**Arrhythmia Disease Classification and Mobile Based System Design**

Soha Samir AbdElMoneem\(^1,a\), Hany Hanafy Said\(^2,b\), and Amani Anwar Saad\(^3,c\)

\(^1,2,3\) Computer Engineering Department, Arab Academy for Science, Technology & Maritime Transport Abukir, Alexandria, Egypt

E-mail: \(^a\) soha.samir90@gmail.com, \(^b\) amani.saad@aast.edu, \(^c\) hany.said.1980@ieee.org

**Abstract** – Heart Rate Variability (HRV) is a measure of variation in the time interval between consecutive heart beats. HRV analysis is highly sensitive for risks linked with cardiovascular diseases which are main causes of death in Egypt and all over the Middle East. Early detection of cardiac arrhythmia diseases achieves effective treatment by making it easy to choose appropriate anti-arrhythmic drugs, also very important for improving arrhythmia therapy and preventing number of death in individuals. In this paper, an efficient cardiac arrhythmia detection algorithm is introduced. Different classifiers are deployed and examined on ECG signals. Various oversampling techniques are investigated to handle imbalanced dataset. The ensemble classifier; support vector machine and Random forest with random sampling show accuracy of 98.18 % in 0.145 sec which is the best accuracy among all other classifiers. In addition, this paper also proposes a mobile based system architecture integrated with the algorithm for diagnosis and classification of cardiac arrhythmia diseases. The proposed system can be easily used by patients to check their heart health remotely and easily.

**Keywords:** HRV, ECG Classification, Arrhythmia, Random Forest, Logistic Regression, Ensemble Method.

1. **Introduction**

Machine learning is the science of getting computers to act without being explicitly programmed. Nowadays, Machine learning in medical field is a domain of challenge and it is rapidly growing. In medical organizations, dealing with a huge amount of patient history data that can lead to misdiagnosis and uncertain results. Deploying machine learning algorithms will exploit useful information from this huge data. Cardiovascular diseases are one of the leading chronic diseases in Egypt and all over the Middle East. According to the latest World Health Organization, data published in 2017 coronary heart disease deaths in Egypt reached 24.58% of total deaths [1]. Egypt ranks #18 in the world, also 54% of deaths are from non-communicable disease in the Eastern Mediterranean Region are due to cardiovascular diseases. Cardiac arrhythmia causes disorder of electrical heart rhythm, which indicates a heart disease, stroke or sudden cardiac death. Early diagnosis of cardiac arrhythmia makes it easy to choose appropriate anti-arrhythmic drugs that consequently reduce the number of death in individuals. An electrocardiogram (ECG) is a non-invasive way used in measuring the electric activity of the heart.
and detecting heart diseases as several supervised learning algorithms have been deployed to classify these types for heart diseases [2-5].

This research applies a comprehensive study using different machine learning algorithms, features selections and oversampling techniques. The outcomes from this study will be integrated as a diagnosis tool for the proposed mobile-based system for Arrhythmia disease classification. The proposed system will help patients to check their heart health remotely. The rest of the paper is organized as follows. Dataset description and literature review are presented in sections 2 and 3. The proposed system architectures is described in section 4 while Arrhythmia diagnosis algorithm is outlined in section 5. Experimental results is shown in section 6. Conclusions and future work recommendations are provided in section 7.

2. Arrhythmia dataset
ECG is the process of recording electrical activity of heart recorded by skin electrode where it is measured on the surface of human body [6]. The Rhythm and heart rate reflects the cardiac health of human heart beat which is used in detection of the heart diseases. Any variation of heart rate or rhythm is an indication of cardiac arrhythmia, which could be detected by analysis of the recorded ECG waveform. The standard arrhythmia dataset found at the UCI Machine Learning Repository is used in this study [7]. This dataset contains 452 patients and 279 features. The features include patient attributes such as age, height, weight, and gender as well as quantized ECG data. Each record is assigned to 1 of 16 classes, the class distribution is shown in table 1. Class Label 1 indicates normal ECG patterns while classes label between 2 to 16 indicates ECG patterns with different abnormalities. Several studies have considered this dataset as it is highly imbalanced and contains about 0.32% of missing data [2-5].

| Class                                      | Number of instances |
|-------------------------------------------|---------------------|
| Normal                                    | 245                 |
| Ischemic changes                          | 44                  |
| Old Anterior Myocardial Infarction        | 15                  |
| Old Inferior Myocardial Infarction        | 15                  |
| Sinus tachycardia                         | 13                  |
| Sinus bradycardia                         | 25                  |
| Ventricular Premature Contraction         | 3                   |
| Supraventricular Premature Contraction    | 2                   |
| Left bundle branch block                  | 9                   |
| Right bundle branch block                 | 50                  |
| First degree Atrioventricular block       | 0                   |
| Second degree Atrioventricular block      | 0                   |
| Third degree Atrioventricular block       | 0                   |
| Left ventricle hypertrophy                | 4                   |
| Atrial Fibrillation                       | 5                   |
| Others                                    | 22                  |

3. Related work
Several studies in the literature have focused on detection and classification of ECG Arrhythmia [2-5]. Mitra and his co-workers [2] have classified arrhythmia using incremental back propagation neural network and Levenberg-Marquardt (LM) classification. Umale and his co-workers [3] have carried out
extra studies that involve applying neural network to predict cardiac arrhythmias along with other classifiers such as Support Vector Machine (SVM) and, Naive Bayes (NB) classifiers. They reported that deploying feed forward neural network for classification or regression with a single layer of hidden nodes would provide better result as compared to other algorithms.

Mustaqeem and his co-workers [4] have classified patients into one of the sixteen sub-classes among which class one represents absence of arrhythmia disease, whereas the other fifteen classes represent electrocardiogram records of various sub-types of arrhythmia. Wrapper based feature selection technique were used. For multi-class classification SVM based approaches including one-against-one, one-against-all and error correction code (ECC) are employed to detect the presence and absence of arrhythmias. The results show that OAO method of SVM outperforms all other classifiers by 92% using 90/10 data split option. Singh and his co-workers [5] have classified arrhythmia using SVM, Random Forest, and Repeated Incremental Pruning (JRip), along with three filter-based feature selection methods This includes Chi-square Statistic, Symmetric Uncertainty (SU),Gain Ratio (GR) respectively.

The Study shows that random forest obtains highest accuracy of 85.58% using gain ratio feature selection method with a subset of 30 features.

The literature review shows that many techniques were deployed to enhance the accuracy for arrhythmia classification. The best accuracy obtained so far is 92% and it is not sufficient for building an accurate medical application. Since the proposed mobile system relies on the main service (arrhythmia classification), the service has to be more robust and accurate enough. This research studies the UCI dataset comprehensively and proposes an algorithm that would solve its challenges and achieve higher accuracy in detection. One of these challenges is imbalanced dataset where no comprehensive investigation were conducted to tackle that challenge since these studies focused mostly on applying various classifiers rather than considering the effect of applying various oversampling techniques to efficiently handle this type of dataset.

This research examined four different oversampling techniques which are Synthetic Minority Oversampling Technique (SMOTE), Random over-sampling, Adaptive Synthetic (ADASYN) and Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTENC) [8, 9]. SMOTE is a powerful sampling method that goes beyond simple under or over sampling. This algorithm creates new instances of the minority class by creating convex combinations of neighbouring instances. Random over-sampling is used to repeat some samples in the classes, which are under-represented and balance the number of samples between the dataset. So it is less biased toward the majority class. ADASYN generate new samples in by interpolation. However, the samples used to generate new synthetic samples differ. In fact, ADASYN focuses on generating samples next to the original samples which are wrongly classified using a K-Nearest Neighbours classifier while the basic implementation of SMOTE will not make any distinction between easy and hard samples to be classified using the nearest neighbors rule. Therefore, the decision function found during training will be different among the algorithms. The SMOTENC is an extension of the SMOTE algorithm for which categorical data are treated differently [8].

4. Proposed system architecture
In this paper, Arrhythmia diagnosis algorithm is proposed where different classifiers are examined. This algorithm would be served as diagnosis service hosted by cloud service for the proposed system; mobile based System for arrhythmia disease; as shown in Figure 1.
This service can be accessed when needed using a mobile application, which sends ECG signals remotely using Bluetooth from a wearable sensor attached to the body of the patient. This sensor could be attached as a chest belt or a watch. On the other side; the mobile application receives the ECG signal and saves it as a time series of voltage magnitudes for five minutes. This data is sent to the server to be saved, analysed through the Arrhythmia Diagnosis algorithm which diagnosis, and detect the class of heart arrhythmia disease that the patient holds. The server then sends the result to alert the registered Doctor through SMS on his phone; the doctor can view the ECG history of his patient, send a medication or request an appointment through a responsive web application. The following section presents in details the proposed Arrhythmia diagnosis algorithm.

5. Arrhythmia diagnosis algorithm
The block diagram for the proposed arrhythmia diagnosis algorithm is depicted in Figure 2. This algorithm would be integrated as a cloud service to the mobile based system for arrhythmia disease detection as shown in Figure 1. The main stages for the proposed algorithm include data preprocessing and classification. Data preprocessing includes handling missing data, normalization, feature selection, resampling and class reduction. The following subsections present in details the stages of arrhythmia diagnosis algorithm.
5.1. Data Pre-processing
The original data contains columns with both missing values and single valued columns having the same value for all the patient records. In order to pre-process data, this research applied mean function to deal with missing values and where single valued columns were deleted from the dataset. The resulting dataset contains 452 instances and 257 features. Data normalization is an essential phase, where all feature values are normalized using minimum and maximum scaling estimator. Feature Selection is crucial since the dataset contains few numbers of records (452) in comparisons to available features (257). It helps in avoiding overfitting and shows the important features. According to Eduardo José da S. Luz et al. the effective features of the heart rate are P, Q, R, S & T waves along with their respective amplitudes (19 features)[10]. This research has examined select from model algorithm for feature selection (140 features) [11]. The dataset classes are highly imbalanced as its classes’ categories are not approximately equally represented. This would affect the accuracy of the algorithms. So this study examined four different techniques for oversampling which are random oversampling algorithm, SMOTE, ADASYN and SMOTENC [8, 9].

5.2. Training Classifiers and testing
Seven classifiers are examined; these classifiers are Naive Bayes, Logistic Regression, K-Nearest Neighbors, Random Forest, Decision Trees, Support Vector Machines(poly deg2) and ensemble classifier (SVM and Random Forest). The latter classifier uses as voting as it combines the predictions of several base estimators built with a given learning algorithm in order to improve generalizability and robustness over a single estimator respectively [12]. The dataset was split into 80%-20% between training and testing respectively.

6. Results and discussion
Experiments run on a computer with an Intel core i3 processor (2.53 GHz) using 4.00 GB of RAM. Python Scikit learn software package is used via Jupiter Notebook. There are four experiments groups using these oversampling techniques; SMOTE, Random oversampling, ADASYN and SMOTENC. Their results are shown in Table 2 to Table 5. The results shown below depict the performance of different classifiers using all features (257), select from model algorithm (140 features) and most effective features (19). All these results have been undertaken using four different oversampling techniques. From the classifier point of view, it is clear that by applying Naive Bayes classifier, the proposed detection accuracy is in range (69% to 82.4%) where all features are taken into consideration. This poor results is obtained as it assumption has significant weakness, because there will almost never been an independence between every pair of features given enough features. Deploying Logistic regression as classifier shows the lowest result of 69.99 in 1.142 sec with 19 features and 82.6% in 0.017 sec with 140 features. The KNN shows better results (range of 85.7% to 90.7%) with all features because it work well in high dimensions, Decision Trees have shown an accuracy of 96% using 19 features while got 98% using 140 features. Random forest showed an accuracy of 97.5% using 19 features and got accuracy of 99.7% with 140 features. This is due to the ability to tackle issues like ‘pruning’ on contrary to decision trees. SVM was tested using both linear and polynomial kernel with k = 10. Based on this research results polynomial kernel obtains much better results than linear kernel where accuracy using polynomial kernel is 95.53% in 6.086 sec. SVM is a potential candidate for classifier, as it does not depend upon the dimensionality of the input space. That is why SVM is good for high dimensional space as it got accuracy of 97.9% using 140 features. SVM and voting (RF+SVM) shows the best accuracy over all the methods on contrary to the performance of logistic regression and Naïve Bayes.
### Table 2: Summary of the performance for various classifiers using Smote Oversampling

| Algorithm          | All features Accuracy (%) | Prediction time | 140 feature Accuracy (%) | Prediction time | 19 features Accuracy (%) | Prediction time |
|--------------------|---------------------------|-----------------|---------------------------|-----------------|---------------------------|-----------------|
| Naïve bayes        | 69                        | 0.045           | 82.6                      | 0.017           | 71                        | 0.023           |
| SVM                | 95.89                     | 0.612           | 96.5                      | 0.336           | 95.53                     | 6.028           |
| Random Forest      | 97.77                     | 7.758           | 97.1                      | 8.308           | 95.02                     | 6.086           |
| KNN                | 68.63                     | 0.576           | 90.5                      | 0.261           | 89.78                     | 0.164           |
| Logistic regression| 94.35                     | 1.929           | 94                        | 0.751           | 69.99                     | 1.142           |
| RF+SVM             | 98.18                     | 8.315           | 97.5                      | 8.84            | 95.03                     | 13.755          |
| Decision tree      | 93.035                    | 0.291           | 89.1                      | 0.243           | 86.06                     | 0.058           |

### Table 3: Summary of the performance for various classifiers using Random Oversampling

| Algorithm          | All features Accuracy (%) | Prediction time | 140 feature Accuracy (%) | Prediction time | 19 features Accuracy (%) | Prediction time |
|--------------------|---------------------------|-----------------|---------------------------|-----------------|---------------------------|-----------------|
| Naïve bayes        | 82.4                      | 0.082           | 76                        | 0.042           | 59.04                     | 0.63            |
| SVM                | 98.4                      | 0.529           | 97.9                      | 0.294           | 94.83                     | 4.336           |
| Random Forest      | 99.5                      | 6.835           | 99.7                      | 5.93            | 97.46                     | 0.12            |
| KNN                | 85.7                      | 0.571           | 86.7                      | 0.03            | 78.62                     | 0.232           |
| Logistic regression| 97.4                      | 4.446           | 94.9                      | 1.97            | 65.28                     | 5.712           |
| RF+SVM             | 99.4                      | 7.248           | 99.0                      | 0.43            | 98.18                     | 0.145           |
| Decision tree      | 98.1                      | 0.179           | 98                        | 0.123           | 96.244                    | 0.152           |

### Table 4: Summary of the performance for various classifiers using ADASYN Oversampling

| Algorithm          | All features Accuracy (%) | Prediction time | 140 feature Accuracy (%) | Prediction time | 19 features Accuracy (%) | Prediction time |
|--------------------|---------------------------|-----------------|---------------------------|-----------------|---------------------------|-----------------|
| Naïve bayes        | 71.6                      | 0.051           | 79.3                      | 0.017           | 68.4                      | 0.006           |
| SVM                | 96.7                      | 0.581           | 96.7                      | 0.347           | 95.3                      | 15.878          |
| Random Forest      | 97.3                      | 8.892           | 97.1                      | 8.015           | 95.2                      | 5.365           |
| KNN                | 90.7                      | 0.625           | 90.7                      | 0.298           | 88.5                      | 0.037           |
| Logistic regression| 95.1                      | 1.919           | 93.6                      | 0.862           | 70.6                      | 1.064           |
| RF+SVM             | 98.2                      | 8.332           | 98.2                      | 8.385           | 96.2                      | 20.89           |
| Decision tree      | 89.6                      | 0.329           | 91.6                      | 0.295           | 87.4                      | 0.037           |
Table 5: Summary of the performance for various classifiers using SMOTENC

| Algorithm          | All features | Prediction time | 140 feature | Prediction time | 19 features | Prediction time |
|--------------------|--------------|-----------------|-------------|----------------|-------------|-----------------|
|                   | Accuracy(%)  |                  | Accuracy(%) |                  | Accuracy(%) |                  |
| Naïve bayes       | 73.4         | 0.056           | 83.9        | 0.017           | 77.2        | 0.006           |
| SVM                | 96.7         | 0.607           | 96.9        | 0.337           | 94.0        | 20.112          |
| Random Forest      | 97.8         | 7.613           | 97.3        | 7.847           | 94.6        | 4.13            |
| KNN                | 90.2         | 0.599           | 90.8        | 0.271           | 88.2        | 0.036           |
| Logistic regression| 95.1         | 2.66            | 94          | 0.747           | 72.5        | 0.592           |
| RF+SVM             | 97.6         | 8.53            | 97.6        | 8.807           | 94.7        | 26.59           |
| Decision tree      | 91.3         | 0.418           | 91.1        | 0.246           | 85.9        | 0.036           |

Considering different oversampling techniques & features selections, the average classifiers accuracy when considering all features, indicate that the selection of oversampling technique has significant effect of the classifier accuracy where Random sampling get highest accuracy 94.4% while SMOTE gets lowest 88.12%.

When selecting few features 19 features, the selection of oversampling techniques has minor effect where highest accuracy is SMOTENC as it gets 86.7% while lowest is Random sampling (84.24%).

7. Conclusions
This paper presents a comparative study for classification of cardiac arrhythmia diseases dataset. The research is carried out on the standard dataset taken from the University of California. Several classification algorithms were examined alongside different oversampling techniques. The Proposed Arrhythmia diagnosis algorithm is discussed. Currently, a mobile-based system for Arrhythmia disease is being implemented. The proposed system receives the ECG signals from patient via mobile application and then processes it by the Arrhythmia diagnosis algorithm that acts as a cloud service. Through the proposed mobile-based system, patient can identify the type of arrhythmia.

Using 80%-20% train-test split, the SVM classifier was found to have the best performance in terms of prediction accuracy using SMOTE with 19 features got 95.53% in 6.086 sec, also using ADASYN with 19 features got accuracy of 96.2 % in 20.89 sec, SMOTENC with 19 features got accuracy of 94.7% in 26.59 sec and using random sampling with 19 features shows that Voting RF and SVM gives a good accuracy and prediction time of 98.18 % in 0.145 sec with 19 features. Studying those techniques showed that we can obtain mainly the same good accuracy and prediction time with a less number of relevant features. Using a fast and effective system make it easy for patients to use the system and predict cardiac Arrhythmia while Voting RF and SVM have shown high prediction accuracy of 98.18 % in 0.145 sec.

References
[1] World Health Organization (2018, Nov. 11). EGYPT : CORONARY HEART DISEASE. Retrieved from www.worldlifeexpectancy.com/egypt-coronary-heart-disease.
[2] Mitra M and Samanta R K 2013 Procedia Technology 10 76-84
[3] Umale V, Waghe S, Bhlerao D and Gound R 2016 International Research Journal of Engineering and Technology (IRJET) 3 258-262
[4] Mustaqeem A, Anwar S M and Majid M 2018 Computational and Mathematical Methods in Medicine 2018 Article ID 7310496
[5] Singh N and Singh P 2018 *Engineering Vibration, Communication and Information Processing*, Springer, Singapore *478* 469-480

[6] Rajni R and Kaur I 2013 *International Journal of Computer Applications* *84* 22-25

[7] UCI Machine Learning Repository (2018, Nov. 28). Arrhythmia Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Arrhythmia.

[8] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 *Journal of artificial intelligence research* *16* 321-357

[9] Haibo H, Yang B, Eduardo A G and Shutao L 2008 "ADASYN: Adaptive synthetic sampling approach for imbalanced learning" *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* pp. 1322-1328

[10] Eduardo J S L, William R S, Guillermo C-C and David M 2016 *Computer Methods and Programs in Biomedicine* *127* 144–164.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M and Perrot M 2011 *Journal of Machine Learning Research* *12* 2825-2830

[12] Bauer E and Kohavi R 1999 *Machine Learning* (Berlin: Springer) (vol 36) pp 105-139