European Journal of Computer Science and Information Technology 11 (4), 1-21, 2023
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

A New Index for Intelligent Classification of Early Syndromic of Cardiovascular (CVD) Diseases Based on Electrocardiogram (ECG)

Imeh Umoren[1*] Oluwaseyi Abe, Godwin Ansa[3], Saviour Inyang[4], Idongesit Umoh[5]

[1,2,3,4] Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria
[5] *Department of Cardiology, University of Uyo Teaching Hospital, (UUTH) Uyo, Nigeria*

doi:https://doi.org/10.37745/ejcsit.2013/vol11n4121 Published July 16 2023

Citation: Umoren I., Abe O., Ansa G., Inyang S., and Umoh I. (2023) A New Index for Intelligent Classification of Early Syndromic of Cardiovascular (CVD) Diseases Based on Electrocardiogram (ECG), *European Journal of Computer Science and Information Technology*, 11 (4), 1-21

ABSTRACT: *Most disease that affects the heart or blood vessels is referred to as cardiovascular disease(CVD). The main aim of this work is to build a system capable of modeling and predicting early syndromic cardiovascular diseases (CVD) based on electrocardiogram (ECG). The study considers the implementation of computationally intelligent system for detecting and classifying early syndromic assessment of CVD. The clinical and ECG recordings of patients diagnosed with pulmonary hypertension at the University of Uyo Teaching Hospital (UUTH) were obtained. The datasets were segmented into Demographic and ECG datasets. A quantitative research approach was used for the study with examination of several segments based on recommended framework. Three (3) classifier models were adopted to detect cardiac related problems using specified datasets. The classifiers such as; Random Forest Ensemble (RFE), Support Vector (SVM) Classifier and Artificial Neural Network (ANN) was employed for Machine Learning process. The models were implemented using a robust programming languages (Python and Jupyter notebook). The datasets were further segmented into two categories: training sets and testing in the ratio of 80:20 respectively. The test data reflects; precision, recall and sensitivity: Results show Radom Forest Model: 0.50 (50%) accuracy, 0.48 (48%) precision score and 0.65 (65%) recall sensitivity score (RSS), SVM classifier indicated 0.70 (70%) accuracy score, 0.47 (47) % precision score as well as 0.52 (52 %) sensitivity score. The ANN model illustrates 0.50 (50%) score for accuracy, precision and recall. Research Findings demonstrated that, RFE, SVM, ANN illustrate 100% accuracy in precision and recall sensitivity. The interaction effects of the various clinical factors influencing the CVD of patient was appraised and performance evaluation were further done using standard data science measures; Confusion Matrix (CM), MAP, MAPE, RMSE was deployed. The final results obtained shows that RFE, SVM, ANN models support satisfactorily the assessment and classification of early syndromic conditions of CVD.*

KEYWORDS: ECG, UUTH, RFE, SVM Classifier, ANN, accuracy, precision score and recall sensitivity score (RSS).

INTRODUCTION

Any disease that affects the heart or blood vessels is referred to as "cardiovascular disease". CVD is sometimes thought to be an issue that only affects the adults. It is, nevertheless, more frequent than most people believe among teens and young adults. Anyone, at any age, can be affected. Young people are frequently ignorant that they are at risk and may fail to take the necessary precautions to save their life. One of the greatest method to help avoid death and lessen issues connected with heart disease is to educate parents, adolescents, and young adults about the many risk factors with regards to heart disease. Nevertheless, with the advent of technology today different models have been proven effective in health care as such can help actualize the effective and efficient management of different heart diseases such as hypertension. Furthermore, obesity in children, in particular, has fast become a global epidemic, with one out of every ten teen or youth considered to be obese. Dyslipidemia (high cholesterol), hypertension (high blood pressure), type 2 diabetes, and metabolic syndrome are all risk factors for cardiovascular disease caused by obesity. If these diseases are not addressed, young adults may develop early cardiovascular disease, which can lead to serious health concerns. In addition, undiagnosed or untreated congenital heart malformations and anomalies can cause cardiac disease in young adults [18]. There are numerous cases where both healthcare providers or managers and healthcare users going through critical deprivation - on one hand and healthcare managers are unable to do their jobs while on the other hand, the poor masses requiring healthcare services and other healthcare users suffer health challenges and even death. This critical challenge that has been recorded in this region can be traceable to unavailability of very required medical consumables, including medical professionals like nurses and doctors, often resulting from many factors, one of which is lack of logistic framework for Supply Chain in the south-south, Nigeria, resulting in many deaths recorded in the emergency units of the hospitals [19].
Again, a key characteristic of the south-south rural area of this region in terms of its geographical nature, is that it is swampy and thus provides enough breeding ponds for mosquitoes to thrive - as a result the outbreak of malaria is high and the deaths recorded from malaria in this region is higher than those obtained elsewhere within the country [6]. This research project is primarily concerned with hypertension, which is a major risk factor for cardiovascular disease. Despite the fact that reducing blood pressure can reduce hypertension risk in certain people, only approximately half of hypertensive patients receive medical treatment, and persons with high normal blood pressure or mild hypertension account for more than half of all cardiovascular events. As a result, primary hypertension prevention is a promising strategy for lowering cardiovascular morbidity and death. A recent clinical research found that those with prehypertension who were given an angiotensin receptor blocker for two years had a 15% lower risk of developing hypertension over four years than those who were given a placebo [14] Hypertension can also be reduced by changing one's lifestyle, eating a low-salt diet, and exercising regularly. Although the possibility and efficacy of avoiding hypertension has been shown, a strategy that targets all people who do not have hypertension would increase the long-term use of medical and financial resources. A personalized strategy to risk classification and

focused treatment of normotensive individuals who are most at risk of developing hypertension could thus be more effective in avoiding hypertension in the general population. As a result, even a minor rise in ECG voltage, which signals an increase in left ventricular mass, could be a precursor to hypertension. In this study, we will see if electrocardiography (ECG) can be used to predict the development of hypertension in the general population who are not hypertensive. The heart is a primary target organ for elevated blood pressure, and cardiac muscle hypertrophy occurs in response to increased afterload (i.e., systemic blood pressure), [7]. Throughout the development of hypertension, blood pressure progressively rises with large swings, indicating that the heart may be subjected to higher afterload transiently but often during the early stages of hypertension. Basically, medical proof suggests that early detection of hypertension, lifestyle changes, and strict control could help to limit the disease's progression and repercussions [4]. With no well-defined cut points, effective management of hypertension necessitates the establishment of guidelines and thresholds, as well as treatment recommendations, for a condition that has shown a persistent and graded association to the risk of cardiovascular disease (CVD), [4].

RELATED WORKS

In order to contribute to the constant advancement in the field of medical sciences, [18] implemented an automated approach for the diagnosis and give possible recommendations using Fuzzy C Means (FCM) clustering techniques for high precision and reliable clinical supervision. [10], proposed a method for BP estimation using only electrocardiogram (ECG) signals. Before being submitted to a complexity analysis for feature extraction, raw ECG data is filtered and segmented. A machine-learning technique is then utilized to create systolic, diastolic, and mean arterial pressure (MAP) prediction models, which incorporates a stacking-based classification module with a regression module. To adjust the models to a given user in the proposed system, the technique uses a probability distribution-based calibration. The proposed method produced BP estimation results that were comparable to those of a certified medical equipment. [9] designed database of 139 Holter recordings with clinical data of hypertensive patients followed up for at least 12 months were collected ad hoc. Subjects to who experienced a vascular event (i.e., myocardial infarction, stroke, syncopal event) were considered as high-risk subjects. Several data-mining algorithms (such as support vector machine, tree-based classifier, artificial neural network) were used to develop automatic classifiers and their accuracy was tested by assessing the receiver-operator characteristics curve and tested the echographic parameters, which have been showed as powerful predictors of future vascular events. [11], proposed a perceptron on multi-layered method with numerous neural network layers for effective detection of handwritten digits. A model for electrocardiographic strain pattern and prediction of cardiovascular morbidity and death in hypertension patients was put up by [12]. A marker for left ventricular hypertrophy (LVH) and a poor prognosis in population research is the ECG strain pattern of lateral ST depression and T-wave inversion. 8854 hypertensive individuals with ECG LVH who were given blinded atenolol or losartan-based regimens had their ECGs analysed at the start of the trial. 971 individuals had strain, which was identified by the presence of a downsloping convex ST segment and an inverted asymmetrical T wave opposing the QRS axis in leads V5 and/or V6 (11.0%). 1035 participants experienced the composite end outcome of cardiovascular mortality, nonfatal myocardial infarction, or stroke in the Losartan Intervention for Endpoint

Reduction in Hypertension (LIFE) research (11.7%). The article shows additive value of standard ECG for the risk prediction of hypertensive disorders during pregnancy, [1] tested a number of demographic, clinical, and laboratory variables as predictors of a composite pool of pre specified events, including gestational hypertension, preeclampsia, or eclampsia. The link between the numerous demographic, clinical, laboratory, and ECG characteristics and outcomes was examined using logistic regression. To find the univariable predictors of hypertensive diseases, they evaluated a baseline model. [13], proposed an Electrocardiogram-based scoring system for predicting secondary pulmonary hypertension. Between 2006 and 2009, 552 consecutive patients underwent right cardiac catheterization. In the research group, 297 individuals overall (54%) had pulmonary hypertension. The validation cohort was made up of the remaining 220 study cohort patients, while the development cohort was composed of a total of 332 study cohort patients. RAE, LAE, RAD, and R-wave in V1 6 mm were given values of 5, 2, and 1, respectively, based on log odds ratios of association in the development cohort to create a 10-point scoring system called the "Scranton PHT (SP) score. For differentiating pulmonary hypertension, SP scores of 5 points and 7 points in DC exhibited C-statistics of 0.83 and 0.89, respectively. When compared to an SP score of 7, the C-statistic for RAE alone was considerably lower (0.83 vs. 0.89, P = 0.021). The validation cohort's SP score reliability was adequate. When patients have a clinical suspicion of pulmonary hypertension, the SP score is a useful point-of-care tool for predicting the condition. [8], proposed a model for Electrocardiogram (ECG) Classification Using Convolutional Neural Network (CNN). It shows how to classify ECG signals using a convolutional neural network method. The algorithm makes use of an ECG dataset obtained from kaggle.com. The signals in the dataset were preprocessed to ensure that each segment matched a heartbeat. The suggested model was trained employing a convolutional neural network approach that included six total hidden layers, a hidden size of 128, a batch size of 96, and a number of epoch of ten. Individuals can get access to it and make classifications themselves after completing successful training at the epoch level of ten (10). The work of [4] presented metric for efficient prediction and adopted a logistic regression model for predicting hypertension severity index based on five parameters such as Sugar level (FBS), Cholesterol, Age, gender and Body Mass Index (BMI). [17], established that, the implementation of the AI–ECG is still in its infancy, but a continuously growing clinical investigation agenda will determine the added value of these AI tools, their optimal deployment in the clinical arena, their multifaceted and so-far largely unpredictable implications. As with any medical tool, the AI–ECG must be vetted, validated and verified, and clinicians must be trained to use it properly, but when integrated into medical practice, the AI–ECG holds the promise to transform clinical care

PROPOSED SYSTEM ARCHITECTURE

The Proposed System Architecture is presented in in figure 1.

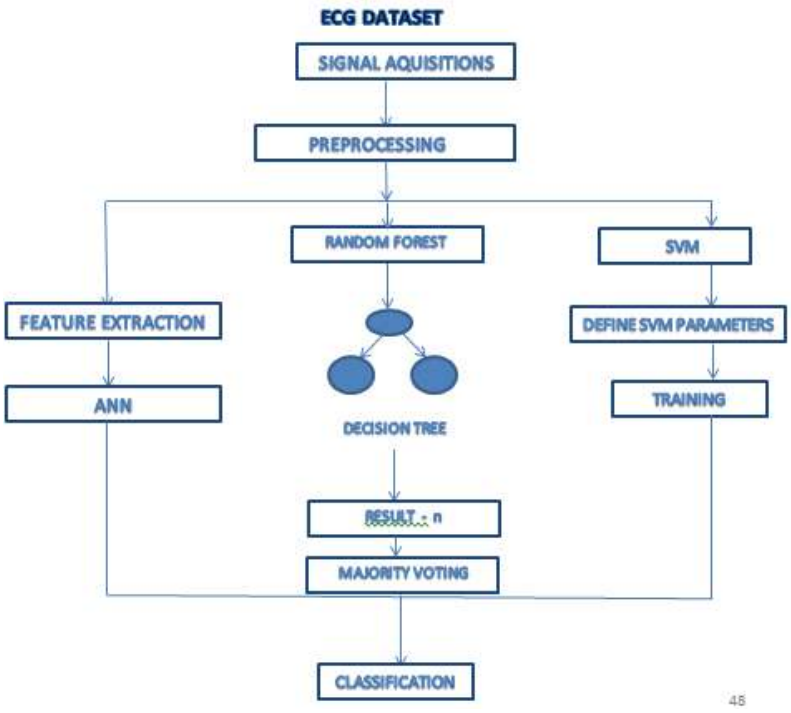

Fig. 1: Proposed System Architecture

MATERIALS AND METHODS

Data Collection

The data used in this research was obtained from the University of Uyo Teaching Hospital. The data consist of electrocardiogram result of 50 patients diagnosed of hypertension disease. A collection of scanned files of the patients' records admitted and diagnosed for hypertension were analyzed and data extraction was performed. The electrocardiogram (ECG) captures electrical activity on the surface of the body that emerge from the heart, and the signals can offer information about the heart's rhythm, anatomy, and function [1]. [2]. Using advanced analysis, ECG data in HT people can be linked to blood pressure readings and even used to identify patients at increased clinical risk [16]. The data collected was deemed to be accurate and verified by cardiac specialist in charge of collection and recording of ECG data. The data was divided into two parts: the demographic information and ECG as the predictive variables. Two variables (age and sex) were used as the demographic information. These two variables are simple and were obtained directly from the patience files. The features of ECG data such as heart rate, presence of atrial fibrillation or flutter, PR interval, QT interval, corrected QT interval, QRS duration, P, R and T- wave axes, Blood pressure were used with the demographic information to develop machine learning algorithms. Table 1 shows a sample row of the raw data gathered for this research project.

Table 1(a) Sample of ECG data I

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
17	63	M	67	152	78	392	404	44	14	1.44	NORMAL	100	65	11	BAD	0
18	73	M	87	166	104	388	466	72	-73	4.219	ABNORMA	140	100	1	GOOD	1
19	60	F	75	180	145	428	480	50	-85	1.355	NORMAL	160	100	4	GOOD	0
20	57	F	90	130	88	344	420	80	35	1.814	ABNORMA	190	100	2	GOOD	0
21	67	F	76	154	82	404	454	52	52	3.669	BODERLIN	130	70	1	GOOD	0
22	59	F	122	155	98	308	438	0	92	2.074	ABNORMA	80	70	3	BAD	2
23	49	F	85	158	94	334	396	60	1	1.61	NORMAL	80	60	4	BAD	0
24	79	F	81	164	86	374	434	48	59	2.132	ABNORMA	160	110	1	GOOD	0
25	42	F	53	250	114	460	432	54	-2	1.428	ABNORMA	110	90	2	GOOD	2
26	69	F	53	140	158	476	446	0	50	1.414	ABNORMA	170	80	2	GOOD	2
27	59	F	86	146	132	400	478	37	27	1.575	ABNORMA	180	100	9	GOOD	1
28	70	M	52	212	82	424	394	56	34	2.117	ABNORMA	110	90	1	GOOD	2
29	52	M	76	146	85	414	467	43	-82	2.345	ABNORMA	200	120	7	GOOD	0
30	71	M	65	228	92	378	392	65	60	2.911	ABNORMA	110	70	22	GOOD	2
31	62	M	122	82	96	378	538	138	60	0.437	ABNORMA	130	85	1	GOOD	0
32	37	M	103	142	96	412	538	33	-154	5.05	ABNORMA	120	80	1	GOOD	1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	AGE	GENDER	HR	PR	QRS	QT	QTC	P	T	RV5+SV1	ANA_RES	FBP_H	FBP_L	YR_INT	RESPONSE	OUTCOME
2	56	M	89	192	94	344	418	54	21	2.012	ABNORMA	130	70	3	GOOD	0
3	43	F	64	136	87	419	436	20	61	2.06	NORMAL	180	110	4	GOOD	0
4	24	F	94	192	94	362	452	53	-20	5.911	ABNORMA	130	90	1	BAD	1
5	69	F	68	140	128	442	470	59	37	2.092	ABNORMA	120	80	1	BAD	2
6	35	F	93	172	86	330	410	61	44	3.571	NORMAL	100	60	8	BAD	0
7	74	F	64	136	87	419	436	20	61	2.06	NORMAL	150	100	11	GOOD	0
8	42	F	53	250	114	460	432	54	-2	1.428	ABNORMA	100	70	1	GOOD	2
9	74	M	68	180	96	390	414	68	52	2.574	ABNORMA	140	70	13	BAD	1
10	42	M	98	148	100	426	544	80	87	2.598	ABNORMA	140	90	1	GOOD	0
11	71	M	33	140	184	264	194	0	-108	0.692	ABNORMA	140	80	2	GOOD	2
12	59	F	122	150	98	308	438	0	92	2.074	ABNORMA	100	80	3	BAD	2
13	49	F	85	158	94	334	396	60	1	1.61	NORMAL	80	60	4	BAD	0
14	54	F	84	167	85	376	447	14	34	1.61	NORMAL	110	60	21	BAD	0
15	73	F	77	135	85	355	404	61	57	2.07	ABNORMA	110	70	7	BAD	1
16	66	M	70	184	104	332	358	29	-32	2.114	ABNORMA	160	90	3	GOOD	1

zTable 1(b) Sample of ECG data II

Dataset

A dataset is a collection of information or data that is presented in an organized manner and may be changed as a unit of data. A dataset may also be thought of as a grouping of data items such as records, observations, variables, events, or entities. Data objects are described using a variety of attributes. Data objects include characteristics such as types, timeseries, dimensions, and so on. The multivariate dataset utilized in this study has 50 observations and 15 feature

variables. The ECG recordings of patients diagnosed with pulmonary hypertension at the University of Uyo Teaching Hospital are included in this collection. The dataset is described in Table 2.

The dataset was divided into two parts: Demographic and ECG data. Two variables described the demographic of the patients while the remain variables are the 12 lead measurement of electrocardiogram machine.

Table 2: Dataset Features and Description

S/No.	Attributes	Description	Data Type	Attribute Category
1	Age	Age of a Patient	Integer	Demographic Attribute
2	Gender	Gender (Male or Female)	Categorical	
3	Heart rate (HR)	Patient's HR-interval between two successive QRS complexes measured in beats per minute (bpm)	Integer	ECG Attribute
4	PR Interval (PR)	Patient's PR interval in milliseconds. The measurement start from the beginning of the P wave to the first part of the QRS complex	Integer	
5	QRS Complex	The QRS duration represent the time for ventricular depolarization	Integer	
6	P wave	The atrial depolarization	Integer	
7	QT	QT interval measures the depolarization and repolarization of the ventricles.	Integer	
8	QTC	QTC is a corrected QT interval. QTC is the QT interval divided by the square root of the PR interval	Integer	
9	R wave	R wave progression across V1 to V6	Float	
10	Analysis result	Rhythm described as abnormal or normal sinus	Categorical	
11	Year of Interval	Year of interval of diagnosis	Integer	
12	FBP	First Systolic/ Diastolic blood pressure	Integer	
13	LBP	Last Systolic/ Diastolic blood pressure	Integer	
14	Response	Patient's response to treatment	Categorical	
15	Outcome	Outcome of detection of pulmonary hypertension	Categorical	

MODEL FORMULATION USING RANDOM FOREST

The random forest is based on applying bagging to decision trees, with one crucial innovation: the algorithm samples the variables in addition to the records. In classic decision trees, the algorithm chooses the variable and split point by minimizing a criterion such as Gini or the residual sum of squares to determine how to form a sub-partition of a partition A. The choice

of variable in random forests is limited to a random subset of variables at each stage of the process. At each split, the random forest algorithm adds two further steps: bagging and bootstrap sampling of variables.

The steps for developing the random forest model are as follows:

1. From the medical records, we take a bootstrap (with replacement) subsample.
2. Sample $p < P$ variables at random without replacement for the first split.
3. The splitting procedure is applied to each of the sampled variables in the dataset $X_{i_1} \dots X_{j(p)}$.
   i. For each split values $s\_j(k)$ of $X_{j(k)}$:
   a. Split the records in partition A, with $X_{j(k)} < s_{j(k)}$ as one partition and the remaining records where $X_{j(k)} \geq s_{j(k)}$ as another partition
   b. We measure the homogeneity of classes within each partition of A
   ii. We choose the $s_{j(k)}$ value that generates the split value $s_{j(k)}$ with the highest class homogeneity within partition.
4. The next step is to choose the variable $X_{j(k)}$and separate the values $s_{j(k)}$ that create the most class homogeneity into partitions.
5. We go on to the next divide and repeat the actions from step 2 to step 3.
6. We repeat the process with more splits until the tree is fully grown.
7. We return to step 1 to take a new bootstrap subsample and repeat the procedure.

The issues of how to pick how many variables to sample at each phase will be essential. At each phase, we utilized three variables to sample. This followed a general rule of thumb of choosing $\sqrt{P}$, where P is the number of predictor variables.

Model Formulation Using Artificial Neural Network (ANN) Perceptron

One of the most basic ANN structures is the perceptron. It is based on a slightly different neuron known as the threshold logic unit (TLU), [6], which is also known as the linear threshold unit (LTU). Each input connection has a weight, and the inputs and outputs are integers rather than binary on/off values.

The TLU computes the weight sum of its inputs $z = w_1x_1 + w_2x_2 + \cdots + w_nx_n = x^Tw$, we can then apply a step function to the sum and the outputs the result:

$h_w(x) = step(z)$, where $z = x^Tw \quad \rightarrow (i)$

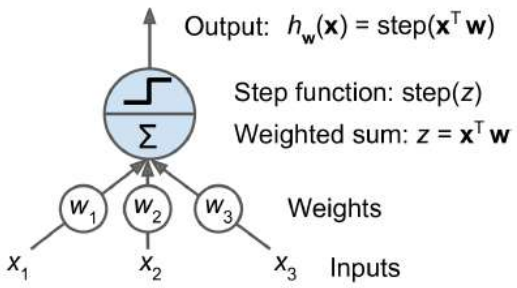

Fig. 2: Threshold Logic Unit

For basic linear binary classification, a single TLU can be utilized. It computes a linear combination of the inputs and outputs the positive or negative class depending on whether the result exceeds a threshold (just like a Logistic Regression classifier or a linear SVM).

A Perceptron is made up of a single layer of TLUs, each of which is coupled to all of the inputs. A completely connected layer, also known as a dense layer, is one in which all of the neurons in that layer are linked to every neuron in the preceding layer, which are the input neurons. It is usual to draw special passthrough neurons called input neurons (figure 3) to reflect the fact that each input is passed to each TLU. They simply output whatever input they are supplied. The input layer is made up of all the input neurons. Furthermore, an extra bias feature (x0 = 1) is commonly added: it is normally represented using a particular type of neuron known as a bias neuron, which always outputs 1. Figure 3 depicts a Perceptron with two inputs and three outputs. This Perceptron is a multi-output classifier since it can classify examples into three separate binary classes at the same time.

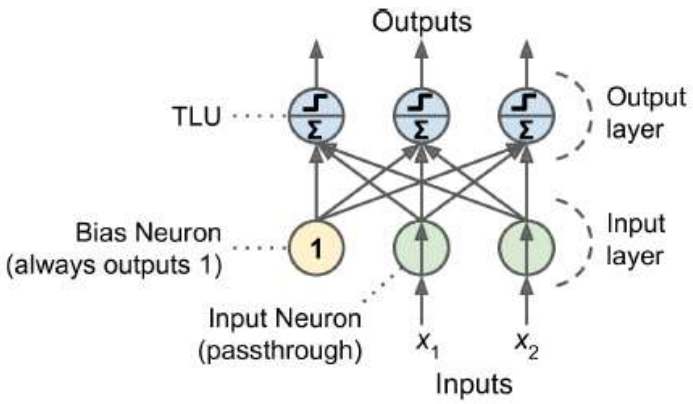

Fig 3: Perceptron Diagram

The outputs of a layer of artificial neurons may then be efficiently computed for several instances at the same time using $h_{wb}(X) = \phi(XW + b)$,

1. The matrix of input features is represented by X, which has one row per instance and one column per feature.
2. Except for the bias neuron's connection weights, the weight matrix W contains all of the connection weights. The layer comprises one row for each input neuron and one column for each artificial neuron.
3. All of the connection weights between the bias neuron and the fake neurons are contained in the bias vector b. Each artificial neuron contains one bias term.
4. The activation function is called the function ϕ: when the artificial neurons are TLUs, it is a step function.

perceptron are trained using a rule that considers the network's error and reinforces connections that help minimize the error. In more detail, the Perceptron is fed one training instance at a time and generates predictions for each one. It reinforces the connection weights from the inputs that would have contributed to the right prediction for every output neuron that made an erroneous prediction. The following equation represents the perceptron learning rule: $w_{ij}^{next\ step} = w_{ij} + \eta(y_j - \hat{y}_i)x_i$ where,

1. w_{ij} is the connection weight between the i^{th} input neuron and the j^{th} output neuron
2. x_i is the i^{th} input value of the current training instance.
3. $\hat{y}_i$ is the output of the j^{th} output neuron for the current training instance.
4. y_i is the target output of the j^{th} output neuron for the current training instance.
5. η is the learning rate.

Multi-Layer Perceptron and Backpropagation

A multi-layer perceptron may be mathematically modeled using the mathematical model of a single perceptron. Figure 4 shows a multi-layer perceptron (MLP) with one (passthrough) input layer, one or more hidden layers of TLUs, and one final layer of TLUs termed the output layer. Lower layers are those closest to the input layer, while upper levels are those closest to the outputs. A bias neuron is included in every layer except the output layer and is fully coupled to the following layer. `

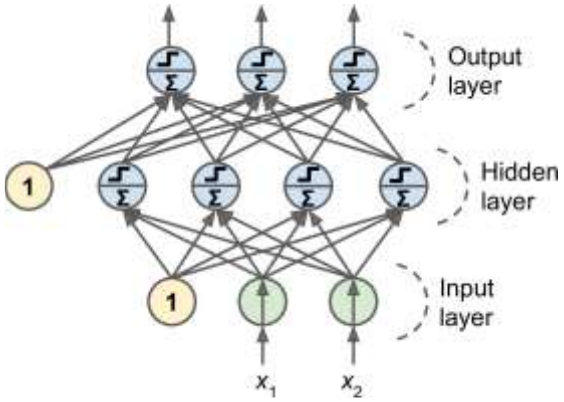

Fig 4. Multi-layer Perceptron

The backpropagation technique is employed in multi-layer perceptron learning. It relies solely on gradient descent. Gradient descent is an effective way for computing gradients automatically: the backpropagation algorithm can compute the gradient of the network's error for each model parameter in just two runs through the network (one forward, one backward). To put it another way, it can figure out how each connection weight and bias term should be adjusted to decrease error. It just conducts a conventional Gradient Descent step once it obtains these gradients, and the procedure is continued until the network converges to the solution.

The following are the complete details of the artificial neural network utilized in this project:

1. It only works with one mini-batch at a time (each having 50 examples), and it runs through the whole training set numerous times. Each pass is referred to as an epoch.
2. The network's input layer receives each mini-batch and delivers it to the first hidden layer. The output of all the neurons in this layer is then computed by the algorithm (for every instance in the mini-batch). The result is sent on to the next layer, which computes and passes on its output to the next layer, and so on until we reach the last layer's output, the output layer. This is the forward pass: it's precisely like making predictions, except that all intermediate outcomes are saved since the backward pass requires them.

3. The algorithm then calculates the output error of the network (that is, it uses a loss function that compares the desired output and the actual output of the network, and returns some measure of the error).
4. It then calculates the contribution of each output link to the error. This is accomplished analytically using the chain rule, which makes this process fast and precise.
5. The algorithm then uses the chain rule to determine how much of these error contributions came from each connection in the layer below, and so on until it reaches the input layer. By propagating the error gradient backward through the network, this reverse pass effectively measures the error gradient over all of the connection weights in the network.

Finally, the method uses the error gradients it just generated to tweak all of the network's link weights in a Gradient Descent step. The artificial neural network (ANN) was used to predict if each patient had a mild, moderate, or severe sign of cardiovascular disease. In this situation, three output neutrons were required. We had only one output neuron per class, and a softmax activation function for the entire output layer was used. Each instance of the prognosis belonged to a single class (the prognosis is mutually exclusive), out of three possible classes (0,1,2 corresponding to mild, moderate, and severe respectively). The softmax function assures that all predicted probabilities are in the range of 0 to 1 and add up to one. When the classes are exclusive, like in the case of the study dataset, this is essential.

Model Formulation Using Support Vector Machine (SVM)

The support vector machine is an extension of the support vector classifier that emerges from a specific approach of extending the feature space using kernels. The answer in a support vector classifier requires just the inner products of the observations rather than the data themselves. The inner product of two r-vectors a and b is defined as

$$(a, b) = \sum_{i=1}^{r} a_i b_i \quad \rightarrow (i)$$

Therefore, the inner product of the two observations x_i, x_j can be expressed as:

$$\left(x_i, x_j\right) = \sum_{j=1}^{p} x_{ij} x'_{ij} \quad \rightarrow (ii)$$

It can be shown that the linear support vector classifier can be represented as

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i (x, x_i) \quad \rightarrow (iii)$$

where there are n parameters α_i, $i = 1, \ldots n$ one per training observation. To estimate the parameters $\alpha_1, \ldots, \alpha_n$ and β_0, all we need are the $\binom{n}{2}$inner products $\langle x_i, x_i' \rangle$ between all pairs of training observations.

Consider replacing the support vector classifier function $f(x)$, with a generalization of the inner product of the form $K(x_i, x_i')$,, where K is a kernel function. A kernel is a function that measures how similar two observations are. For example, a kernel may be represented as

$$K(x_i, x_i') = \sum_{j=1}^{p} x_{ij} x'_{ij} \quad \rightarrow (iv)$$

which returns the support vector classifier, Because the support vector classifier's features are linear, this equation is known as the linear kernel; the linear kernel effectively measures the similarity of two observations using Pearson (standard) correlation. We can now select an alternative kernel form to replace each instance of $\sum_{j=1}^{p} x_{ij} x'_{ij}$with the quality.

$$K(x_i, x_i') = \left(1 + \sum_{j=1}^{p} x_{ij}, x'_{ij}\right)^d \quad \rightarrow (v)$$

The polynomial kernel of degree d, where d is a positive integer, is known as d. In the support vector classifier technique, using such a kernel with $d > 1$ instead of the normal kernel results in a significantly more flexible decision boundary. It simply entails training a support vector classifier in a higher-dimensional space utilizing polynomials of degree d instead of the original feature space. A support vector machine is a classifier that is created by combining a support vector classifier with a non-linear kernel. Aside from the polynomial kernel, another non-linear kernel option is the radial kernel, which has the following form:

$$K(x_i, x_i') = \exp(\gamma \sum_{j=1}^{p} (x_{ij} - x_{ij}')^2) \rightarrow (vi)$$

To identify the optimum estimator model and model parameters, linear polynomial and radial kernels were developed using the grid search function from the sci-kit learn python packages.

MODEL PERFORMANCE AND EVALUATION

In predictive modeling, it is usual to train a variety of models, apply them to a holdout sample, and evaluate their performance. After a number of models have been reviewed and refined, and if there is sufficient data, a third holdout sample, referred to as test data, is sometimes used to estimate how the chosen model would perform with entirely fresh data. The goal of the model evaluation procedure is to figure out which model makes the most accurate and valuable predictions. Some of the approaches utilized to evaluate the performance of the classification algorithms employed in this study are as follows:

1. **Accuracy:** Accuracy is a simple measure of total error and is expressed as the percentage of cases classified correctly. $accuracy = \frac{\sum True\ Positive + \sum TrueNegative}{SampleSize} x\ 100\% \rightarrow (vii)$
2. **Confusion Matrix:** The confusion matrix is a table that categorizes the number of correct and wrong guesses by response type. It's a table that describes how well a classification model performs on a set of test data for which the real values are known. True positives and negatives are included. Also, the ratio of true negative to false negative and true negative is used to determine selectivity or memory. This indicates how many relevant situations the classifier has identified and how many of them have been accurately guessed. This recall ratio is used to determine inclusiveness.
3. **Precision:** This is the percentage of true positives among the positive cases. As a result, any false positives are included. When the outcome is positive and the actual value is positive, it is a real positive. The genuine negative is when a condition (in this example, a sickness) exists but the classifier result is absent. The ratio of positives indicates a classifier's resolute performance based on the number of accurate guesses to the number of guesses it made. $precision = \frac{\sum TruePositive}{\sum TruePositive + \sum FalsePositive}$
4. **Sensitivity:** Sensitivity, also called as recalls, measures the model's ability to anticipate a good outcome by the percentage of 1s it properly recognizes. It is written as $\frac{|Trueclass\ 1|}{|class\ 1|}$
5. **Specificity:** The ability of a model to predict a negative event is measured by its specificity, which is expressed as: $specificity = \frac{\sum TrueNegative}{\sum TruePositive + \sum FalsePositive}$

DISCUSSION OF RESULTS

Three classifier models were used to detect cardiac problems using the dataset during the classification phase. Random forest ensemble, SVM classifier, and artificial neural network are the classifiers employed. Data was divided into two categories: training and testing, with training accounting for 80% of the data and testing accounting for 20%. The dataset was divided into five folds using cross-validation as the sampling method. The classifier was trained on four folds before being tested on the fifth. The performance metrics were determined after exploring all conceivable combinations of four folds. The model parameters for the classifiers are summarized in Table 3.

Table 3: Model Parameters for the Applied Classifiers

Random Forest	SVM	ANN
1000 tress, criterion = entropy	SVM type: C-SVM, C=1.0, Kernel: Linear, Probability = True, tolerance: 0.001	Hidden layers =2, hidden units per layer: 200 and 100 respectively, hidden layer activation function: Relu, output layer function: softmax, loss: spare categorical cross entropy, optimizer: stochastic gradient descent

Table 4: summarizes the results of the two classifiers, whereas figure 5 backs up these results by presenting the classification accuracy for each classifier on an unseen test dataset.

Table 4: A Summary of the Classifier Performance on Test Data

Classifier	Accuracy (%)	Precision (%)	Recall (%)	Sensitivity (%)
Random Tree Ensemble	50	48	65	65
SVM	70	47	52	52
ANN	50	50	50	50

The model performances were measured based on their accuracy, precision, recall and sensitivity. Radom forest model had a 50% accuracy, 48% precision score and 65% recall (sensitivity) score while SVM classifier had 70% accuracy score, 47% precision score and 52% sensitivity score. The ANN model had 50% score for accuracy, precision and recall. When the models were evaluated on the training data, all the three models obtained 100% in accuracy, precision and recall.

European Journal of Computer Science and Information Technology 11 (4), 1-21, 2023
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

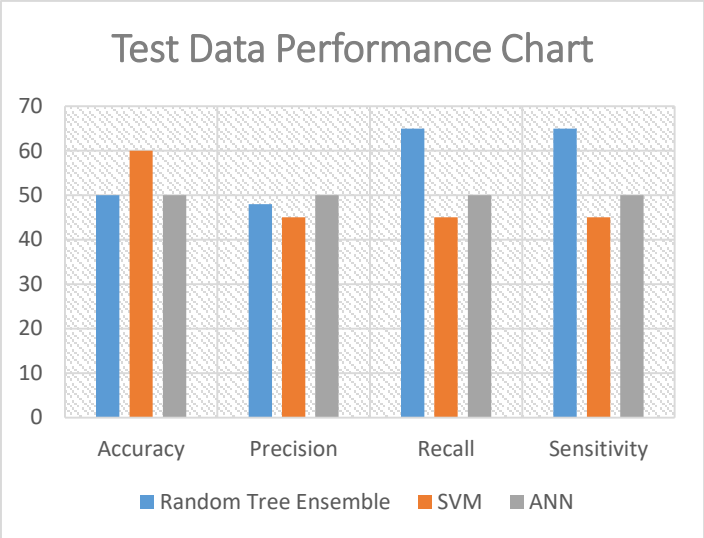

Fig 5: Test Data Performance Chart

For both the training and test datasets, confusion matrices for all models are shown. For the three classes of disease prognosis, the Random forest ensemble was able to forecast 100 percent correctly. When used on an unknown test dataset, it properly identified 75% of mild instances, 20% of intermediate cases, and 100% of severe cases. It misclassified one mild case and classified it as moderate, as well as four moderate cases as mild and severe.

When it came to the training data, the SVM classifier had a 97 percent overall accuracy. When the model was applied to the test data for prediction, it accurately identified 75% of mild cases, 80% of intermediate cases, and none of the severe cases. It misclassified one case as mild and classified it as moderate, one case as moderate and classified it as mild, and one case as severe and classified it as moderate.

The total accuracy of ANN on the training data was 100 percent. When the model was applied to the test data for prediction, it accurately identified 75% of mild cases, 20% of intermediate cases, and failed to categorize a severe case. It misclassified one light case as severe, three moderate cases as mild or severe, and a severe case as mild.

		Random Forest			
		PREDICTED			
		MILD	MODERATE	SEVERE	**Total**
Actual	MILD	**17**	**0**	**0**	17
	MODERATE	**0**	**11**	**0**	11
	SEVERE	**0**	**0**	**11**	11
	Total	17	11	11	39

		PREDICTED			
		MILD	MODERATE	SEVERE	**Total**
Actual	MILD	**3**	**1**	**0**	4
	MODERATE	**2**	**1**	**2**	5
	SEVERE	**0**	**0**	**1**	1
	Total	5	2	3	10

Fig. 6a: Random Forest on Training Data

			ANN			
			PREDICTED			
			MILD	MODERATE	SEVERE	**Total**
Actual		MILD	**17**	**0**	**0**	17
		MODERATE	**0**	**11**	**0**	11
		SEVERE	**0**	**0**	**11**	11
		Total	17	11	11	39

Fig. 6 b: Random Forest on Test Data

			PREDICTED			
			MILD	MODERATE	SEVERE	**Total**
Actual		MILD	**3**	**0**	**1**	4
		MODERATE	**2**	**2**	**1**	5
		SEVERE	**1**	**0**	**0**	1
		Total	6	2	2	10

Fig. 7a: SVM on Training Data

			SVM			
			PREDICTED			
			MILD	MODERATE	SEVERE	**Total**
Actual		MILD	**17**	**0**	**0**	17
		MODERATE	**0**	**11**	**0**	11
		SEVERE	**0**	**1**	**10**	11
		Total	17	12	10	39

Fig. 7b: SVM on Test Data

			PREDICTED			
			MILD	MODERATE	SEVERE	**Total**
Actual		MILD	**3**	**1**	**0**	4
		MODERATE	**1**	**4**	**0**	5
		SEVERE	**0**	**1**	**0**	1
		Total	4	6	0	10

Fig. 8a: ANN on Training Data

Fig. 8b: ANN on Test Data

DATA VISUALIZATION AND EXPLORATORY ANALYSIS

Visualization systems provide visual representations of datasets to assist individuals in doing activities more efficiently. When augmenting human skills rather than replacing humans with algorithmic decision-making processes, visualization is appropriate. When we don't know exactly what questions we need to ask ahead of time, data visualization helps us to examine data. The promise of better decision making through access to more data than ever before characterizes the modern era. We may employ purely computational approaches from domains like statistics and machine learning when we have well-defined questions to ask about data. Visualization is centered on using the human visual system as a form of communication, as the name indicates.

The following data visualization and exploratory analysis were carried on the research dataset
Figure 9: shows box plot of QRS and outcome

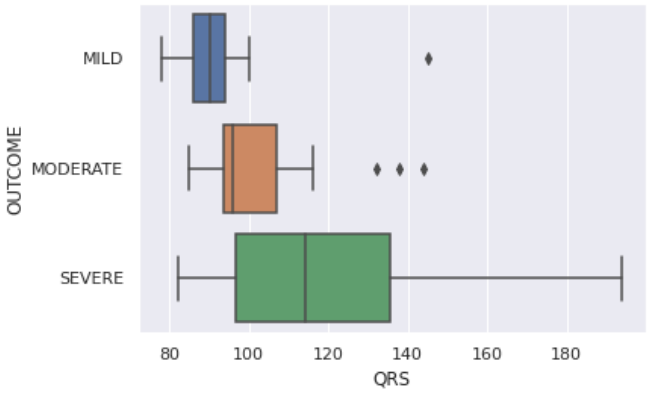

Fig 9: QRS vs Outcome

Figure 10 illustrates Bar plot of Heart Rate and Outcome

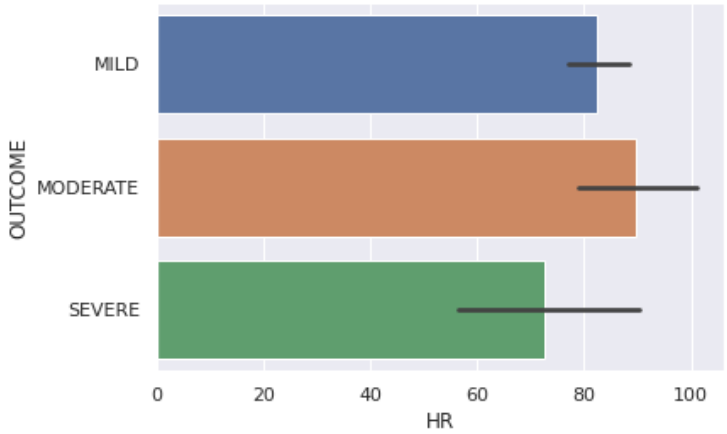

Fig 10: Heart rate vs Outcome

Figure 11: indicated Bar plot of Age and Outcome

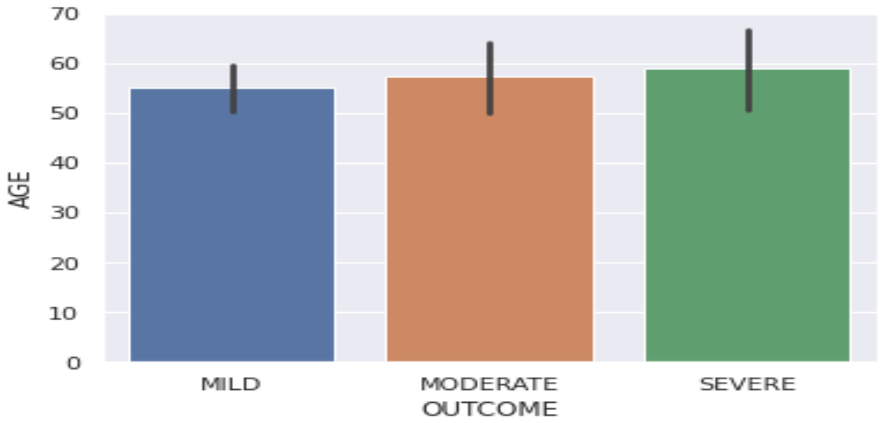

Fig 11: Age vs Outcome

Figure 12 illustrate Box plot of QTC and outcome

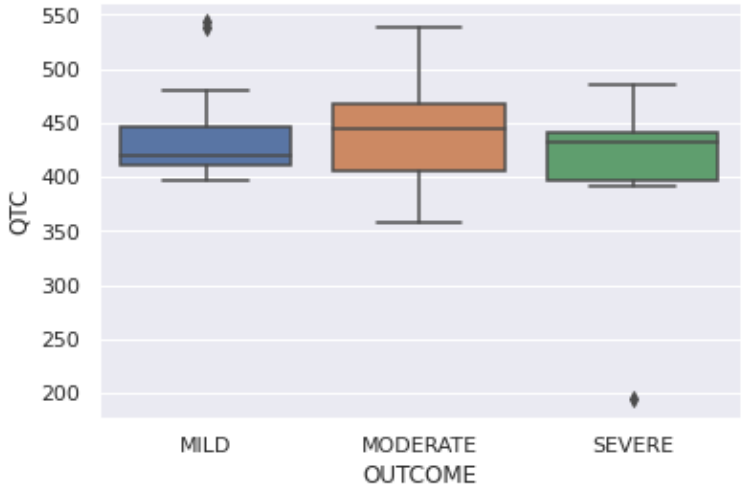

Fig 12: QTC vs Outcome

Figure 13 shows Bar plot of pulse rate vs outcome

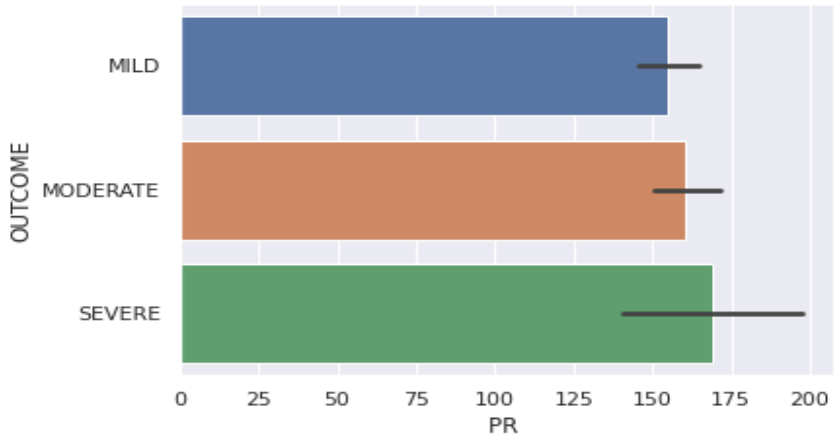

Fig 13: Pulse rate vs outcome

Figure 14 presents Bar plot of RV5+SV1 and Outcome

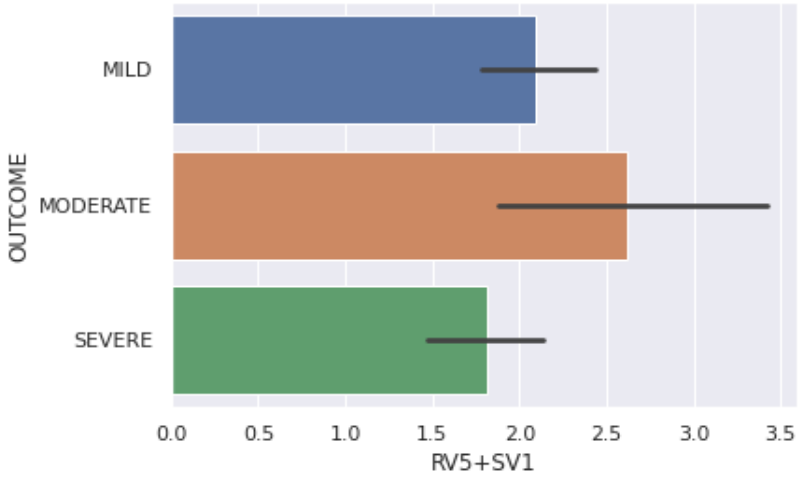

Fig 14: Outcome vs RV5+SV1

CONCLUSION

Using patient ECG and demographic information as a collection of data, we developed a new index for modeling early diagnosis of cardiovascular illnesses based on data and measurements from electrocardiogram (ECG) system. The clinical dataset was subjected model training and tested using three algorithms: random forest (RF), support vector machine (SVM), and artificial neural network (ANN). The random forest approach employed a 1000-tree estimator to categorize instances into mild, moderate, and severe categories. The SVM was trained using three different methods. Linear kernel, polynomial kernel, and radial kernel methods. The grid search scikit-learn python function was deployed to archive the implementation of various algorithms. The grid search cross validation approach iterated over all of the SVM kernels until the most accurate estimator was found. Given the patient's ECG data, the linear kernel was chosen as the estimator with the greatest accuracy score in predicting the diagnostic outcome. The sequential technique was adopted to train the ANN algorithm. All three approaches functioned admirably, although they vary in terms of subsets of training, validation, and testing data. The RF approach was employed for the classification, Using the test dataset give a classification accuracy of 50%. Again, the classifier produced varying results when categorizing data into various classes. This indicates that if an ensemble random forest had been trained on a bigger dataset, it would have produced better results than what was obtained. The demographic and ECG medical data of patients were supplied as input along with the diagnosis data for prediction of early syndromic of cardiovascular disease as mild, moderate, and severe. Furthermore, the test dataset with SVM and ANN achieved an accuracy of 70% and 50% respectively on the prediction probability combination. Finally, visualization techniques deployed to boost human capacities, improve computational decision-making, and describe the functional relationship between feature variables and medical diagnosis outcome flexible. The work established that, the implementation of the AI–ECG as a tool in the clinical domain holds the promise to transform clinical care and offer efficient medical service provisioning.

ACKNOWLEDGMENTS
The authors would like to thank Department of Computer Science, Faculty of Physical Sciences, Akwa Ibom State University, and University of Uyo Teaching Hospital (UUTH), Nigeria for the support during the research work.

Author Contributions: Conceptualization, I. Umoren; O. Abe, methodology, I. Umoren; O. Abe, S. Inyang; software, S. Inyang; validation, I. Umoren, O. Abe; formal analysis, I. Umoren, G. Ansa, S. Inyang; investigation, O. Abe, S. Inyang, I. Umoh, software; resources, S. Inyang, O. Abe; data curation, I. Umoh, S. Inyang, O. Abe; writing original draft preparation, I. Umoren, O. Abe; writing review and editing, I. Umoren; visualization, I. Umoren, O. Abe, Supervision, I. Umoren; project administration, I. Umoren; G. Ansa, funding acquisition,
All authors have read and agreed to the published version of the manuscript.

Funding: This paper did not receive funding.
Institutional Review Board Statement: Not applicable.
Informed Consent Statement: Not applicable.
Data Availability Statement: Not applicable.
Acknowledgments: The authors would like to thank Department of Computer Science, Faculty of Physical Sciences, Akwa Ibom State University, MkpatEnin, Nigeria for the support during this research work.
Conflicts of Interest: The authors declare no conflict of interest.

DATA AVAILABILITY: Data supporting these findings are available within the article or upon request.

REFERENCES

[1] Enrica, A., Verdecchia, P., Narducci, P., & Angeli, F. Additive value of standard ECG for the risk prediction of hypertensive disorders during pregnancy. Hypertension Research, 2011. 34(6), 707–713. https://doi.org/10.1038/hr.2011.12

[2] Eduardo José da S. Luz, William Robson Schwartz, Guillermo Cámara-Chávez, David Menotti,
ECG-based heartbeat classification for arrhythmia detection: A survey, Computer Methods and Programs in Biomedicine, 2016; 127, Pages 144-164, ISSN 0169-2607, https://doi.org/10.1016/j.cmpb.2015.12.008.

[3] Estrada, G., Luis, M., Mendoza, l.e., Sc, M., Molina, V., Relationship of blood pressure with the electrical signal of the heart using signal processing. 2014

[4] Essien V., Umoren I., Umoh I. A Bio-Informatics System for Intelligent Classification of Severity Index of Hypertension, International Journal of Innovative Research in Sciences and Engineering Studies (IJIRSES), 2021; ISSN: 2583-1658 | Volume: 1 Issue: 2, pp. 1-8. www.ijirses.com

[5] Geron A.. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd. ed.). O'Reilly Media, Inc. 2019.

[6] Henry, N., Ginsberg, and Paul, R., MacCallum, The Obesity, Metabolic Syndrome, and Type 2 Diabetes Mellitus Pandemic: Part I. Increased Cardiovascular Disease Risk and the Importance of Atherogenic Dyslipidemia in Persons with the Metabolic Syndrome and Type 2 Diabetes Mellitus J Cardiometab Syndr. Spring; 2009; 4(2): 113–119. doi: 10.1111/j.1559-4572.2008.00044.x

[7] Hiroyuki, T., Tomonori, S., Shunsuke, M., Sumiyo , Y., Nobuyuki, O. and Yasuaki, D., Use of Electrocardiography to Predict Future Development of Hypertension in the General Population. 2016; doi: 10.1097/MD.0000000000003483 PMCID: PMC4998710

[8] Justice, P., F, Gaage., Igulu, K., A model for electrogram classification using convolutional neural network. DOI 10.17148/IJ ARCCE.2021.10405

[9] Melillo P, Izzo R, Orrico A, Scala P, Attanasio M, Mirra M, et al. Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis. PLoS ONE, 2015; 10(3): e0118504. https://doi.org/10.1371/journal.pone.0118504

[10] Monika, Martin, Ana, Bojana et al., ECG-derived Blood Pressure Classification using Complexity Analysis-based Machine Learning 11th International Conference on Health Informatics, 2018; DOI:10.5220

[11] Moradi, M., & Zulkernine, M. November). A neural network based system for intrusion detection and classification of attacks. In Proceedings of the IEEE international conference on advances in intelligent systems-theory and applications, 2004; (pp. 15-18). IEEE Lux-embourg-Kirchberg, Luxembourg.

[12] Peter, O.M., Devereux, R. B., Nieminen, M. S., Jern, S., Oikarinen, L., Viitasalo, M., Toivonen, L., Kjeldsen, S. E., Julius, S., Snapinn, S., & DahlöF, B. Strain Pattern and Prediction of Cardiovascular Morbidity and Mortality in Hypertensive Patients. Hypertension, Electrocardiographic 2004; 44(1),48–54.

[13] Samir Shantha, G.P., Patel, N., Chandran, S., Nanavaty, S., Boruah, P., Sheth, J., & Pancholy, S. (2014). Electrocardiogram Based Scoring System for Predicting Secondary Pulmonary Hypertension: A Cross-Sectional Study. Journal of The American College of Cardiology, 63(12), A1497.

[14] Setareh, A., Williams, Michelson, L., Valerie, A., An Evaluation of the Effects of an Angiotensin Receptor Blocker on Health-Related Quality of Life in Patients with High-Normal Blood Pressure(Prehypertension) in the Trial of Preventing Hypertension (TROPHY). (2008). doi: 10.1111/j.17517176.2008.07837

[15] Sharma, M., Tan, R.S., Acharya, U.R. A novel automated diagnostic system for sclassification of myocardial infarction ECG signals using an optimal biorthogonal filter bank. Computer. Biol. 2018.

[16] Simjanoska, M., Gjoreski, M., Madevska Bogdanova, A., Koteska, B., Gams, M., Tasic, J., 2018; ECG-Derived Blood Pressure Classification Using Complexity Analysis-Based Machine Learning. In Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF, Madeira, Portugal, 282–292.

[17] Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. Nat Rev Cardiol. 2021 Jul;18(7):465-478. doi: 10.1038/s41569-020-00503-2. Epub, PMID: 33526938; PMCID: PMC7848866.

[18] Umoren I, Usua G. and Osang F. Analytic Medical Process for Ophthalmic Pathologies Using Fuzzy C-Mean Algorithm, Innov. Syst. Softw. Eng, 2019.

[19] Umoren, I., Udonyah K., and Isong E. A Computational Intelligence Framework for Length of Stay
Prediction in Emergency Healthcare Services Department, 2019; 14th International Conference on Computer Science & Education (ICCSE), pp. 2019; 539-549, doi: 10.1109/ICCSE.2019.8845332.

[20] Umoren I. Etuk U. Ekong A. and Udonyah K. Healthcare Logistics Optimization Framework for Efficient Supply Chain Management in Niger Delta Region of Nigeria" International Journal of Advanced Computer Science and Applications (IJACSA), 2021; 12(4)
http://dx.doi.org/10.14569/IJACSA.2021.0120475

[21] Wong, M. Y.-M., Poon, C. C.-Y., and Zhang, Y.-T. An Evaluation of the Cuffless Blood Pressure Estimation Based On Pulse Transit Time Technique: a half year study on normotensive subjects. Cardiovascular Engineering, 2009; 9(1):32–38.