Journal of Big Data

RESEARCH

Open Access

Analysis and classification of heart diseases using heartbeat features and machine learning algorithms

Fajr Ibrahem Alarsan[1*] and Mamoon Younes[2]

*Correspondence: fajr.ib.alarsan@gmail.com
[1] Informatics and Decision Supporting Systems, Higher Institute for Applied Sciences and Technology, Damascus, Syria
Full list of author information is available at the end of the article

Abstract

This study proposed an ECG (Electrocardiogram) classification approach using machine learning based on several ECG features. An electrocardiogram (ECG) is a signal that measures the electric activity of the heart. The proposed approach is implemented using ML-libs and Scala language on Apache Spark framework; MLlib is Apache Spark's scalable machine learning library. The key challenge in ECG classification is to handle the irregularities in the ECG signals which is very important to detect the patient status. Therefore, we have proposed an efficient approach to classify ECG signals with high accuracy Each heartbeat is a combination of action impulse waveforms produced by different specialized cardiac heart tissues. Heartbeats classification faces some difficulties because these waveforms differ from person to another, they are described by some features. These features are the inputs of machine learning algorithm. In general, using Spark–Scala tools simplifies the usage of many algorithms such as machine-learning (ML) algorithms. On other hand, Spark–Scala is preferred to be used more than other tools when size of processing data is too large. In our case, we have used a dataset with 205,146 records to evaluate the performance of our approach. Machine learning libraries in Spark–Scala provide easy ways to implement many classification algorithms (Decision Tree, Random Forests, Gradient-Boosted Trees (GDB), etc.). The proposed method is evaluated and validated on baseline MIT-BIH Arrhythmia and MIT-BIH Supraventricular Arrhythmia database. The results show that our approach achieved an overall accuracy of 96.75% using GDB Tree algorithm and 97.98% using random Forest for binary classification. For multi class classification, it achieved to 98.03% accuracy using Random Forest, Gradient Boosting tree supports only binary classification.

Keywords: Heartbeats classification, Electrocardiogram (ECG), Machine-learning libraries (MLlib), Spark–Scala

Introduction

An electrocardiogram (ECG) is a complete representation of the electrical activity of the heart on the surface of the human body, and it is extensively applied in the clinical diagnosis of heart diseases [1], it can be reliably used as a measure to monitor the functionality of the cardiovascular system. ECG signals have been widely used for detecting heart diseases due to its simplicity and non-invasive nature. Features of ECG signals can be computed from ECG samples and extracted using some softwares (ex: Matlab). For instance, millions of people suffer from irregular heartbeats which can be lethal in some cases. Therefore, accurate and low-cost diagnosis of arrhythmic heartbeats is highly desirable [2]. Many studies have developed arrhythmia classification approaches that use automatic analysis and diagnosis systems based on ECG signals. The most important factors for the analysis and diagnosis of cardiac diseases are features extraction and beats classification. Numerous techniques for classifying ECG signals were proposed in recent years and good results achieved [3–5].

The performance of ECG pattern classification strongly depends on the characterization power of the features that are extracted from the ECG signal and the design of the classifier (classification model).

Automated classification of heartbeats has been previously reported by many investigators using a variety of features to represent the ECG and a number of classification methods. In general, heartbeat features include ECG morphology, heartbeat interval features (temporal features), beats correlations and summits values [6].

The target of classification process is obtaining an intelligent model, that is capable to class any heartbeat signal to specific type of heartbeats. Experiments have been conducted on the well-known MIT-BIH Arrhythmia database using obtained model, and results have been compared with the previous scientific literature. The final results show that our model is not only more efficient than related works in terms of accuracy, but also competitive in terms of sensitivity and specificity.

Big data analytic plays a vital role in managing the huge amount of health-care data and improving the quality of health-care services offered to patients. In this context, one of the challenges lies in the classification of data, which relies on effectively distributed processing platforms, advanced data mining and machine learning techniques. Therefore, a Big data technique is introduced in this work to meet the challenges faced by classify the ECG beats. Recently, deep learning techniques have been used by many companies, including Facebook, Google, IBM, Microsoft, NEC, Netflix, and NVIDIA [7, 8], and in a very large set of application domains such as customer churn prediction in telecom company [9]. In this paper, a novel deep learning approach for ECG beats classification is presented.

Background and related work

There are many works related to ECG classification without using big data tools when size of dataset is not large. On other hand, there are several studies that depend on big data techniques. In [10], Indonesia has high mortality caused by cardiovascular diseases. To minimize the mortality, a tele-ecg system was built for heart diseases early detection and monitoring using Hadoop framework, in order to deal with big data processing. The system can classify the ECG data using decision tree (DT) and random forest (RF), it was the first real system for heartbeats classification using big data tools. The system was build on cluster computer with 4 nodes. The server was able to handle 60 requests at the same time. The accuracy was 97.14% and 98,92% for decision tree and random forest respectively. In [11], Neural networks and dimensionality reduction technique was used and the approach was tested on the Massachusetts Institute of Technology arrhythmia database. The classification performance on a test set of only 18 ECG records of 30 min each achieved an accuracy of 96.97%. In [1], Many types of heartbeat were extracted and used for classification, classification method is used to classify independent type (3 records for each type); each type of heartbeats has its own model (model to classify normal heartbeats, model to classify type 1 of heartbeats and so forth). Neural network and SVM were applied and the accuracy of results was high good (more than 90%), but there was not unified model to classify all multi-types together at once. In [12], Feed-forward and fully connected artificial neural networks aided by particle swarm optimization technique are employed to recognize two patterns of heartbeats [Ventricular ectopic beats (VEBs) and supra-VEBs (SVEBs)]. Tuning parameters of proposed method has improved accuracy of classification to 96% comparing to the same method with default value of parameters. Not all features were used (only morphological and temporal) and a total of 83,648 beats were selected for training and testing.

In [13], Hidden Markov Models and Spark were used to mine ECG data, combining accurate Hidden Markov Model (HMM) techniques with Apache Spark to improve the speed of ECG analysis. The paper has proven that there is potential for developing a fast classifier for heartbeats classification.

In [8], DNN (Deep Neural Network) has been used for deep learning to classify heartbeats. The author has compared results with many studies, accuracy of classifying has reached 99%, but the classification was only two types (Normal and Abnormal) and dataset size was almost 85,000 records.

In [2], multiply types of heartbeats have been studied and the author has reached accuracy 93.4%. Convolutional neural network for classification of ECG beat types has been developed by the author.

All studies have proven that machine learning algorithms are very effective in heartbeats classification.

Objective of the paper

In this paper, multiples classifiers are proposed for ECG classification, these classifiers are used mostly in Big Data and Machine Learning fields by the weighted voting principle. Each classifier influences the final decision according to its performance on the training data. Parameters of each classifier are adjusted on the basis of an individual classifier's performance on the training data by applying the pseudoinverse technique. The proposed approach is validated in the MIT BIH Arrhythmia Database. The classification performance was validated on a set of 51 ECG records with different temporal length. So our work is distinguished by:

- Number of tested records (205,146 records of 51 patients).
- Complexity of heartbeat types in training and testing (training records contains Normal and Abnormal beats).
- Using Machine learning algorithms for classification.
- Using big data tool (Spark–Scala).
- Using local host pc (according to the lack of requirements).
- Binary and Multi Classification.

In general, previous studies are using known methods (SVM, NN, PCA, Adaptive methods, etc.) and limited number of records for testing and training.

Methods

Experimental study for whole work will be introduced in the following points:

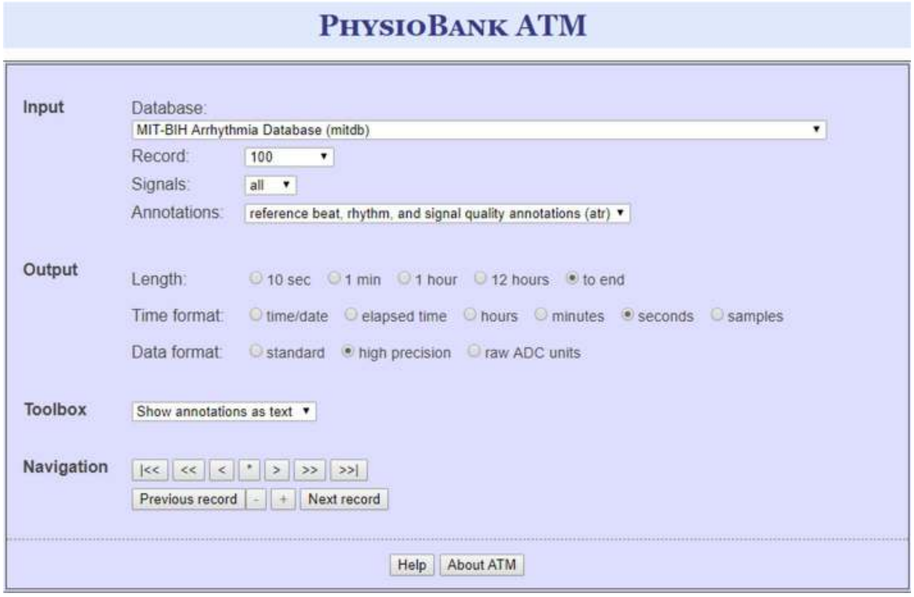
Heartbeat dataset

Data set preparing

The development of this work requires a database with digital ECG records for computational analysis of many different patients with different pathologies. Accordingly, we employed the widely known Massachusetts Institute of Technology (MIT) arrhythmia database. The original dataset is the MIT-BIH Arrhythmia Dataset. Using physionet ATM Bank [14], to get records' annotations with specific configuration as shown in Fig. 1.

As shown in Fig. 1, each record was extracted to its end (ex: record 100 is 1805 s). In general, completed dataset contains information of 51 patients (totally 205,146 rows for all patients). Each row has 16 columns of features. Each record of data has three main files with three different extentions .atr, .hea and .dat, atr file contains annotations, dat file is the digitalized signals and hea file is header file. The recordings were digitized at 360 samples per second per channel with 11-bit resolution over a 10 mV range. Two or more cardiologists independently annotated each record; disagreements were resolved to obtain the computer-readable reference annotations for each beat included with the database. More details about dataset in [15]. Annotations was saved as text files from Physionet website. Many of the considered features based on the Discrete Wavelet Transform (DWT) of the continuous ECG signal, final dataset was saved as csv file.

In addition to the ECG signal, annotations contain the beat localization and the beat class [16]. Normal beat pattern is shown in Fig. 2. Firstly, two classes (Normal and Abnormal) were classified, then multi classes (4 classes) were classified.



PhysioBank ATM

Input

Database: MIT-BIH Arrhythmia Database (mitdb)

Record: 100

Signals: all

Annotations: reference beat, rhythm, and signal quality annotations (atr)

Output

Length: ○ 10 sec ○ 1 min ○ 1 hour ○ 12 hours ● to end

Time format: ○ time/date ○ elapsed time ○ hours ○ minutes ● seconds ○ samples

Data format: ○ standard ● high precision ○ raw ADC units

Toolbox: Show annotations as text

Fig. 1 Physionet configuration

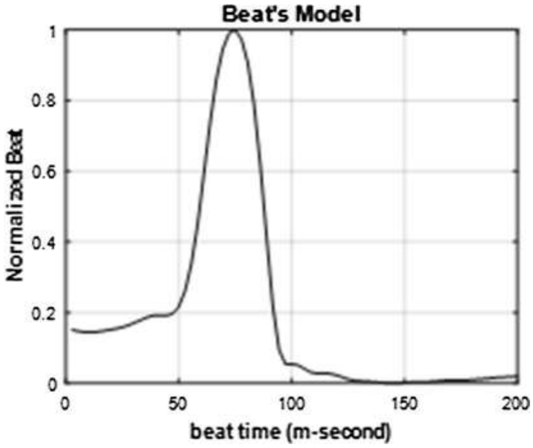


Fig. 2 Normal beat pattern

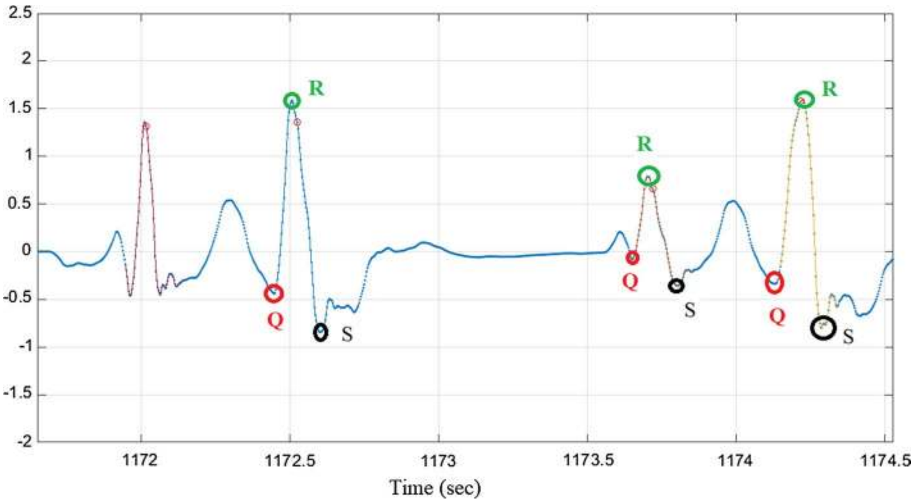


Fig. 3 Beat properties (not normalized)

The selected features are beat properties, Fig. 3 shows 3 summits (Q, R and S) of heartbeat

Features can be divided into three classes: Summits features, Temporal features and Morphological features.

Feature extraction and selection

Discrete Wavelet Transform (DWT) was used to get features from downloaded annotations. All chosen features were used in classification model. Features can be divided into three types as described in the following.

Summits features

Three main summits were considered in this paper as features, these features are related to the amplitude of three summits QRS as shown in Fig. 3.

Temporal features

Nine temporal features were calculated and used, one of them is the RR interval, defined as the time delay between two QRS peaks. Two other features are the interval between the current and previous beat and the one between the current and subsequent beat, which are called RR1 and RR2 respectively. Another interval is defined as the distance between the previous beat and its predecessor, called RR0. Figure 4 shows all RR intervals:

Three other features were extracted based on previous intervals. These features are called Ratio1, Ratio2 and Ratio3. They are defined as below:

$$Ratio1 = \frac{RR_0}{RR_1} \quad Ratio2 = \frac{RR_2}{RR_1} \quad Ratio3 = \frac{RR_m}{RR_1} \tag{1}$$

where

$$RR_m = mean(RR_0, RR_1, RR_2) \tag{2}$$

Three other features are chosen and selected to define each summit period as shown in Fig. 5.

Morphological features

Using the QRS summits (after normalization), the maximum of cross-correlation function between each detected beat and the following beat was calculated, as well as the maximum of cross-correlation between the current beat and the previous beat detected, called respectively Corr1 and Corr2 [17]. Another feature was the maximum of cross-correlation between a template of normal beat, with each QRS complex detected, called C_{xy}, was computed. For each record, the template was calculated as the averaged beat of a sequence of many normal sinus beats.

Finally, a feature was defined as the QRS duration when QRS beat equals to 0.5 in the normalized QRS complex, as shown in Fig. 5.

Morphological features are 4 features. The total features are 16 features (3 QRS amplitude, 9 temporal, 4 morphological) as shown in Fig. 6.

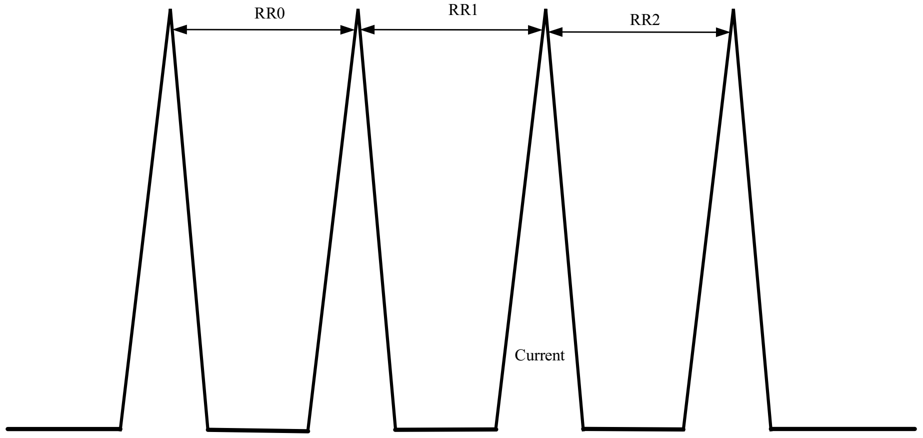


Fig. 4 RR intervals

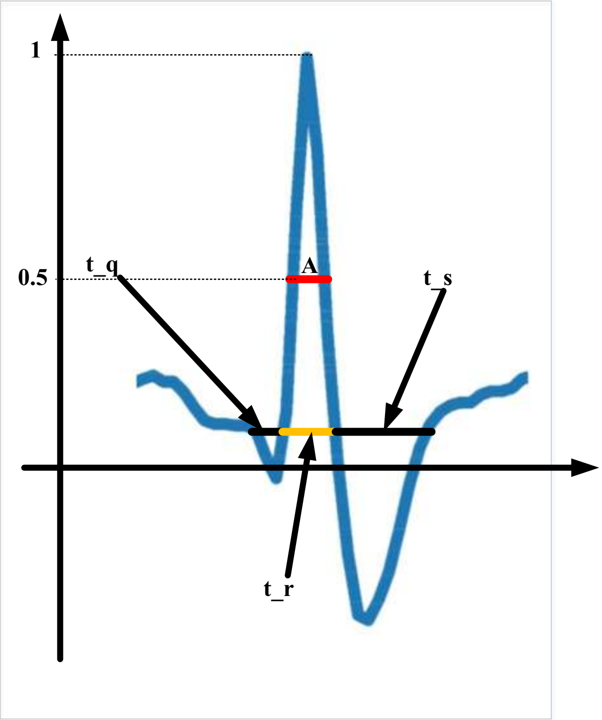


Fig. 5 QRS summits periods

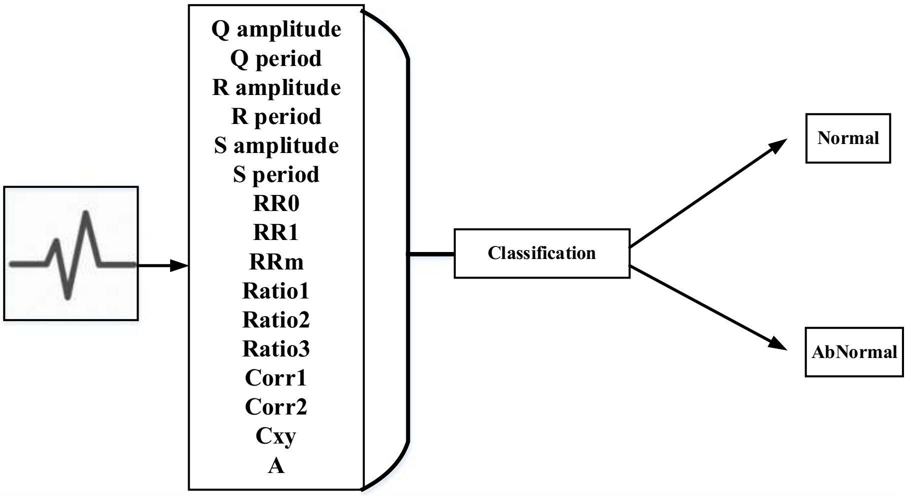


Fig. 6 Total features for classification

For features selection, many tests have been done to get best features for final model. Basically, all extracted features were selected for classification process.

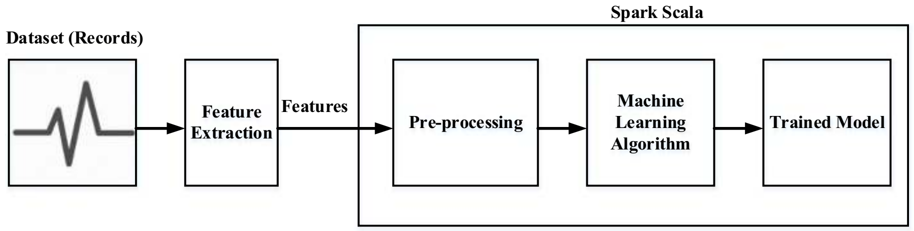


Fig. 7 Dataset to model stages

Table 1 Used records

Records to get final model									
101	102	103	104	105	107	108	109	111	112
113	114	117	118	121	122	123	124	201	202
203	205	207	208	210	212	213	214	215	217
219	222	230	231	232	233	234			
Records for testing the model									
100	106	115	116	119	200	209	220	221	223
228	234	809	812						

Heartbeat classification using machine learning

Features acquisition and storing

Dataset was used after downloading it from [18], there are a lot of patients records in this website and many types of databases; they belongs to real patients. As mentioned before, MIT-BIH Arrhythmia and MIT-BIH Supraventricular Arrhythmia databases were chosen. Figure 7 shows stages from dataset to get model to classify heartbeats.

The records are described in Table 1.

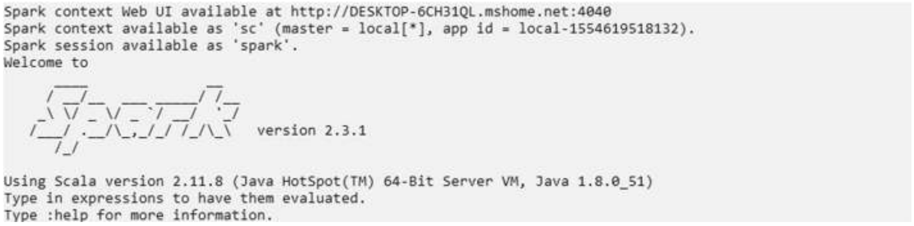
Features were obtained using Matlab software and stored in csv file with known columns types; some columns are integer type and others are double types, columns types is needed when reading csv file in Spark–Scala.

Processing

Processing was implemented using Spark–Scala firmware; most of papers used Matlab software for classification, Matlab software is very helpful tool in classification problems but when size of dataset is too large, other techniques are preferred. On other hand, machine learning algorithms are not implemented easily in Matlab. This case can be summarized as follows:

- Dataset size is too large.
- Need to implement algorithm like: Decision Tree, Random Forests, Gradient-Boosted Trees.
- Processing speed.

So, using big data tools would be helpful in this case. There are a lot of big data tools. For our case, Spark-shell and Scala were used in local host PC (Fig. 8: Local host PC Spark).



```
Spark context Web UI available at http://DESKTOP-6CH31QL.mshome.net:4040
Spark context available as 'sc' (master = local[*], app id = local-1554619518132).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.1
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_51)
Type in expressions to have them evaluated.
Type :help for more information.
```

Fig. 8 Local host PC Spark

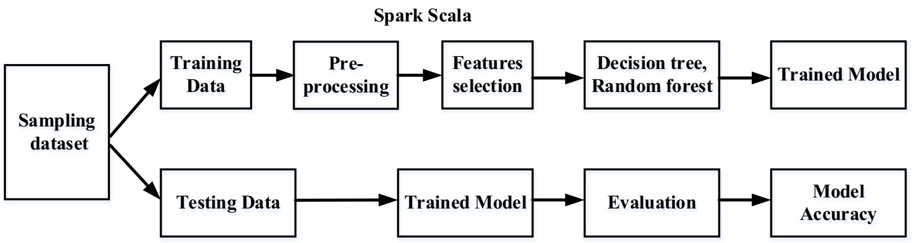


Fig. 9 Classification model

Table 2 Dataset schema

val Schema = StructType(Array(StructField("Var1", DoubleType,true)
StructField("Var2", DoubleType, true),StructField("Var3", DoubleType, true)
StructField("Var4", DoubleType, true),StructField("Var5", DoubleType, true)
StructField("Var6", DoubleType, true),StructField("Var7", DoubleType, true)
StructField("Var8", DoubleType, true),StructField("Var9", DoubleType, true)
StructField("Var10", DoubleType, true),StructField("Var11", DoubleType, true)
StructField("Var12", DoubleType, true),StructField("Var13", DoubleType, true)
StructField("Var14", DoubleType, true),StructField("Var15", DoubleType, true)
StructField("Var16", DoubleType, true),StructField("Var17", IntegerType, true)))

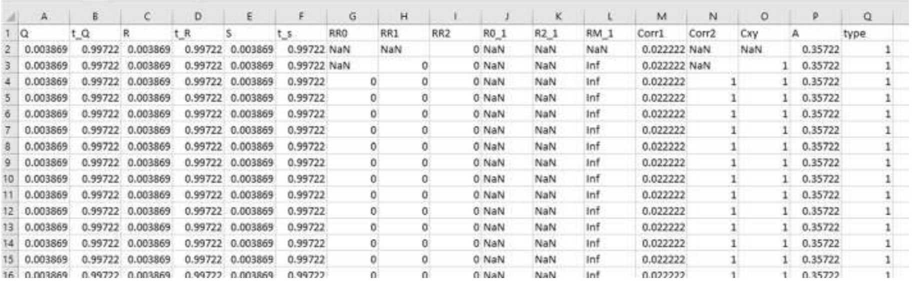
Figure 9 shows stages frame sampling data to get final model

CSV file can be read by Spark–Scala easily, we just built a schema for its columns. The schema defines type of each column in the csv file as shown in Table 2.

All columns are double values except the last column; it is label column and its type is integer. Figure 10 shows sample of dataset, it contains 16 columns as features and the column 17 contains beat type.

Every generic model of machine learning consists of some components independent of the algorithm adopted [19]. In our case, they are:

- Sampling dataset: divide dataset into two groups, one for training the model and the other for testing the model. Random sampling is used
- Pre-processing: all needed operations to get the data ready for classification model and it depends on dataset structure:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Q	t_Q	R	t_R	S	t_s	RR0	RR1	RR2	R0_1	R2_1	RM_1	Corr1	Corr2	Cxy	A	type
2	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	NaN	NaN	0	NaN	NaN	NaN	0.022222	NaN	NaN	0.35722	1
3	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	NaN	0	0	NaN	NaN	Inf	0.022222	NaN	1	0.35722	1
4	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
5	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
6	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
7	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
8	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
9	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
10	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
11	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
12	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
13	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
14	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
15	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1
16	0.003869	0.99722	0.003869	0.99722	0.003869	0.99722	0	0	0	NaN	NaN	Inf	0.022222	1	1	0.35722	1

Fig. 10 Sample of dataset

  - Fill null values: null values in columns might yield to mis-classification, so replacing all null values in columns with other value is very important in classification mode. Null values might replaced with static value (such as 0 value) or with values like mean value of all column or max value of column values. In our case, null values are filled with 0 value.
  - Process column 17 (heartbeat type): labeling this column with two classes (Normal and Abnormal) or multi classes (Normal and specified types of irregular heartbeat types).
  - Over fitting handling: it means when training data has many rows with type 1 and few rows with type 2. In our case more than 10000 rows have type “Normal” , while type “Abnormal” are less than that. Mapping dataset with fractions is done for fitting data.
  - Separate columns according to their type (Integer, Double)
  - Using String-indexer for labeling beat type.

- Features Selection: Select columns from data columns as features.
- Using algorithm GBT: Gradient-Boosted Trees with parameters MaxDepth and MaxIter, Or RF: Random Forest with parameters MaxDepth and NumTrees.
- Training model: after this step, a trained model is generated and is ready for testing on testing data.
- Evaluation trained model: evaluators are needed to calculate accuracy for each trained model, each algorithm has its own evaluator.

In this paper, Gradient-Boosted Trees model (GBT) and Random Forest model (RF) are implemented and tested.

Gradient-Boosted Trees model

Gradient-Boosted (GDB) Tree is a machine learning technique for regression and classification issues, which produces a prediction model in the form of an ensemble of weak prediction models. The idea of gradient boosting originated in the observation that boosting can be interpreted as an optimization algorithm on a suitable cost function [20]. The built model basically depends on two parameters of gradient boosted tree; these two parameters are most important parameters of GBT. The GBT model is in Table 3.

Table 3 GBT tree

```
var  dTree-am  =new  GBTClassifier()
.setLabelCol("label")//beat type
.setFeaturesCol("features")//chosen features
.setMaxDepth(20)//depth of GBT
.setMaxIter(15)//iteration of GBT
```

Table 4 values of iteration and depth

Max Depth	Max iteration	Extracted model
20	10	GBT Model1
10	10	GBT Model2
5	10	GBT Model3
5	5	GBT Model4
5	15	GBT Model5
10	15	GBT Model6
10	20	GBT Model7

Table 5 RF model

```
var  dTree-am  =new  RandomForestClassifier()
.setLabelCol("label")//beat type
.setFeaturesCol("features")//chosen features
.setNumTrees(20)//depth of GBT
.setMaxDepth(15)//iteration of GBT
```

GBT trained model was built according to many values of Max iteration and Max Depth; values were changed manually as in Table 4.

Random Forest model (RF model)

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [16, 21]. The built model depended basically on two parameters of random forest; these two parameters are most important parameters of RF. The RF model is in Table 5.

RF trained model was built according to many values of Number of Trees and Max Depth; values were changed manually as in Table 6.

Results and discussion

The dataset contains 205,146 rows, they were randomly split into two parts: training and testing. After that, the built model was tested validated on different dataset (32,168 rows). To validate this work, accuracy of model was calculated using binary evaluator BinaryClassificationEvaluator for GBT model and

Table 6 RF models

Max Depth	Number of trees	Extracted model
15	15	RF Model1
5	15	RF Model2
10	15	RF Model3
15	10	RF Model4
5	10	RF Model5
10	10	RF Model6
20	20	RF Model7

Table 7 GBT results

Model	Accuracy%	SP%	SE%	CC%
GBT Model1	96.75	96	96	96
GBT Model2	92.6	94	91	92
GBT Model3	84.11	80	88	84
GBT Model4	81.75	75	88	82
GBT Model5	84.98	82	87	85
GBT Model6	93.21	94	91	93
GBT Model7	93.63	94	92	93

MulticlassClassificationEvaluator for RF model. In addition, Sensitivity and specificity were calculated based on Eq. (3). Where [22, 23]:

SE (Sensitivity): The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease.

SP (Specificity): The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease.

To calculate SE and SP, these terms should be defined as follows:

- TP True positive: the patient has the disease and the test is positive.
- FP False positive: the patient does not have the disease but the test is positive.
- TN True negative: the patient does not have the disease and the test is negative.
- FN False negative: the patient has the disease but the test is negative.

And equations of SE and SP:

$$SE = \frac{TP}{TP + FN} * 100 \quad SP = \frac{TN}{TN + FP} * 100 \tag{3}$$

CC: Correct Classification is computed as below:

$$CC = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{4}$$

Tables 7 and 8 summarized results of GBT and RF algorithms.

Table 8 RF results

Model	Accuracy%	SP%	SE%	CC%
RF Model1	96.31	96	96	96
RF Model2	80.65	70	89	80
RF Model3	91.67	89	93	91
RF Model4	95.15	96	96	96
RF Model5	79.83	70	87	79
RF Model6	91.63	90	92	91
RF Model7	97.98	97	98	97

Table 9 4 classes classification

Class	Label	Information
1	1	Normal
2	5	PVC: Premature ventricular contractions
3	9	PAC: Premature atrial contraction
4	0	Other

Training process in random forest is faster than decision tree, while testing process in decision tree is faster than in random forest. Parameters of both algorithms were changed manually. The optimal values for tuned parameters can be obtained by running methods with cross validation, but they need too much time. For production system, cross validation can be used and the resulted optimal values can be used instead.

Tables above show that built models of both algorithms are capable to predict types of heartbeats with accuracy 96.75% and 97.98 for GBT model and RF model respectively.

4-Classes classification

After two classes classification, multi classes classification was validated using RF Algorithm. RF Algorithm supports multi classes classification, while GBT supports only binary classification. Originally, the dataset has a column named label, it has many different integer values such as 1, 5, 9, 2 and others. Each value labels a class such as 1 labels Normal beat and all other values labels Abnormal beat (5 labels PVC and 9 labels PAC).In binary classification, this column is handled to be just two classes in the pre-processing stage (Normal and Abnormal). In multi classification, this column is handled to be 4 classes in the pre-processing stage (Normal, PVC, PAC and Other) as explained in Table 9.

Table 10 shows results using random forest algorithm:

Table above shows that built model for multi classification is able to predict multi types of heartbeats with accuracy 98.03%, this contribution is very useful; it predicts 4 classes of heartbeats at once.

Table 10 RF results for multi classification

Model	Accuracy%
Number of trees = 15, Max depth = 10	87.74
Number of trees = 15, Max depth = 15	95.89
Number of trees = 10, Max depth = 15	95.94
Number of trees = 10, Max depth = 20	97.85
Number of trees = 10, Max depth = 25	98.03

Max accuracy of binary classification in our case was 97%, it is better the results in [11, 12]. In [11], the accuracy of classification was 96.97% and 96% in [12]. Max accuracy of multi classification in our case was 98.03%, it is better comparing to [7], where multiply types of heartbeats have been studied and classified using convolutional neural network and the accuracy of classification was 93.4%.

Conclusion and future scope

In summary, This work has validated an ability to classify heartbeats. Classification process is using some features of heartbeats and machine learning classification algorithms with local host pc working using only one node, which are crucial for diagnosis of cardiac arrhythmia. The developed GBT and RF models can classify different ECG heartbeat types and thus, can be implemented into a CAD ECG system to perform a quick and reliable diagnosis. The proposed model has the potential to be introduced into clinical settings as a helpful tool to aid the cardiologists in the reading of ECG heartbeat signals and to understand more about them. The occurrence, sequential patterns and persistence of the classes of ECG heartbeats considered in this work can be grouped under three main categories which represents normal, PVC, PAC, and other. As a future work, implemented methods can be rebuilt to work with many classes (Ex: more than 5 types of heartbeats), the work can be developed to be used in real time and be trained continuously to enhance it and increase its accuracy. Moreover, the whole process of classification can be used with other types of datasets such as stress and clinical datasets.

Abbreviations

ECG: electrocardiogram; SVM: Support Vector Machine; NN: neural network; PCA: Principal Component Analysis; MIT: Massachusetts Institute of Technology; DWT: Discrete Wavelet Transform; GBT: Gradient-Boosted Trees; RF: Random Forest.

Acknowledgements

Thanks for Mr. Ammar Asaad, and Mr. Ahmad Ali for their co-operation and help.

Authors' contributions

FIA took the role of performing the literature review, finding best model for heartbeat classification. MY took on a supervisory role and oversaw the completion of the work, Proposed data set for implementation and validation. Both authors read and approved the final manuscript.

Funding

The authors declare that they have no funding.

Availability of data and materials

The data set is available to public and can be found in https://physionet.org/physiobank/, it is simple to extract the data in many formats using physionet ATM Bank https://physionet.org/cgi-bin/atm/ATM.

Ethics approval and consent to participate

The authors declare that this study contain real data for real patients obtained from famous dataset on internet and this data is available for all.

Consent for publication

The database is free to use by all users.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Informatics and Decision Supporting Systems, Higher Institute for Applied Sciences and Technology, Damascus, Syria.
[2] Faculty of Computer and Automation Engineering, Damascus University, Damascus, Syria.

Received: 13 May 2019 Accepted: 20 August 2019
Published online: 31 August 2019

References

1. Li H, Yuan D, Ma X, Cui D, Cao L. Genetic algorithm for the optimization of features and neural networks in ECG signals classification. Sci Rep. 2017;7:41011.
2. Kachuee M, Fazeli S, Sarrafzadeh M. ECG heartbeat classification: a deep transferable representation. In: 2018 IEEE international conference on healthcare informatics (ICHI). New York: IEEE; 2018. p. 443–4.
3. Zhao Z, Yang L, Chen D, Luo Y. A human ECG identification system based on ensemble empirical mode decomposition. Sensors. 2013;13(5):6832–64.
4. Valenza G, Citi L, Lanatá A, Scilingo EP, Barbieri R. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. Sci Rep. 2014;4:4998.
5. Valenza G, Greco A, Citi L, Bianchi M, Barbieri R, Scilingo E. Inhomogeneous point-processes to instantaneously assess affective haptic perception through heartbeat dynamics information. Sci Rep. 2016;6:28567.
6. Christov I, Jekova I, Bortolan G. Premature ventricular contraction classification by the kth nearest-neighbours rule. Physiol Meas. 2005;26(1):123.
7. Sannino G, De Pietro G. A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. Fut Gener Comput Syst. 2018;86:446–55.
8. Celesti F, Celesti A, Carnevale L, Galletta A, Campo S, Romano A, Bramanti P, Villari M. Big data analytics in genomics: the point on deep learning solutions. In: 2017 IEEE symposium on computers and communications (ISCC). New York: IEEE; 2017. p. 306–9.
9. Ahmad AK, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data. 2019;6(1):28.
10. Ma'sum M.A, Jatmiko W, Suhartanto H. Enhanced tele ECG system using hadoop framework to deal with big data processing. In: 2016 international workshop on Big Data and information security (IWBIS). New York: IEEE; 2016. p. 121–6.
11. Dalvi RdF, Zago GT, Andreão RV. Heartbeat classification system based on neural networks and dimensionality reduction. Res Biomed Eng. 2016;32(4):318–26.
12. Ince T, Kiranyaz S, Gabbouj M. A generic and robust system for automated patient-specific classification of ECG signals. IEEE Trans Biomed Eng. 2009;56(5):1415–26.
13. O'Brien J. Using hidden Markov models and spark to mine ECG data.
14. PhysioBank ATM. 2019. https://physionet.org/cgi-bin/atm/ATM. Accessed 21 June 2019.
15. Moody GB, Mark RG. The impact of the mit-bih arrhythmia database. IEEE Eng Med Biol Mag. 2001;20(3):45–50.
16. Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol. 1. New York: IEEE; 1995. p. 278–82.
17. Gharaviri A, Dehghan F, Teshnelab M, Moghaddam HA. Comparison of neural network, ANFIS, and SVM classifiers for PVC arrhythmia detection. In: 2008 international conference on machine learning and cybernetics, vol. 2. New York: IEEE; 2008. p. 750–5.
18. PhysioBank ATM Dataset. 2019. https://physionet.org/physiobank/. Accessed 21 June 2019.
19. Alzubi J, Nayyar A, Kumar A. Machine learning from theory to algorithms: an overview. In: Journal of physics: conference series, vol. 1142. Bristol: IOP Publishing; 2018. p. 012012.
20. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38(4):367–78.
21. Barandiaran I. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998;20(8):122.
22. Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008;56(1):45.
23. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. Continuing Educ Anaesth Crit Care Pain. 2008;8(6):221–3.

Publisher's Note