Fast Parallel SVM based Arrhythmia Detection on Multiple GPU Clusters

Ghazanfar Latif
Department of Computer Science,
Prince Mohammad bin Fahd University,
Al-Khobar, Saudi Arabia.
glatif@pmu.edu.sa

Jaafar Alghazo
Department of Computer Engineering,
Prince Mohammad bin Fahd University,
Al-Khobar, Saudi Arabia.
jghazo@pmu.edu.sa

Mohsin Butt
Prep Computer Science Department,
King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia.
mohsinbutt@kfupm.edu.sa

Zafar A. Kazimi
Department of Information Technology,
Prince Mohammad bin Fahd University,
Al-Khobar, Saudi Arabia.
zkazimi@pmu.edu.sa

***Abstract*—Regression analysis and classification can be done using a supervised learning technique called Support Vector Machine (SVM) which is one of many such methods. The method creates hyperplanes which are used to analyze data patterns and separate data into multiple classes. The computation complexity of the algorithm is very high for training and testing of large multidimensional datasets. In this work, we propose a scalable and cost-effective method to run SVM that reduces memory usage and computing power. The process uses distributed cloud GPU's cluster nodes to run the algorithm in parallel on data which is divided into "n" parts. The results obtained from each of the cluster nodes are merged on a master node and the SVM algorithm is applied once more for classification. The study tackles the ECG classification using parallel SVM to investigate heartbeats and brain traces linked with different types of Arrhythmia and Seizure. Experiments performed on real ECG datasets (MIT BIH Diagnostic database and EEG Seizure database) resulted in a classification accuracy of 97.45%. The technique is proven both efficient by reducing training time and with high classification accuracy. The results achieved show that the proposed technique outperforms similar methods proposed in previous literature.**

Keywords- ECG Classification; Support Vector Machine (SVM); Parallel SVM; Arrhythmia Classification; Cloud Computing

I. INTRODUCTION

Cardiac and brain problems have been basic health issues on the basis of the number of admissions caused by the cardiac and brain cases all over the world. Around 295 thousand emergency medical services are treated in US for cardiac arrest annually [1] and this number is still increasing. Seizure is a sudden electrical disorder in the brain caused by sudden abnormal behavior of neurons [2]. The top right side of the human hearts generates an electrical signal from the Sino Atrial node simulating a heartbeat [3].

This research to improve the efficiency by utilizing cloud GPU clusters for scientific research to get the live reports for Arrhythmia and Seizure analysis, classification, and detection with high accuracy. The classification and recognition process concentrates on the abnormal ECG signals. Due to the physical, geographical and computation limits of the centralized system, Distributed Parallel Classification has gained momentum [4]. Support Vector Machine (SVM) is a very popular classification method across various fields. SVM was originally based on the concept of hyper places to act as a class divider in the binary classification. Compared with other classification methods such as K-nearest neighbors, SVM achieves better accuracy when parameters are selected properly. However, even though it achieves better accuracy but this comes at a cost of greater computational time in the training phase in particular for non-linear kernel based SVM used on large datasets.

The rest of this paper is organized as follows: section 2 details the background of ECG and arrhythmia disease. Section 3 state of the art research pertaining to ECG classification are presented. In section 4, the proposed methodology along with different classification techniques used at detailed. In section 5, the MIT-BIH arrhythmia datasets and experimental results are explained. A summary and conclusion are shown in section 6. Section 7 contains the references used in this work.

II. BACKGROUND

Unnatural heart rhythm (beat) referred to as Arrhythmia is the most common cause of death. In this case, the heart beats increase or decrease in a sudden and irregular manner inhibiting the heart supplying the body with regular levels on blood. The lack of blood supply to all parts of body can damage different body organs. On average, a normal heart rate is 50 to 100 beats per minute depending upon age and the heartbeat is controlled by electrical impulses. Arrhythmia is most common in middle age or above persons starting at the age of 25 years and the rate increase with age. There are several types of arrhythmia and most of them are not that dangerous to the extent of causing death. The harmless and inconsequential arrhythmias are known as benign but there are several types of arrhythmias which are dangerous to the health and can cause death. These types of potentially fatal arrhythmia are known as malignant arrhythmia which could lead to heart attack, stroke, dizziness and heart failure. The malignant arrhythmia requires urgent medical treatment and proper care. The main causes of arrhythmia include heart disease, stress, extensive use of alcohol and tobacco, and other causes as well.

In this paper, we consider nine different categories of arrhythmia considered a health hazard. Out of the 9 types, right bundle branch block (RBBB), premature ventricular contractions (PVC) and left bundle branch block (LBBB) are the most common malignant types. PVC is most common in middle-aged persons and may occur suddenly with or without heart disease. The causes of PVC are usually stress, extensive exercise and due to some heart conditions. RBBB arrhythmia is a condition in which electric impulses that regulate the heartbeats pumping blood to the body are delayed along the right side of the heart. RBBB mostly occurs in athletes and the people who do extensive exercise and usually aged less than 40 years. LBBB arrhythmia is a condition in which electric impulses are delayed or stopped in the left bundle branch of heart. In LBBB and RBBB, heart two ventricles are being stimulated with electrical impulse in sequence instead of simultaneously. Most of the time, LBBB are more dangerous and required urgent medical treatment. LBBB occurs in older adults under the age of 50 years. The list of 9 arrhythmias considered in this paper based on the MIT-BIH database are described in Table 1.

TABLE I. SUMMARY OF ECG CLASSES

#	**Symbol**	**Name**
1	PVC	Premature Ventricular Contractions
2	APB	Atrial Premature Beat
3	PB	Paced Beat
4	NEB	Nodal (junction) escape beat
5	FPNB	Fusion of Paced and Normal Beat
6	LBBB	Left Bundle Branch Block
7	RBBB	Right Bundle Branch Block
8	FVNB	Fusion of ventricular and normal beat
9	VFW	Ventricular Flutter Wave

A. *Distributed Computing (Systems)*

Simultaneous program execution on multiple computers is possible through dividing the program tasks for different computers. Thus, distributed computing is a network of systems for which each has a task from the original program [4]. Each computer on the network is referred to as a node that will solve one or more the tasks and are then passed from one node to another through a process called message passing. Distributed computing (systems) therefore consists of multiple computers, network, network topology and designed in such a way to tolerate failure and recover accordingly. Distributed systems are superior is terms of reliability and performance and allows for parallel use of GPU's and resources [5]. As a whole, distributed system can grow to form a Supercomputer. Supercomputers allow tasks to be run on different servers and can be made more secure with less cost.

B. *Cloud Computing*

Cloud computing is similar in concept to distributed computing in the sense of dividing tasks to different CPUs or GPUs with the difference that resources are remotely located in data centers spread across the world wide web [6]. Cloud computing has gained momentum recently with different services added to the so-called cloud computing industry. Various vendors offer cloud-computing services including but not limited to Amazon, Oracle, and many others. The complete potential of cloud computing is yet to be explored. This paper focus on performance analysis of using multiple Cloud GPU's for scientific computing purposes. Cloud computing comes with the advantage ready computing facilities meaning that the end user does not need to install operating systems and programs nor worry about maintenance and updates. The question whether this concept results in reduction in cost of computer hardware and datacenter operations for medium to large-scale institutions is still debatable in literature. Cloud computing once full potential is realized is expected to offer a performance advantage and cost advantage over other technologies such as super computers, clusters, and grids.

C. *Support Vector Machines (SVM)*

SVMs are learning models that analyze information and supports data classification. SVM can classify the data points within n-dimensional space depending on the feature numbers (n = number of features) by finding the hyper-plane. In addition, the hyper-plane position can be affected by the data points (support vector). So when the data points are located close to the hyper-plane, the orientation and the position will be affected accordingly. The main use of the hyper-plane is as a separator; putting the data into classes depending on the data point locations [7]. SVM construct a Platt scaling model with a representation of data as points in space meaning it divides the data clearly with wide gaps. Additionally, SVMs can perform both linear and non-linear classification via kernel method that works by matching and connecting data inputs to the high-dimensional spaces.

SVM is a supervised learning technique used for classification [8]. The primary objective of supervised classification achieving higher accuracy with lower number of sample for training, making the classification as cost-effective and simple as possible. SVM is considered more accurate when compared to other classification methods such as maximum likelihood. SVM works by fitting an optimal separating hyperplane (OSH) between the different classes. This is done by finding the selecting training samples that are the edge of each class. Thus only training samples deemed within the boundary of a certain class can be used in SVM.

In order to enable SVM to be applied more appropriately, suggested an approach in which all classes are considered at once. n two classes rules, while the mth function w_(y_i)^T φ(x_i)+b is used to separate the training vectors of the class m from other constructed classes creating only n decision function which can be obtained by:

$$\min_{w,b,\xi} \frac{1}{2} \sum_{m=1}^{n} w_m^T w_m + C \sum_{i=1}^{l} \sum_{m \neq y_i} \xi_i^m \, w_{y_i}^T \varphi(x_i) + b_{y_i} \geq w_{y_i}^T \varphi(x_i) + b_m + 2 - \xi_i^m$$

III. RELATED WORK

ECG-EEG classification and recognition of different heart diseases got significant attention of researchers in pattern recognition and biomedical engineering domains. Few supervised and semi-supervised classification techniques for ECG and EEG using heart beats as input records produce good results. In [9], the authors investigated different designs of neural networks and experimented with various learning algorithms for ECG classification. In [10], the authors proposed a method based on Multilayer Perceptron (MLP) which achieved an average 99.02% accuracy. The features of the ECG signals were extracted based on heart beat intervals, spectral entropy and RR intervals. Each data set is supposed to contain at least 1500 sample beats. In [11], the authors use SVM with enhanced kernel-based approach to classify the ECG signals and compared the results with and Linear Discriminant Analysis (LDA) based approach. In [12], linear and non-linear features of the ECG signals are used to classify the data into five types of beat classes of arrhythmia and achieved 98.91% accuracy.

Different studies that are based on SVM to detect arrhythmia in ECG signals were explored. A comparison between one-against-one and one-against-all multiclass SVM classification were performed and the one-against-all produced the best results in the case of arrhythmia detection [13-14]. In [15], a hybrid SVM system was implemented to classify different types of heart beat arrhythmias. The hybrid SVM model was composed of two multiclass SVMs and a binary SVM each dedicated to a different classification task. In this model an overall validation accuracy of 90.89% was achieved on the MIT-BIH data.

In [16], the authors proposed Random Forest (RF) along with data resampling method to enhance the arrhythmia classification. They used correlation-based feature selection technique for features selection of ECG dataset. The testing was done for 14 arrhythmias classes using a dataset containing 452 samples and 90.0% accuracy was achieved.

G. Sannino and G. Pietro used deep neural network for heartbeat classification. They used tools like tensor flow and Google deep learning. There proposed model used 7 hidden layers and the results have shown an effective output in terms of sensitivity [17]. Özal Yıldırım et al in their paper used a new approach based on long duration of ECG Signal fragments. 17 ECG class was used in their experiments using one dimensional convolutional neural network. Their model obtains 91.33% accuracy [18]. Pawel and Acharya in their paper [19] used deep genetic ensemble classifier. Their approach was a hybrid approach using advantages of ensemble learning, deep learning and evolutionary computation. Their model can be applied in cloud computing and cardiac health using high precision.

IV. Proposed Method

In this research, we present the use of Support Vector Machine for training and classification on a parallel and widely available inexpensive computing platform. Accelerating SVM can be done by excluding the non-support vectors early in the process. Using this concept, we propose a novel filtering process that enables the use of SVM on a parallel platform efficiently. A performance evaluation method is designed that permits the assessment of Parallel Support Vector on the cloud clusters connected through high-speed networks for large dataset classification. All measurements are performed on the EC2 environment [20].

A. Parallel SVM Algorithm

Cloud based Parallel SVM clusters operate as follow:

- Automated script for the EC2 High Performance Computing (HPC) instances initiate "N" N GPU clusters [4]. The first GPU's server act as the Master Node.
- The dataset is divided into equally sized subsets and distributed to the cluster nodes. The input dataset "D" with size "L" is used where i = 0, 1, 2… N-1.

$$L_i = \frac{i \times L}{N}$$

$$D_{i+1} = D(L_i + 1, L_{i+1})$$

- The output of the cluster after processed is merged and automated script will ensure that only the Master node remains on. The merged output is then input to the Master node.

$$D_{new} = \sum_{i=1}^{N} SVout_i$$

- The master node input the merged support vectors and applies the SVM to obtain more refined results.

B. Feature Extraction

Extracting features and selecting appropriate features is an important step achieve high classification results. In ECG dataset, clinical features are very important based on domain knowledge, which can help for arrhythmia detection. Features like P-wave, T-wave, amplitude, inter beat timing and QRS heights are very important. The time domain based selected ECG features in the proposed system includes PR, QRS, QT and RR intervals. The other selected feature includes mean, standard deviation, QT interval and RR interval.

V. Experimental Dataset

For experiments, MIT-BIH ECG dataset is used for the arrhythmia classification is used which consists of 109000 beats having labels of total 48 classes [2]. The summary of the used classes with their samples count is presented in Table 2. An algorithm that is part of the ecgpuwave PhysioToolkit software was used for segmentation of the QRS wave and waveform boundary intervals. The preprocessing is performed to discard the beats with missing segmentation values and/or missing class labels. Due to this preprocessing, out of about 109,400 total beats, 108,232 beats were available for use, or 98.6% of the original data. The training phase utilized the following classes: the normal beats and 9 types of arrhythmia.

TABLE II. Used ECG Classes with their Number of Samples

Type	Each Class Beat Samples
Normal	74384
APB	2356
PVC	6730
PB	6969
FPNB	974
Others	337
NEB	227
LBBB	8033
FVNB	790
RBBB	7205
VFW	226

VI. Results and Analysis

Table 3 shows the results obtained from the single node experiment. The processing time for the single node is represented by PT while the identified support vectors processing time is represented by ISV. Table 4 shows the initial results of proposed method using four cloud GPU clusters.

Table 5 summarizes the final results of the proposed method after combining the recognized support vectors from multiple GPUs and inputting them into an SVM for the refined results. Table 5 shows also the performance and accuracy comparison. TPT refers to the total processing time of the SVM on multiple parallel cloud cluster nodes.

TABLE III. SINGLE NODE PERFORMANCE ANALYSIS

#	Data Samples	# of Features	Single Node		
			PT (Seconds)	ISV (Seconds)	Accuracy %
1	10000	2	982	3620	85.12
2	16000	2	21422	5715	84.84
3	24000	2	79195	8407	84.97
4	22400	4	53052	8647	85.96
5	59535	8	83517	25074	96.79
6	108,232	11	136252	48544	97.32

A comparison of the accuracy for a single node and multiple cluster nodes is shown in Figure 1. S-accuracy is used to refer to the single node accuracy while M-accuracy is used to represent the multiple cluster node's accuracy. The efficiency achieved using the multiple cloud cluster nodes is shown in Figure 2. The results clearly indicate a reduction of processing time by up to 60% through the use of multiple cluster nodes.

Table 5 was used to show the results of the proposed method in comparison with similar techniques. Results clearly indicate that the proposed technique outperforms similar methods proposed in previous literature. As shown in the result in table 6 above the Parallel SVM was able to achieve 97.03% accuracy, sensitivity 97.43% and 95.87% specificity with a processing time of 68,984 seconds.

TABLE IV. FIRST STEP SUMMARY OF MULTIPLE CLOUD GPU'S PERFORMANCE IN TERM OF TIME (IN SECONDS)

#	Data Samples	# of Features	Multiple GPU's Clusters								
			GPU 1		*GPU 2*		*GPU 3*		*GPU 4*		*TSV*
			PT	*ISV*	*PT*	*ISV*	*PT*	*ISV*	*PT*	*ISV*	
1	10000	2	31.021	1001	24.772	964	18.939	1039	20.824	1015	4019
2	16000	2	58.139	1526	61.31	1591	52.27	1577	45.71	1566	6260
3	24000	2	200.94	2303	123.21	2286	135.26	2272	227.79	2219	9080
4	22400	4	1054.898	2428	1231.171	2420	910.6977	2363	2246.163	2500	9711
5	59535	8	13931	7979	14037	8773	8606.2	6046	12018	8254	31052
6	108,232	11	22371	14882	20476	13232	22981	15423	21767	14211	57748

TABLE V. FIRST STEP SUMMARY OF MULTIPLE CLOUD GPU'S PERFORMANCE IN TERM OF TIME (IN SECONDS)

#	Data Samples	# of Features	Multiple GPU's Clusters (P2)						
			Results of Multiple GPU Nodes to single GPU Node						
			TSV	PT	ISV	Acc.	TPT	Efficiency %	Acc. Effect
1	10000	2	4019	313.1	3494	85.09	344.121	64.88	0.035%
2	16000	2	6260	2102.75	5603	84.8	2164.06	89.89	0.047%
3	24000	2	9080	4959.9	8259	85.021	5187.69	93.45	-0.06%
4	22400	4	9711	25815.7	7959	85.92	28061.87	47.1	0.10%
5	59535	8	31052	36007	24467	96.67	50044	46.01	0.131%
6	108,232	11	57748	46003	46334	97.01	68984	49.37	0.0996%

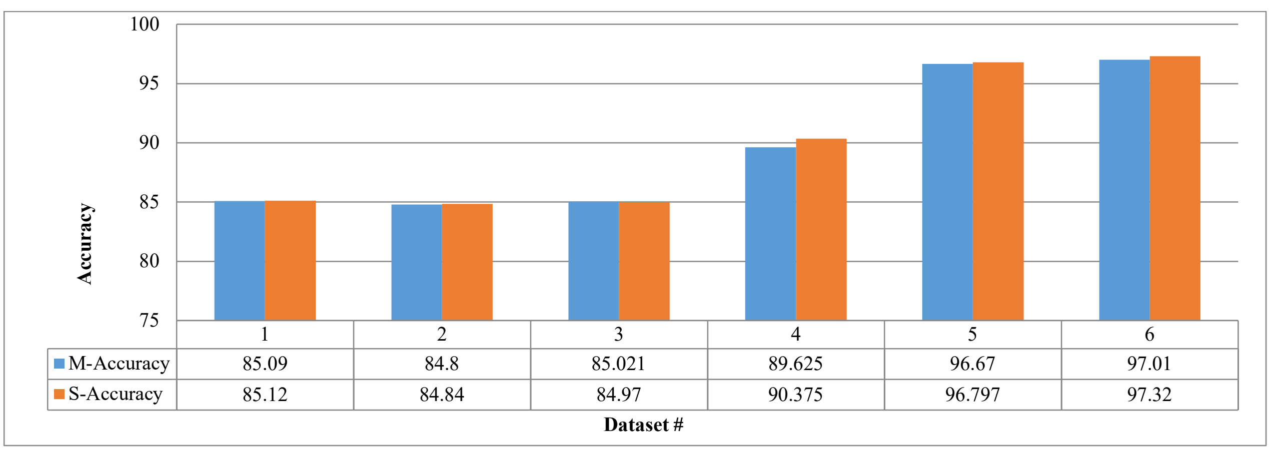

Figure 1. The comparision of expiremental results with Single GPU cluster and Multiple GPU Cluster nodes in terms of Accuracy.

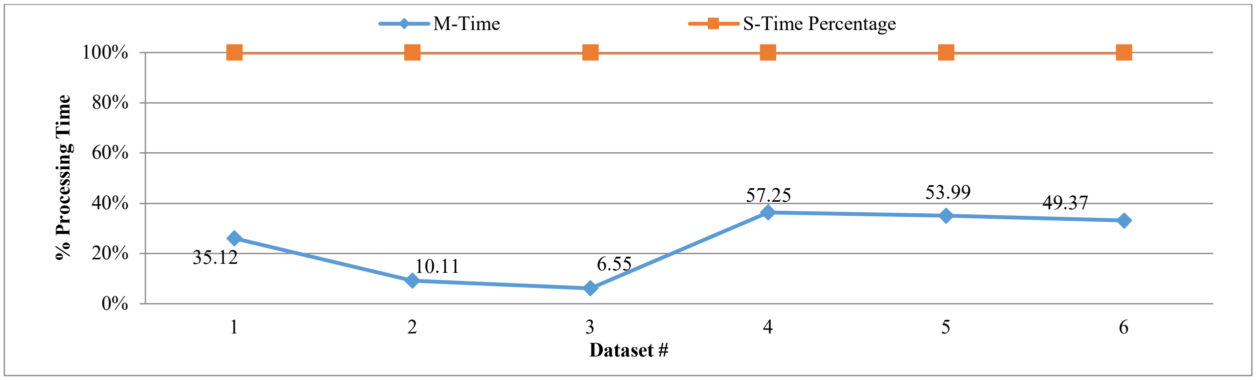

Figure 2. Time efficiency between single node and multiple nodes

TABLE VI. SUMMARY OF PERFORMANCE AND ACCURACY COMPARISON FOR ECG REAL DATASET FOR DIFFERENT CLASSIFIERS

Classifier	Average Accuracy(%)	Average Sensitivity(%)	Average Specificity(%)	Processing Time (Seconds)
RBF-Network	92.00	91.90	90.80	156395
FFBP NN	95.54	95.50	94.80	95602
Single Node SVM	97.32	97.86	96.23	136252
Parallel SVM	97.03	97.43	95.87	68984

CONCLUSION

This research proposes the use of parallel SVM on cloud-based cluster for the purpose of big data classification. The results clearly indicate that the proposed method in this paper outperforms similar techniques in terms of training time and classifies the data with excellent accuracy. The method is both robust and scalable. Experimental results were generated to show the superiority of the proposed method over its others proposed in previous literature. Future work will include more sophisticated ensemble methods for the classification of the dataset used in this paper. The research will also include applying the proposed method in this paper for the classification of other datasets in various interdisciplinary fields.

REFERENCES

[1] Yune, Ho Young, et al. "Delta neutrophil index as a promising prognostic marker in out of hospital cardiac arrest." PloS one 10.3 (2015): e0120677.

[2] Van der Lende, Marije, et al. "Cardiac arrhythmias during or after epileptic seizures." J Neurol Neurosurg Psychiatry 87.1 (2016): 69-74.

[3] Rajpurkar, Pranav, et al. "Cardiologist-level arrhythmia detection with convolutional neural networks." arXiv preprint arXiv:1707.01836 (2017).

[4] Sadooghi, Iman, et al. "Understanding the performance and potential of cloud computing for scientific applications." IEEE Transactions on Cloud Computing 5.2 (2015): 358-371.

[5] Haque, M. S. M., Latif, G., Hasan, M. R., Arifuzzaman, M., Shafin, S. S., & Rahman, Q. A. (2019). Scalable parallel SVM on cloud clusters for large datasets classification.

[6] Chen, Xuhui, et al. "Real-time personalized cardiac arrhythmia detection and diagnosis: A cloud computing architecture." 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2017.

[7] Luz, Eduardo José da S., et al. "ECG-based heartbeat classification for arrhythmia detection: A survey." Computer methods and programs in biomedicine 127 (2016): 144-164.

[8] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm.

[9] Jambukia, Shweta H., Vipul K. Dabhi, and Harshadkumar B. Prajapati. "Classification of ECG signals using machine learning techniques: A survey." 2015 International Conference on Advances in Computer Engineering and Applications. IEEE, 2015.

[10] Issac Niwas, S.; Shantha Selva Kumari, R.; Sadasivam, V.; , "Artificial neural network based automatic cardiac abnormalities classification," Computational Intelligence and Multimedia Applications, 2005. Sixth International Conference on , vol., no., pp. 41- 46, 16-18 Aug. 2005.

[11] Varatharajan, R., Gunasekaran Manogaran, and M. K. Priyan. "A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing." Multimedia Tools and Applications 77.8 (2018): 10195-10215.

[12] Elhaj, Fatin A., et al. "Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals." Computer methods and programs in biomedicine 127 (2016): 52-63.

[13] C. Caragea, D. Caragea, and V. Honavar, "Learning support vector machine classifiers from distributed data sources," in Proc. 20th Nat. Conf. Artif. Intell. Student Abstract Poster Program, Pittsburgh, PA, 2005, pp. 1602–1603.

[14] Latif, G., Al Anezi, F. Y., Zikria, M., & Alghazo, J. (2020, January). EEG-ECG Signals Classification for Arrhythmia Detection using Decision Trees. In *2020 Fourth International Conference on Inventive Systems and Control (ICISC)* (pp. 192-196). IEEE.

[15] A. Navia-Vazquez, D. Gutierrez-Gonzalez, E. Parrado-Hernandez, and J. Navarro-Abellan, "Distributed support vector machines," IEEE Trans. Neural Netw., vol. 17, no. 4, pp. 1091–1097, Jul. 2006.

[16] Özçift, Akin. "Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis." Computers in biology and medicine 41, no. 5 (2011): 265-271.

[17] G. Sannino and G. Pietro. "A deep Learning approach for ECG-Based heartbeat classificaition for arrythmia detection". Future Generation Computer Systems, Volume 86, September 2018, Pages 446-455. 2018.

[18] ÖzalYıldırım et al. "Arrhythmia detection using deep convolutional neural network with long duration ECG signals". Computers in Biology and Medicine. Volume 102, Pages 411-420, 2018.

[19] Paweł Pławiak & U. Rajendra Acharya. "Novel deep genetic ensemble of classifiers for arrhythmia detection using ECG signals". Computer aided medical diagnosis. Neural Computing and Applications (2020) 32:11137–1116. 2020.

[20] Gallagher, J. H. R., and N. Potter. "AWS-Glacier As A Storage Foundation For AWS-EC2 Hosted Scientific Data Services." AGU Fall Meeting Abstracts. 2016.