 1357-1365, Available: http://www.chalcogen.infim.ro/1357_Moldovanu.pdf

[9] D. Bibicu and L. Moraru, “Cardiac cycle phase estimation in 2D echocardiographic images using Artificial Neural Networks,” IEEE Transactions on Biomedical Engineering (T-BME), 60 (5), pp. 1273-1279, 2013, doi: 10.1109/TBME.2012.2231864.

[10] Dhananjaya V, Geetha M, “Effective Analysis and Accurate Detection of Common Diseases in ECG Signals and Classifications and Monitoring Through Cloud Computing Technology,” WSEAS Transactions on Biology and Biomedicine, vol. 19, pp. 107-117, 2022.

[11] Sreeja M. U., Supriya M. H., “A Deep Convolutional Model for Heart Disease Prediction based on ECG Data with Explainable AI,” WSEAS Transactions on Information Science and Applications, vol. 20, pp. 254-264, 2023.

[12] J.R. Annam, B.R. Surampudi, “Inter-patient heart-beat classification using complete ECG beat time series by alignment of R-peaks using SVM and decision rule,” presented at the 2016 International Conference on Signal and Information Processing (IConSIP), 2016, pp. 1-5, doi: 10.1109/ ICONSIP .2016.7857480.

[13] F.A. Elhaj, N. Salim, T. Ahmed, A.R. Harris, T.T. Swee, “Hybrid classification of Bayesian and extreme learning machine for heartbeat classification of arrhythmia detection,” presented at the 6th ICT International Student Project Conference (ICT-ISPC-2017), Johor, Malaysia,

2017, pp. 1-4, doi: 10.1109/ICT-ISPC.2017.8075320.

[14] S. Irfan, N. Anjum, T. Althobaiti, A.A. Alotaibi, A.B. Siddiqui, N. Ramzan, “Heartbeat Classification and Arrhythmia Detection Using a Multi-Model Deep-Learning Technique,” vol. 22, Sensors, 2022, doi: https://doi.org /10.3390/s22155606.

[15] Q. Qin, J. Li, L. Zhang, Y. Yue, C. Liu, “Combining low-dimensional wavelet features and support vector machine for arrhythmia beat classification,” Scientific Reports, vol. 7 (1), 2017, pp. 6067, doi: 10.1038/s41598-017-06596-z.

[16] Z. Ge, Z. Zhu, P. Feng, S. Zhang, J. Wang and B. Zhou, ”ECG-Signal Classification Using SVM with Multi-feature”, presented at the 8th International Symposium on Next Generation Electronics (ISNE) in 2019, Zhengzhou, China, 2019, pp. 1-3, doi: 10.1109/ISNE.2019.8896430.

[17] R. Thilagavathy, R. Srivatsan, S. Sreekarun, D. Sudeshna, P.L. Priya, B. Venkataramani, ”Real-time ECG signal feature extraction and classification using support vector machine,” In Proceedings of the 2020 International Conference on Contemporary Computing and Applications (IC3A), Lucknow, India, 5–7 February 2020; pp. 44–48. doi: 10.1109/IC3A48958.2020.233266.

[18] H. Li, D. Yuan, X. Ma, D. Cui, L. Cao, ”Genetic algorithm for the optimization of features and neural networks in ECG signals classification,” Scientific Reports vol. 7, 2017, doi: https://doi.org/10.1038/srep41011.

[19] S. Kilicarslan, M. Celik, Ș. Sahin, ”Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification”. Biomedical Signal Processing and Control, vol. 63, 2020, doi:10.1016/j.bspc.2020.102231.

[20] Chernbumroong S, Cang S, Yu H., “Genetic algorithm-based classifiers fusion for multisensor activity recognition of elderly people”. IEEE J Biomed Health Inform. 2015 Jan;19(1):282-289. doi: 10.1109/JBHI.2014.2313473.

[21] Del Valle, A.M., Mantovani, R.G. & Cerri, R. “A systematic literature review on AutoML for multi-target learning tasks. Artif Intell Rev (2023). https://doi.org/10.1007/s10462-023-10569-2.

[22] Kaggle database: Cardiac Arrhythmia Database, available at: https://www.kaggle.com/datasets/bulentesen/cardiac-arrhythmia-database

[23] H. Guvenir, B. Acar, H. Muderrisoglu, and R. Quinlan, Arrhythmia. UCI Machine Learning Repository, 1998, doi: https://doi.org/10.24432/C5BS32.

[24] G.T. Reddy, M.P.K. Reddy, K. Lakshmanna, D.S. Rajput, R. Kaluri, and G. Srivastava, “Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis”, Evolutionary Intelligence, vol. 13(2), 2019, pp.185–96, doi:10.1007/s12065-019-00327-1.

[25] W. M. Zuo, W. G. Lu, K. Q. Wang, and H. Zhang, “Diagnosis of cardiac arrhythmia using kernel difference weighted KNN classifier”, 2008 Computers in Cardiology, Bologna, Italy, 2008, pp. 253-256, doi: 10.1109/CIC.2008.4749025.

[26] An. Agarwal, A. Agarwal, “Classification and Diagnosis of Cardiac Arrhythmia using an ECG-based Ensemble Approach”, Journal of Biomedical Engineering and Medical Imaging, vol. 5, no 5, 2018, pp. 20-26, http://dx.doi.org/10.14738/jbemi.55.5452.

[27] S. Dalal, and V.P. Vishwakarma, “GA based KELM optimization for ECG classification”, Procedia Computer Science, vol. 167, 2020, pp. 580-588, https://doi.org/10.1016/j.procs.2020.03.322.

[28] N. Safdarian, N.J. Dabanloo, G. Attarodi, “A new pattern recognition method for detection and localization of myocardial infarction using T-wave integral and total integral as extracted features from one cycle of ECG signal”, Journal of Biomedical Science and Engineering, vol.7 no.10, 2014, pp. 818-824. doi: 10.4236/jbise.2014.710081.