Turkish Journal of Science & Technology
19(1), 147-159, 2024

Research Paper
https://doi.org/10.55525/tjst.1324854

Classification of the Cardiac Arrhythmia Using Combined Feature Selection Algorithms

Murat TUNÇ[1]*, Gülnur Begüm CANGÖZ[2]
[1,2]Elektrik-Elektronik Mühendisliği, Mühendislik Fakültesi, Başkent Üniversitesi, Ankara, Türkiye
[1]murattunc0110@gmail.com

(Geliş/Received: 12/07/2023; Kabul/Accepted: 26/03/2024)

Abstract: The prediction of heart disease has gained great importance in recent years. Efficient monitoring of cardiac patients can save tremendous number of lives. This paper presents a method for classification and prediction of electrocardiogram data obtained from 452 patients representing the risk of cardiac arrhythmia. The aim of the study is to select highly related features with arrhythmia risk by using three different feature selection algorithms. In addition, various machine learning models are utilized for the classification task such as k-Nearest Neighbors (k-NN), Support Vector Machines (SVM) and Decision Tree (DT). The experimental results show that combination of a purposed feature selection method which later is called "Matched Selection" using SVM classifier outperforms other combinations and have an accuracy of 81.27% while k-NN and DT classifiers have an accuracy of 69.66% and 73.50% respectively. The study, in which detailed analyses are presented comparatively, is promising for the future studies.

Keywords: Arrhythmia, Classification, Feature Selection, Matched Selection.

Öznitelik Seçim Algoritmalarının Kombinasyonu ile Kardiyak Aritminin Sınıflandırılması

Öz: Kalp hastalıklarının önceden tahmin edilmesi son yıllarda büyük önem kazanmıştır. Kalp hastalarının etkin bir şekilde izlenmesi, sayısız hayatın kurtarılmasını sağlayabilir. Bu makale, kardiyak aritmi riski taşıyan 452 hastadan elde edilen elektrokardiyogram verilerinin sınıflandırılması ve hastalıkların tahmin edilmesi için yenilikçi bir yöntem sunmaktadır. Bu çalışmanın ana hedefi, üç farklı öznitelik seçme algoritmasını kullanarak aritmi riski ile yüksek derecede bağlılık gösteren özniteliklerin seçilmesidir. Ayrıca, sınıflandırma görevi için En yakın komşular algoritması(k-NN), Destek Vektör Makineleri (SVM) ve Karar Ağaçları (DT) gibi çeşitli makine öğrenimi modelleri kullanılmaktadır. Deneysel sonuçlar, Destek Vektör Makineleri (SVM) sınıflandırıcısı ve "Eşleştirilmiş Seçim" öznitelik seçim yönteminin diğer kombinasyonları geride bıraktığını göstermektedir. Bu kombinasyon %81.27 doğruluk oranına sahipken, k-NN ve DT sınıflandırıcılarının doğruluk oranları sırasıyla %69.66 ve %73.50'dir. "Detaylı analizlerin karşılaştırmalı olarak sunulduğu bu çalışma, gelecekteki araştırmalar için umut vadetmektedir.

Anahtar Kelimeler: Aritmi, Sınıflandırma Öznitelik Seçimi, Eşleştirilmiş Seçim.

1. INTRODUCTION

Cardiac diseases are considered as life threatening among chronic diseases which are responsible for many casualties around the world. Prevention or early detection can save many lives by monitoring heart activity. There are some technological methods such as ECG to examine these cardiac activities. An electrocardiogram (ECG) is the basic technique used to monitor heart activity in human beings. This technique can detect and diagnose various heart conditions, including arrhythmias (irregular heartbeats), heart attacks, and heart disease [1]. During an ECG, electrodes are placed on the chest, arms, and legs, and they pick up the electrical impulses generated by the heart. The impulses are then transmitted to a machine that records the data and produces a graph of the heart's activity. The results of an ECG can help doctors determine if there are any problems with the heart's rhythm or if the heart muscle is damaged [2]. The P, Q, R, S, and T waves are components of an electrocardiogram (ECG) that represent different phases of the cardiac electrical cycle [3]. Certain features of these waves such as peaks, duration and pauses between two waves provide key information to diagnose any abnormalities. In an arrhythmia, the electrical impulses that control the heart's rhythm are disrupted, causing the heart to beat too fast, too slow, or irregularly. There are many types of arrhythmias, including atrial fibrillation, atrial flutter, supraventricular tachycardia, ventricular tachycardia, and ventricular fibrillation [4]. Arrhythmia can cause a range of symptoms, including palpitations (a feeling of fluttering or racing in the chest), dizziness, shortness of breath, and chest pain. In some cases, arrhythmias can be life-threatening. Thus, early detection of arrhythmia is essential. If left untreated, arrhythmia can also lead to more serious complications, such as stroke, heart attack, or even sudden cardiac death. However, if detected early, arrhythmia can often be effectively managed with medication, lifestyle changes, or medical procedures such as ablation or implantation of a pacemaker or defibrillator [4].

* Corresponding author: murattunc0110@gmail.com ORCID numbers of authors [1]0009-0008-4994-3858; [2]0000-0001-8469-5484

In summary, early detection of arrhythmia is important because it can lead to effective management and prevention of serious complications. For the reasons described, there are many studies in the literature. In a work, Güvenir et. al. used VFI5 algorithm which is majority deterministic algorithm for classification of Arrhythmia data set. Arrhythmia data set is also used in this study. By using the VFI Algorithm with 10-fold cross validation technique, the accuracy was obtained as 62%. Authors also used genetic algorithm to make VFI5 Algorithm effective and increased the accuracy to 68%. Moreover, 50% and 53% accuracy were obtained in regarding study, respectively, by applying Naive Bayes and Nearest Neighbors classifiers [3]. In another work, Niazi et. al. analyzed the same data set, Arrhythmia. They classified the data set using SVM and k-NN algorithms. They applied improved F-score and sequential forward search (IFSFS) for the feature selection process. After performing 20-fold cross validation on their presented model, average accuracy of 73.8% in case of KNN, and 68.8% in case of SVM [5]. Isin and his colleague Ozdalili used pre-trained deep learning models for ECG arrhythmia diagnosis. They employed AlexNet as a feature extractor and basic neural network for the classification of the extracted features. They obtained 98.51% accuracy while the testing accuracy is around 92%. In the light of their findings, they claimed that transfer learning can be an efficient automatic cardiac arrhythmia detection method [6]. Sannino and Pietro presented a study of ECG beat classification using MIT–BIH Arrhythmia Database. They proposed a deep learning approach developed by using Tensor Flow and Google deep learning library. Their experimental results show that the accuracy, sensitivity, and specificity metrics of the model are more efficient than the state of the art [7]. Alfaras et. al. proposed an automatic ECG arrhythmia classifier based on machine learning technique known as Echo State Networks which are a type of recurrent neural network (RNN) that have gained popularity in the field of machine learning. They used the MIT-BIH AR and the AHA ECG databases. They declared that the obtained results are comparable with the state of the art [8]. In another study Togaçar M. and his colleagues classified chest CT's by using different ML models and improved the results by integrating MRMR algorithm to their model [9].

One of the most important parameters effecting the success of a machine learning (ML) application is the data set used in the model training process. The reliability, precision and efficiency of ML model is directly dependent on the data set. In order to achieve high performance, the data set with high numbers of features and samples must be pre-processed before the training phase. These improvements in the data set can be achieved by using feature selection methods [10]. Feature selection algorithms reduce the size of the data and make the data set suitable for training phase by choosing the most relevant features in the data set. Yasar Çıklacandır and her colleague reduced their sample size by 90% by using feature selection methods such as; CHI2, MRMR and ReliefF in their work. Authors obtained 83.58% and 89.89% using DT and SVM classifiers respectively. [11]

The main goal of feature selection is to identify the most informative and relevant features which are most strongly associated with the target variable. By reducing the number of features, feature selection can help to improve the accuracy and generalizability of the model, as well as reduce the risk of overfitting [12]. In this paper, an improved feature selection method is proposed achieving higher classification accuracies for the Arrhythmia data set. Three different feature selection methods are applied on the data set for obtaining the most discriminative features. After the feature selection process, ML algorithms are implemented for the classification.

The aim of the study is to determine the most effective combination of the feature selection method and the classification algorithm. The study divided into three parts consists of the following scheme: In the first part, the data set and the methods used in this study are presented. Additionally, the usage of feature selection algorithms and classification processes are explained. In the next section, the findings obtained from the applied methods are presented. In the final section, the obtained results are reveled and recommendations for further studies are given.

This study aimed to effectively diagnose cardiac arrhythmias using the data set from the UCI Machine Learning Repository. By employing different feature selection methods and classification algorithms. Our research experiments led to the development of an improved feature selection method, achieving superior classification accuracies. Notably, the matched selection method consistently outperformed its alternatives, with a maximum accuracy of 81.27% when used with the SVM classifier. A similar result is obtained by Alshamlan H. and his colleagues. They identified and classified genes involved in the progression of the Alzheimer's Disease. Authors used SVM classifier with various feature selection methods such as; MRMR, CFS, CHI2 and F-Score. They achieved classification accuracy of 84% with the MRMR and F-Score filter methods. [13] Despite slightly longer completion times, our proposed method demonstrated remarkable accuracy and robustness, offering promising implications for improving cardiac arrhythmia diagnosis.

2. MATERIALS AND METHODS

2.1. Data Set: The purpose of this study is to classify the presence and the types of cardiac arrhythmia disease by using an ECG data set created from various patients. To serve this purpose, the data set taken from UCI Machine Learning Repository, contains ECG recordings of 452 patients. There are 279 features for all measurements, represented by the columns of the data set. Labels of the classes from 1 to 16 in the 280th column is representing the type of arrhythmia patients encounter [14]. The data set contains missing measurements. Regarding missing data is handled by setting as zero. The most occurring three classes in the data set are normal rate, right bundle branch block and coronary artery disease with their respective occurrence rates are 54.2, 11.1, 9.7 per hundred subjects. All the features of the data set are given in Table 1. below.

Table 1. Features in the data set

Number of Feature	Features	Number of Feature	Features
1	f_1 : Age	19	f_{19} : R' wave, small peak just after R
2	f_2 : Sex	20	f_{20} : S' wave
3	f_3 : Height	21	f_{21} : Number of intrinsic deflections, linear
4	f_4 : Weight	22	f_{22} : Existence of ragged R wave, nominal
5	f_5 : QRS duration	23	f_{23} : Existence of diphasic derivation of R wave, nominal
6	f_6 : P-R interval	24	f_{24} : Existence of ragged P wave, nominal
7	f_7 : Q-T interval	25	f_{25} : Existence of diphasic derivation of P wave, nominal
8	f_8 : T interval	26	f_{26} : Existence of ragged T wave, nominal
9	f_9 :P interval	27	f_{27} : Existence of diphasic derivation of T wave, nominal
10	f_{10} : QRS	160.	f_{160} : JJ wave, linear
11	f_{11} : T	161.	f_{161} : Q wave, linear
12	f_{12} : P	162.	f_{162} : R wave, linear
13	f_{13} : QRST	163.	f_{163} : S wave, linear
14	f_{14} : J	164.	f_{164} : R' wave, linear
15	f_{15} : Heart rate	165.	f_{165} : S' wave, linear
16	f_{16} : Q wave	166.	f_{166} : P wave, linear
17	f_{17} : R wave	167.	f_{167} : T wave, linear
18	f_{18} : S wave		

All of the above features from 16 to 27 measured for the DI channel are also measured for the DII channel (f_{28}-f_{39}), DIII channel (f_{40}-f_{51}), AVR (Augmented Vector Right) channel (f_{52}-f_{63}), AVL (Augmented Vector Left) channel (f_{64}-f_{75}), AVF (Augmented Vector Foot) channel (f_{76}-f_{87}), V1 channel (f_{88}-f_{99}), V2 channel (f_{100}-f_{111}), V3 channel (f_{112}-f_{123}), V4 channel (f_{124}-f_{135}), V5 (f_{136}-f_{147}), V6 channel (f_{148}-f_{159}). AVL, AVR, AVF, V1, V2, V3, V4, V5, and V6 are the names of the channels used in a standard 12-lead ECG to record the electrical activity of the heart from different perspectives. AVL is a modified chest lead that provides a view of the heart's electrical activity from the left lateral perspective. Likewise, AVR provides this activity from a right side. AVF is a modified limb lead that provides a view of the heart's electrical activity from the foot or inferior perspective. V1 to V6 are chest leads that are placed on specific locations on the chest to provide a view of the heart's electrical activity from different angles. Together, these leads or channels allow healthcare professionals to assess the electrical activity of the heart from multiple perspectives, which helps them diagnose a wide range of cardiac conditions, including arrhythmias, and other abnormalities. The following measurements from 160 to 167 are obtained from the DI channel. Features 168 (f_{168}) and feature 169 (f_{169}) named QRSA and QRSTA, respectively, are derived from some of the metrics detailed as follows:

QRSA = sum of areas of all segments divided by 10. (Area = width × height ÷ 2)

QRSTA = QRSA + 0.5 × width of T wave × 0.1 × height of T wave. (If T is diphasic then the bigger segment is considered).

The features from 160 to 167 measured for the DI channel are also measured for the DII channel (f_{170}-f_{179}), DIII channel (f_{180}-f_{189}), AVR channel (f_{190}-f_{199}), AVL channel (f_{200}-f_{209}), AVF channel (f_{210}-f_{219}), V1 channel (f_{220}-f_{229}), V2 channel (f_{230}-f_{239}), V3 channel (f_{240}-f_{249}), V4 channel (f_{250}-f_{259}), V5 (f_{260}-f_{269}), V6 channel (f_{270}-f_{279}).

2.2. Feature Selection: The data set contains many features, and using all of these features can lead to long processing times, incorrect classification and low prediction results. It may be sufficient to consider only some of the features in determining the disease that the patient will encounter. The selection of relevant features is an important step in machine learning as it can reduce the complexity of the model and increase its accuracy, interpretability, and efficiency [15]. There are several methods for feature selection, and they can be classified into three categories: filter methods, wrapper methods, and embedded methods.

2.2.1. Wrapper Methods: Wrapper methods select features by evaluating the performance of a machine learning model with different subsets of features. They can select the most relevant features and account for the interaction between features. However, they can be computationally expensive and prone to overfitting. Examples of wrapper methods include recursive feature elimination, forward selection, and backward elimination [16].

2.2.2. Embedded Methods: These methods select features during the training of a ML model. They can be computationally efficient and select the most relevant features for a specific model. However, they may not always select the most relevant features for other models, and they can be sensitive to the choice of the model. Examples of embedded methods are Lasso Regularization, Ridge Regularization, and Decision Tree-Based feature selection [17].

2.2.3. Filter Methods: Filter methods select features based on their statistical properties without considering the performance of the model. They can be fast and efficient. However, like the embedded methods, they may not always select the most relevant features. Examples of filter methods include chi-squared, mutual information, correlation-based feature selection, and variance threshold [17].

In a nutshell, filter methods are computationally efficient but may not always select the most relevant features. Wrapper methods can select the most relevant features but can be computationally expensive. Embedded methods can be efficient and select the most relevant features for a specific model but may not always be suitable for other models. Therefore, two different filter methods and a purposed hybrid approach are performed in the paper. Feature selection algorithms are applied before the classification process, such as a pre-processing filter operation. In this study, three different feature selection algorithms were applied including Minimum Redundancy Maximum Relevance, Chi-square, and Matched Feature Selection Method. These methods are briefly described in the following subsection.

Minimum Redundancy Maximum Relevance (MRMR) Feature Selection Method: This method creates the smallest set of features which consists of highly related features with class labels. This algorithm calculates relevance and redundancy of the respective feature with class labels and create a score list by using these calculations. In every loop, algorithm reduce the size of the data set until it reaches the determined set size or maximum number of repetitions. Therefore, processing time of the algorithm increases as the desired size of the data set is decreased [18].

Chi-square Feature Selection (Chi2) Method: Chi-Square test to all features and create list of features with their respective scores. Desired features can be obtained via setting a number of features or setting threshold score to take all features with higher score values. The processing time of this method is independent from the desired number of features as it calculates for all features once [19].

Matched Feature Selection (MFS) Method: This purposed method combines the two different methods mentioned the previous subsections: MRMR and Chi-square. The algorithm works by calculating both MRMR and Chi-square test scores for all of the features once. After this stage, algorithm calculates the mean scores for all of the features. Finally, as in the second method, a threshold can be set for selecting the relevant features for prediction. This algorithm provides a robust set of features, also has a similar completion time to Chi-square test

and similar performance to the MRMR algorithm. All of the algorithms have the same inputs including features, labels and the threshold level that can be set by the user. Output of the algorithms are the first set scores assigned to each feature. Number of features they selected according to the threshold level, the reduced data set is created with the important features. Afterwards, the created data set is employed for the train and the test procedures of the selected ML models.

2.3. Classification: Machine learning classification is a process of predicting a categorical label or class for a given input based on the patterns and relationships learned from a labeled training dataset. The classification process involves several steps:

- Data preparation
- Feature selection
- Model selection.
- Training, validation, and test the selected model.

Overall, the purpose is to create an accurate and efficient model which can generalize well to unseen data and automate the classification task. The classification process is carried out by using the ML models which described in the following subsections.

2.3.1. k-Nearest Neighbors (k-NN): The k-NN algorithm is one of the easiest and simplest algorithms for classification and prediction tasks. k-NN classifies new data based on its respective distances with the other samples. In order to utilize k-NN algorithm, it is necessary to determine the number of the neighbors which is demonstrated as 'k'. The key point of the algorithm is the selection of the k value, which directly affects the performance of the output [20].

Algorithm calculates all the distance between new data and each existing sample, these distances are listed and sorted from low to high. First k samples are considered from the list. Number of k class labels are counted, and new data are assigned to a class with the highest number of a certain class labels.

The algorithm is a distance based and there are numerous ways to calculate distances. However, the most common method is Euclidian Distance calculation and normalization. The visual representation of the algorithm can be seen in the Figure 1.

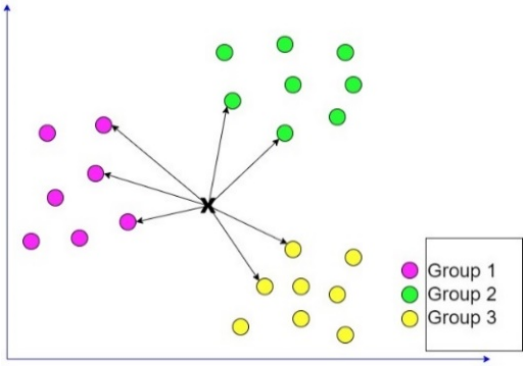

Figure 1. Visual representation of the k-NN algorithm.

2.3.2. Support Vector Machines (SVM): Support Vector Machines Algorithm is capable of determining difference between the samples. SVM uses decision functions to classify new data. Decision functions are created during the training process of the model [21]. SVM uses different methods for classification of multi-class data sets. In this study, two different SVM classification methods are concerned: one vs. one (OvO), and one vs. all (OvA). One vs. one considers only two classes while creating the support vectors which helps to separate the classes. For every combination of two classes, there must be a different decision vector in order to separate classes from each other. On one hand, this method results in high precision of classification and prediction rate. On the

other hand, process needs a high requirement of processing time since every two class combinations require a decision vector, particularly in a great number of classes.

The second method, one vs all, considers all classes while creating the support vectors. Thus, every class requires only one decision function. This method results in low precision of classification and prediction, also high chance of errors in the process. However, all classes need only one decision function completion time of this model would be lower than the first method. Main differences between these methods are their completion time and the precision. Both algorithms classify the new data by inserting the data in the decision functions created during the training process.

2.3.3. Decision Trees (DTs): This algorithm is based on the comparison of class parameters. This algorithm is used tree-like structures for classifying the new data [22]. The algorithm uses root-node-leaf relation. Every leaf of the tree is corresponding to a class label. Nodes of the tree are representing the features of the data set. For the root node feature which has the highest relation with the class label must be used.

Entropy and gain calculations are utilized in the algorithm in order to determine the level of feature – class relation. First entropy of the class labels is calculated, then entropy of all features with the class labels are computed. Finally, the gain is calculated by subtracting the second value from the first. Feature with the highest gain is used for the regarding specific node. This algorithm mostly uses with great amounts of samples in the data set and have small amounts of features. Structure of the algorithm is visualized in the Figure 2.

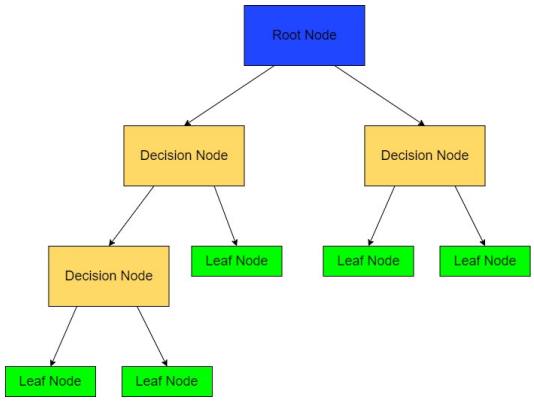

Figure 2. Structure of the Decision Tree Algorithm.

2.4. Proposed Feature Selection Method: The proposed model operates by subjecting the dataset to the MRMR and CHI2 feature selection algorithms. These algorithms compute relevancy score values for each feature, which are subsequently normalized. After normalization, these scores are multiplied by specific coefficients to create a combined value. This combined value is compared to the predetermined threshold value. Those exceeding the threshold value are stored to include in the newly created dataset. Finally, a new dataset is created with the original ranks and contents of the stored features. This dataset, along with the number of features, is returned as output to the main function.

3. FINDINGS AND DISCUSSION

Purpose of this study is to determine the risk of cardiac arrhythmias based on the ECG data from the patients. For realizing this purpose, experimental setup is created by implementing all described algorithms. A block diagram consists of the functions created and their respective input and output parameters also the interconnections between them are given Figure 3.

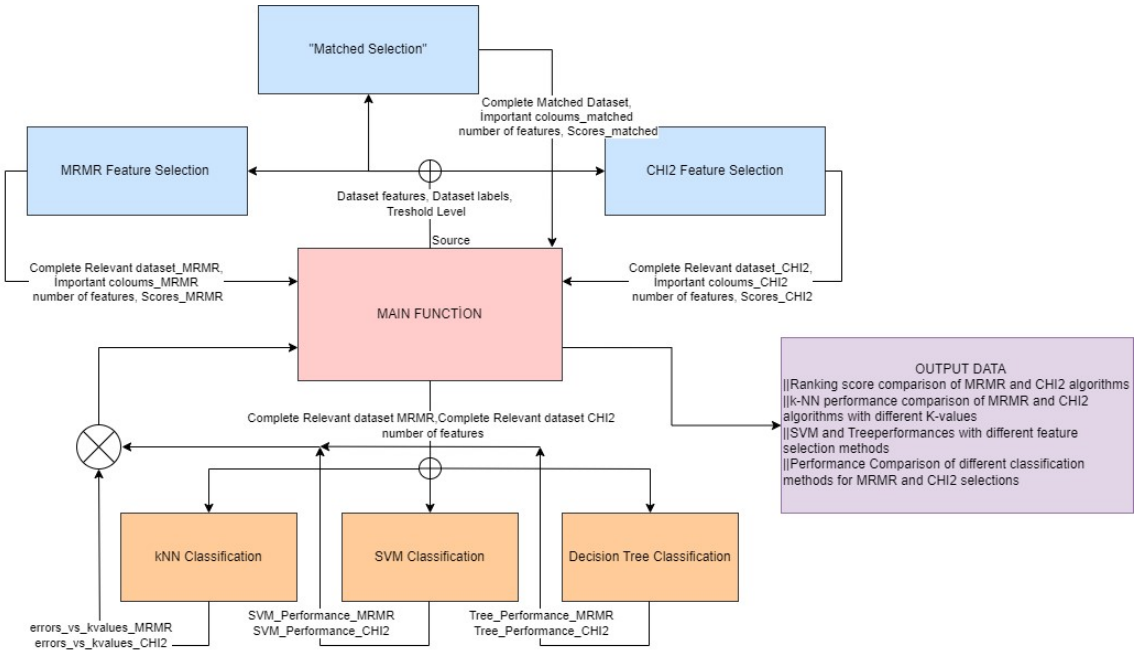

Figure 3. Block diagram of the experimental setup.

Upon successful implementation of the code and addition of a result evaluation section, the setup undergoes testing. The test results for a threshold level of 0.35 are given in the Figure 4 and Figure 5.

The test procedure is repeated for various threshold values, resulting in varying success rates. The table consisting of the results from different tests is created and it can be seen in Table 2. Using the results from the table, performance comparison graphs are created and illustrated in Figure 6, Figure 7, and Figure 8 for the k-NN, SVM and DT algorithms respectively. Comparison of different feature selection algorithms for each classification algorithm can be seen in those figures.

While the model tested with different threshold values, the completion time of feature selection algorithms were recorded to use in the performance evaluation. These records can be seen in the Table 3.

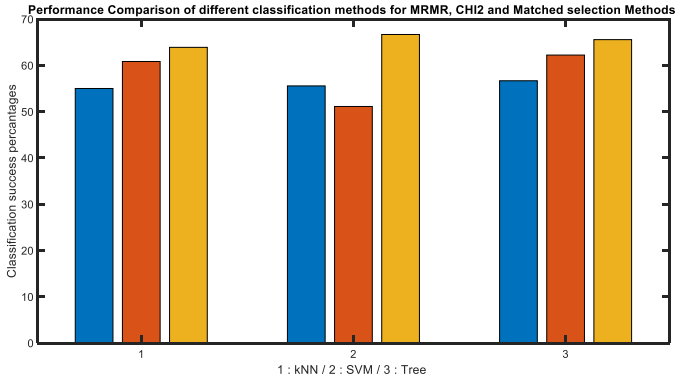

Figure 4. Accuracy of the classifiers for different feature selection algorithms.

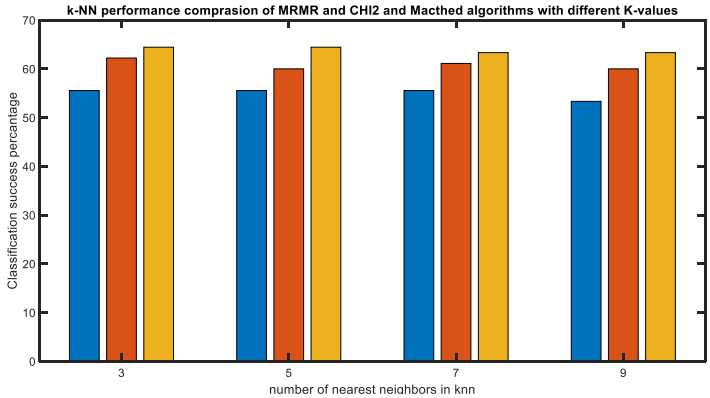

Figure 5. Accuracy of k-NN classifier for the different feature selection algorithms and different k values.

Table 2. Results of the algorithm tests

Threshold level	Completion time(Total)	K-NN			SVM			DT		
		MRMR	CHI2	Matched	MRMR	CHI2	Matched	MRMR	CHI2	Matched
0,1	5,33	55,87	44,99	51,72	55,47	62,94	62,28	59,00	55,00	63,34
0,15	5,06	47,36	53,41	58,67	56,71	51,82	60,81	50,65	59,28	57,60
0,2	4,87	54,51	52,18	60,66	53,89	64,67	69,49	54,40	55,47	57,71
0,25	5,52	55,25	59,00	49,27	66,50	59,12	58,80	57,00	55,95	68,60
0,3	5,06	47,79	59,42	53,79	46,35	66,82	70,96	54,77	64,07	73,24
0,35	4,89	60,80	50,94	52,37	51,17	54,31	72,10	59,28	63,38	64,09
0,4	4,94	53,07	58,67	69,66	53,84	53,84	74,56	49,87	53,78	73,34
0,45	5,47	49,25	47,00	60,37	47,02	62,36	81,27	56,71	55,73	53,71
0,5	5,05	52,18	51,67	59,59	59,28	49,06	70,88	51,00	57,00	58,66
0,55	5,77	59,15	48,53	68,43	56,05	54,97	69,03	46,48	52,41	60,56
0,6	5,97	57,09	55,28	62,65	49,09	50,13	64,09	44,00	54,75	64,09
0,65	7,79	50,81	43,73	63,60	46,48	54,39	74,20	51,42	50,44	73,02
0,7	9,04	50,40	59,46	66,21	55,95	52,78	68,84	46,40	54,13	73,50
0,75	7,79	48,56	56,32	49,50	59,93	54,77	63,80	44,00	54,75	46,20
0,8	8,68	53,15	54,39	56,53	68,20	52,70	59,73	53,40	46,48	51,20
0,85	10,25	52,80	49,38	68,95	54,40	68,27	63,01	44,44	49,60	51,12
0,9	9,75	59,68	59,93	53,88	59,34	59,34	75,38	42,47	50,56	55,35
0,95	7,72	49,60	53,48	57,00	61,43	54,97	52,78	52,78	48,55	55,95
1	4,85	52,23	36,03	51,42	44,98	56,22	59,45	49,06	44,98	47,55
Mean		53,13	52,31	58,65	55,06	57,03	66,92	50,90	54,02	60,46
Standart Deviation		3,92	6,17	6,46	6,40	5,46	6,99	5,11	4,87	8,62
Maximum Value		60,80	59,93	69,66	68,20	68,27	81,27	59,28	64,07	73,50

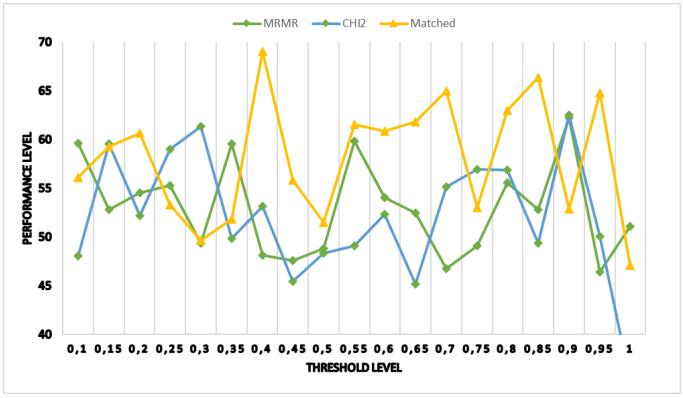

Figure 6. Performance of different feature selection algorithms with k-NN classification algorithm.

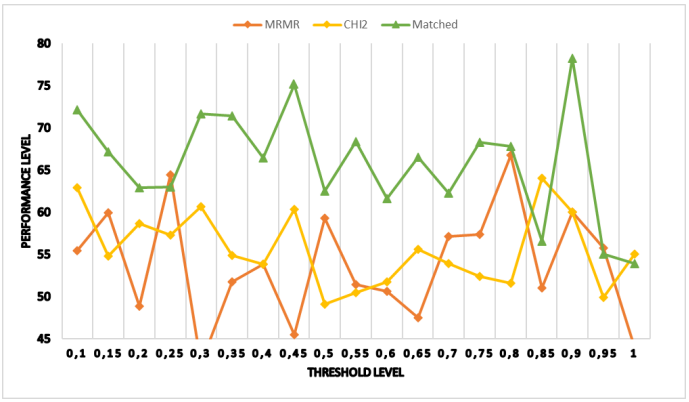

Figure 7. Performance of different feature selection algorithms with SVM classification algorithm.

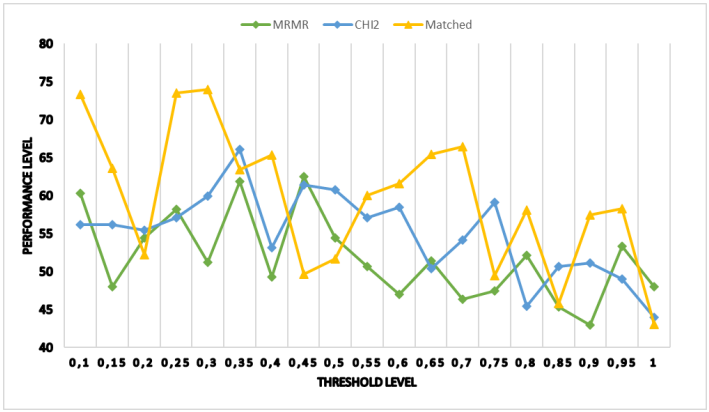

Figure 8. Performance of different feature selection algorithms with DT classification algorithm.

Table 3. Completion time of individual feature selection algorithms

Threshold level	Completion time(Total)	Completion time(Matched)	Completion time(MRMR)	Completion time(CHI2)
0,1	5,33	0,6179	0,2791	0,0386
0,15	5,06	0,5984	0,216	0,0377
0,2	4,87	0,5647	0,3237	0,0422
0,25	5,52	0,5745	0,2067	0,0378
0,3	5,06	0,5597	0,2103	0,0369
0,35	4,89	0,7307	0,2185	0,0389
0,4	4,94	0,5703	0,2026	0,0384
0,45	5,47	0,6006	0,1942	0,0368
0,5	5,05	0,5729	0,0264	0,038
0,55	5,77	0,5768	0,2057	0,0379
0,6	5,97	0,5748	0,2038	0,0371
0,65	7,79	0,57	0,2039	0,0375
0,7	9,04	0,5732	0,2022	0,0373
0,75	7,79	0,5646	0,2123	0,0371
0,8	8,68	0,6105	0,2132	0,0375
0,85	10,25	0,5635	0,2084	0,0374
0,9	9,75	0,6238	0,2199	0,0416
0,95	7,72	0,5576	0,2007	0,038
1	4,85	0,5668	0,2123	0,0366

Performance of the system is evaluated using the accuracy and algorithm completion time. The visualization of this evaluation metrics is depicted in the Figure 9, Figure 10, and Figure 11.

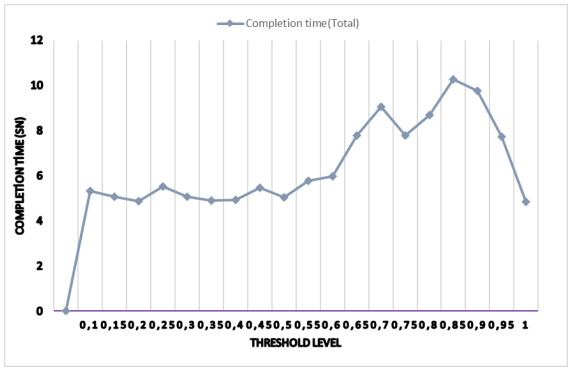

Figure 9. Variation of total completion time according to threshold level.

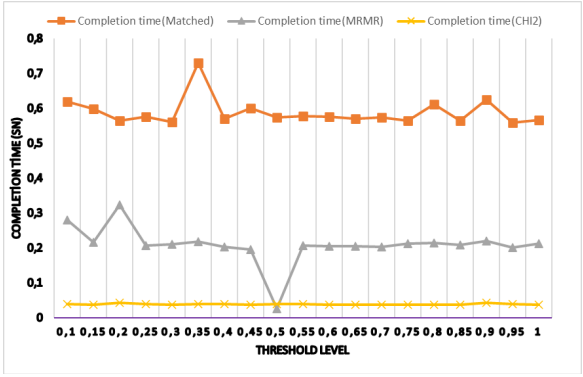

Figure 10. Change in individual completion time of feature selection algorithms with threshold level.

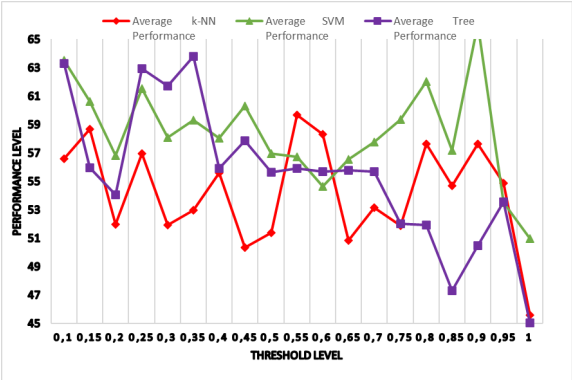

Figure 11. Average classifier performances for all the feature selection algorithms.

As it is stated before Figure 6, Figure 7, and Figure 8 reveals differences between performance of the feature selection algorithms with all classification algorithms.

For k-NN classification method MRMR and CHI2 algorithms shows almost identical performance in terms of mean accuracy. However, the proposed matched selection algorithm outperforms these two algorithms by an average of 6%. Considering standard deviation, the MRMR feature selection method is the most precise method. When considering both accuracy and precision simultaneously, the most suitable model to use with k-NN classification is the proposed matched selection model. Upon comparing the feature selection algorithms for the SVM classification model, it is evident that the most accurate feature selection algorithm is matched selection. Similar to results of the k-NN algorithm MRMR and CHI2 feature selection algorithms show close performance. Again, matched selection algorithm outperforms both algorithms by 11%. When considering precision, the CHI2 method outperforms other methods. The matched selection algorithm demonstrates similar precision performance when used with k-NN and SVM classification algorithms. Upon evaluating both accuracy and precision, the matched selection algorithm provides superior results compared to other feature selection algorithms. Additionally, it can be said that. Matched selection algorithm performs best when paired with the SVM classification algorithm. When the performance of feature selection algorithms used with the DT classification algorithm is investigated, it is apparent that the matched selection algorithm outperforms other feature selection algorithms by approximately 7% in the matter of accuracy. However, on the subject of precision matched feature selection algorithms algorithm demonstrates significantly inferior performance. The most precise feature selection model to use with DT classification is CHI2 method. Based on these it can be said that matched selection algorithm is not the most suitable model to use with DT classification algorithm.

Reviewing the information provided in Table 1, it can be concluded that the matched selection algorithm demonstrates superior performance in terms of accuracy across all classification algorithms. Moreover, proposed model generally exhibits precision performance comparable to others except when used in conjunction with the DT algorithm. Consequently, it can be asserted that the model demonstrates a noticeable superior performance compared to other models in this scenario.

Based on the data presented in Table 2, the performance of feature selection algorithms can be evaluated based on their completion times. The proposed matched selection model takes approximately 2.5 times longer to complete compared to the MRMR model and about 15 times longer compared to the CHI2 model. The completion time of all models, including the proposed model, generally decreases as the threshold level increases. Under these circumstances, it would be more sensible to use the proposed model in situations requiring high accuracy, such as healthcare, rather than time-sensitive scenarios.

The proposed matched selection algorithm combines various feature selection methods, making it versatile and robust. It offers comprehensive selection criteria, enhancing adaptability across tasks. Moreover, it consistently achieves high accuracy, especially when used with k-NN and SVM methods, making it ideal for accuracy-focused applications. Overall, the proposed algorithm represents an advancement, offering higher accuracy, with the potential to enhance classification systems in various domains.

The proposed model combines various existing feature selection algorithms to form a hybrid algorithm. This approach enables achieving higher accuracy at the expense of completion time. This innovative methodology leverages the strengths of individual algorithms while mitigating their limitations, resulting in a more comprehensive and effective solution to feature selection. Emphasizing higher accuracy, at the expense of completion time, underlines the innovative nature of this study. In summary, the proposed model represents a significant advancement in feature selection, promising improved performance.

4. CONCLUSION

The objective of this study is to determine the presence and classify the types of cardiac arrhythmia disease effectively by using the ECG data set taken from UCI Machine Learning Repository. To achieve this purpose, different feature selection methods and classification algorithms are engaged. Determining the most effective combination of the feature selection method and the classification algorithm for the data set, named Arrhythmia, is the motivation of the work.

In this paper, an improved feature selection method that achieves higher classification accuracies for the Arrhythmia data set is proposed. Besides, different feature selection methods are applied on the data set for obtaining the most discriminative features. After the feature selection process, machine learning algorithms are implemented for the classification task.

Experimental evaluation focused on determining the performance of feature selection methods and classification algorithms in predicting arrhythmias. The matched selection feature method consistently outperformed alternatives in terms of accuracy with different classification algorithms, achieving superior accuracy of 81.27% when used with SVM classifier. In addition to accuracy and precision, computational efficiency was evaluated by recording completion times of induvial feature selection methods. Despite slightly longer completion times, the proposed matched selection method's superior accuracy and robustness make it a viable choice for real-world applications prioritizing accuracy rather than time-sensitivity.

References

[1] Krikler DM. "Historical aspects of electrocardiography." Cardiol Clin, vol. 5, no. 3, pp. 349-355, Aug. 1987.

[2] Zimetbaum PJ, Josephson ME. "Use of the electrocardiogram in acute myocardial infarction." N Engl J Med, vol. 348, no. 10, pp. 933-940, Mar. 06, 2003.

[3] Güvenir HA, Acar B, Demiroz G, Cekin A. "A supervised machine learning algorithm for arrhythmia analysis." Computers in Cardiology 1997, pp. 433-436.

[4] Fu, Dg. "Cardiac Arrhythmias: Diagnosis, Symptoms, and Treatments." Cell Biochem Biophys, vol. 73, pp. 291–296, 2015. DOI: https://doi.org/10.1007/s12013-015-0626-4.

[5] Niazi KAK, Khan SA, Shaukat A, Akhtar M. "Identifying best feature subset for cardiac arrhythmia classification." In Proceedings of the 2015 Science and Information Conference, SAI 2015, London, UK, 28–30 July 2015, pp. 494–499.

[6] Isin, A., Ozdalili, S. "Cardiac arrhythmia detection using deep learning." Procedia Computer Science, vol. 120, pp. 268-275, 2017. DOI: https://doi.org/10.1016/j.procs.2017.11.238.

[7] Sannino, G., De Pietro, G. "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection." Future Generation Computer Systems, vol. 86, pp. 446-455, 2018. DOI: https://doi.org/10.1016/j.future.2018.03.057.

[8] Alfaras, M., Soriano, M.C., Ortín, S. "A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection." Frontiers in Physics, vol. 7, 2019. DOI: https://doi.org/10.3389/fphy.2019.00103.

[9] Toğaçar, M., Ergen, B., Cömert, Z. "Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks." Biocybernetics and Biomedical Engineering, vol. 40, no. 1, pp. 23-39, 2020.

[10] Güney, S., Ergün, G.B. "Classification of Canine Maturity and Bone Fracture Time Based on X-Ray Images of Long Bones." IEEE Access, vol. 9, pp. 109004-109011, 2021. DOI: 10.1109/ACCESS.2021.3101040.

[11] Çıklaçandır Y., Karabiber Cura F., Özlem O. "A Comparative Study on Different Feature Selection Methods for Malaria Detection." 1-4, 2023. DOI: 10.1109/TIPTEKNO59875.2023.10359193.

[12] Ergün GB, Güney S. "A Comparison of the Multivariate Calibration Methods with Feature Selection for Gas Sensors' Long-Term Drift Effect." Uluslararası Teknolojik Bilimler Dergisi, c. 11, sayı. 3, ss. 170-176, Ara. 2019.

[13] Alshamlan, H., Omar, S., Aljurayyad, R., Alabduljabbar, R. "Identifying Effective Feature Selection Methods for Alzheimer's Disease Biomarker Gene Detection Using Machine Learning." Diagnostics, 13, 1771, 2023. DOI: 10.3390/diagnostics13101771.

[14] Newman D, Hettich S, Blake C, Merz C (1998). "UCI Repository of machine learning databases." http://www.ics.uci.edu/~mlearn/MLRepository.html.

[15] Ergün GB, Güney S. "A Comparison Study for Image Classification and Feature Selection." 4th International Conference on Computational Mathematics and Engineering Sciences, Antalya, 20-22 April 2019.

[16] Chen G., Chen J. "A novel wrapper method for feature selection and its applications." Neurocomputing, vol. 159, pp. 219-226, 2015. DOI: https://doi.org/10.1016/j.neucom.2015.01.070.

[17] Chandrashekar G., Sahin F. "A survey on feature selection methods." Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28, 2014. DOI: https://doi.org/10.1016/j.compeleceng.2013.11.024.

[18] Bugata P., Drotar P. "On some aspects of minimum redundancy maximum relevance feature selection." Sci. China Inf. Sci., vol. 63, p. 112103, 2020. DOI: https://doi.org/10.1007/s11432-019-2633.

[19] Thaseen IS., Kumar CA. "Intrusion detection model using fusion of chi-square feature selection and multi class SVM." Journal of King Saud University - Computer and Information Sciences, vol. 29, no. 4, pp. 462-472, 2017. DOI: https://doi.org/10.1016/j.jksuci.2015.12.004.

[20] Cover TM., Hart PE. "Nearest neighbor pattern classification." IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967. DOI: 10.1109/TIT.1967.1053964.

[21] Cristianini N., Ricci E. "Support Vector Machines." In: Kao MY. (eds) Encyclopedia of Algorithms. Springer, Boston, MA, 2008. DOI: https://doi.org/10.1007/978-0-387-30162-4_415.

[22] Quinlan JR. "Induction of Decision Trees." Mach. Learn., vol. 1, no. 1, pp. 81-106, Mar. 1986.