Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease Classification

Talha Ahmed Khan[1]
British Malaysian Institute, Universiti Kuala Lumpur Kuala Lumpur, Malaysia, Usman Institute of Technology (NED), Karachi, Pakistan

Kushsairy A. Kadir[2]
British Malaysian Institute, Universiti Kuala Lumpur Kuala Lumpur, Malaysia

Shahzad Nasim[3]
Faculty of Engineering Sciences, Technology and Management, Ziauddin University Karachi, Pakistan

Muhammad Alam[4]
CCSIS, Institute of Business Management (IoBM), Karachi, Pakistan, MIIT-Universiti Kuala Lumpur, Kuala Lumpur, Malaysia

Zeeshan Shahid[5]
Electrical Engineering, Institute of Business Management (IoBM), Karachi, Pakistan

M.S Mazliham[6]*
Malaysian France Institute, Universiti Kuala Lumpur Kuala Lumpur, Malaysia

***Abstract*—Breast cancer and heart disease can be acknowledged as very dangerous and common disease in many countries including Pakistan. In this paper classifiers comparative study has been performed for the tumor and heart disease classification. Around one lac women are diagnosed annually with this life-threatening disease having no family history of the disease. If it is not treated on time it may grow and spread to the other parts of human body. Mammograms are the X-rays of the breast which can be used for the screening of cancer tumor. Prior identification of breast cancer may increase the chance of survival up to 70 percent. Tumors which causes cancer can be categorized into two types: a) Benign and b) Malignant. Benign tumor can be explained as the tumor which are not attached to neighbor tissues or spread in the other parts of the body. In Malignant tumor, other parts may be affected by it as it can grow and spread in the other parts of the body. To classify the tumor as Malignant or Benign is very complex as the similarities of cancer tumor and tumor caused by the skin inflammation are almost same. The early identification of Malignant is mandatory to protect the patient life. Diversified medical methods based on deep learning and machine learning have been developed to treat the patients as cancer is a very serious and crucial issue in this era. In this research paper machine learning algorithms like logistic regression, K-NN and tree have been applied to the breast cancer data set which has been taken from UCI Machine learning repository. Comparative study of classifiers has been performed to determine the better classifier for the robust prediction of breast tumors. Simulated results proved that using Logistic regression, ninety-one percent accuracy was achieved. The research showed that logistic regression can be applied for the accurate and precise early prediction of breast cancer. Cardiovascular disease is very common throughout the world. It has been noticed that health in cardiac patients that there are so many factors which causes heart disease or heart attack. The factors leading to the heart failure includes varying blood pressure, high sugar, cardiac pain, and heart rate, high cholesterol level (LDL), artery blockage and irregular ECG signals. Many researchers proved that stress in patients can also be the reason for the heart disease. Higher numbers of cardiac surgeries like angioplasty and heart by-pass are performed on annual basis. Actually, people don't care about their lifestyle and diet and fully ignore the symbols. It can be early predicted and cured if proper testing and medication for heart is done. Sometimes there is a false pain which has the same feeling like angina pain depicting cardiovascular disease. To reduce the false alarm and robustly classify the heart disease, several machine learning approaches have been adopted. In proposed research for the accurate classification of heart disease comparison has been performed among support vector machine (SVM), K-nearest neighbors K-NN and linear discriminant analysis. Simulated results demonstrated that Support vector machine was found to be a better classifier having an accuracy of 80.4%.**

Keywords*—Breast cancer; benign; malignant; logistic regression; cardiovascular disease; heart disease diagnosis; support vector machine; classifiers; k-nearest neighbors

I. Introduction

Twenty-five percent of women die due to the breast cancer in the ages of thirty-five to forty. Mammography is usually performed to enhance the radiographic decisions. SENOLOG was developed for the breast therapy assistance using SENOBASE [1]. RF-ELM classifier was applied to find out the tumor from the digital mammogram. Mammogram images were taken from MIAS database. Kurtosis, mean, standard deviation, correlation coefficient, entropy and variance were chosen for the accurate classification. RF-ELM was found to be very competent classifier for the diagnosis of breast cancer [2]. This research paper is divided into four main sections. Section one explains the introduction. Second part discusses the implementation of classifier algorithms for breast tumor identification. Section three elaborated the heart disease

*Corresponding Author

classification and its implementation. Results and conclusion have been discussed in section four.

A. Existing Methods for the Identification of Malignant Tumor

Local binary patterns (LBP) were applied using mammograms. Data set was collected from DDSM. LBP using mammograms achieved the accuracy of 84% [3]. Mammography is acknowledged as a good strategy for the screening of tumors. Generally, mammogram analysis is very complex as the image comprises of various little differences of different tissues [4]. A novel hybrid approach of digital image processing was adopted to analyze the mammograms. Using this novel approach, the early identification of breast cancer at the stage of micro calcification was achieved leading to the higher accuracy of proposed technique [5]. Digital image based elasto-tomography was developed as a prototype for the evaluation of breast cancer. Segmentation was performed to identify the model of breast. Using this system up to 10 mm tumor could be detected in a silicon phantom breast [6]. Ensemble empirical mode decomposition (EEMD) has been proposed for the prior determination of breast tumors using ultra-wide band (UWB) microwave imaging. Approximately 4 mm tumor has been identified in inside the glandular tissue whose di-electric constant was 35 in a breast model [7]. The pulsed confocal approach was proposed to improve the identification of breast tumors. Two-dimensional finite difference time domains (FDTD) analysis was conducted to determine the 2mm tumor in the presence of clutter [8]. It has been observed that rumors possess different permittivity and conductivity with respect to the surrounding tissues. Electrical impedance spectroscopy (EIS) was used to classify the normal tissues and malignant [9].

Fig. 1(a) and 1(b) elaborated that homogenous breast model was designed. Incident wave of 6 Ghz having vertical polarization was tested. Artificial Neural network was designed to evaluate the scattered electromagnetic waves. The dielectric values for malignant and normal tissue was randomly chosen [11-12].

B. Problem Statement

Early detection of Breast cancer has become very crucial issue in the medical science as 30% women die annually due to the breast cancer. Women usually ignore the tumor because of the lack of awareness for the breast cancer. To classify the tumor as Malignant or Benign is very complex as there is a misconception and confusion regarding these two classes [1-4]. The last stage symptoms are also similar with the normal inflammatory conditions. Therefore, vigorous early breast cancer detection was needed. Mammography based screening is usually used to evaluate the cancer tumor on early basis as well when it is small. Many clinical laboratories are there to record the mammograms of breast which is the X-ray of breast. Data acquisition for the breast tumor can also be possible as the size, shape adhesion, location and other attributes related to cancer tumor can also be recorded [2-6]. To make it more certain and enhance the accuracy of Malignant tumor identification Machine Learning based decision making was required. The similarities of the tumor's symptoms are almost same as the inflammation of the skin problem. Breast pain, swelling, and reddish skin are very common symptoms of cancer but people ignore it as they take it as normal skin inflammatory problem [9-11].

C. Methodology

Fig. 2 elaborates the main fundamental block diagram for the proposed breast cancer classification using machine learning. Clinical data acquisition was performed and collected from the UCI machine learning repository for applying the proposed classification models. Logistic regression, K-NN and decision tree classifiers were applied to determine the best predictive model for the breast cancer. Cross validation curve was also obtained for the comparison. Results were obtained in terms of accuracy, precision, prediction speed, ROC, true positive rate and false positive rate.

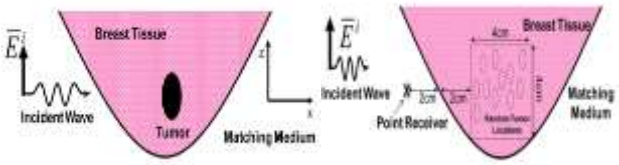

Fig. 1. (a); and (b); Breast Tumors at Random Locations [10].

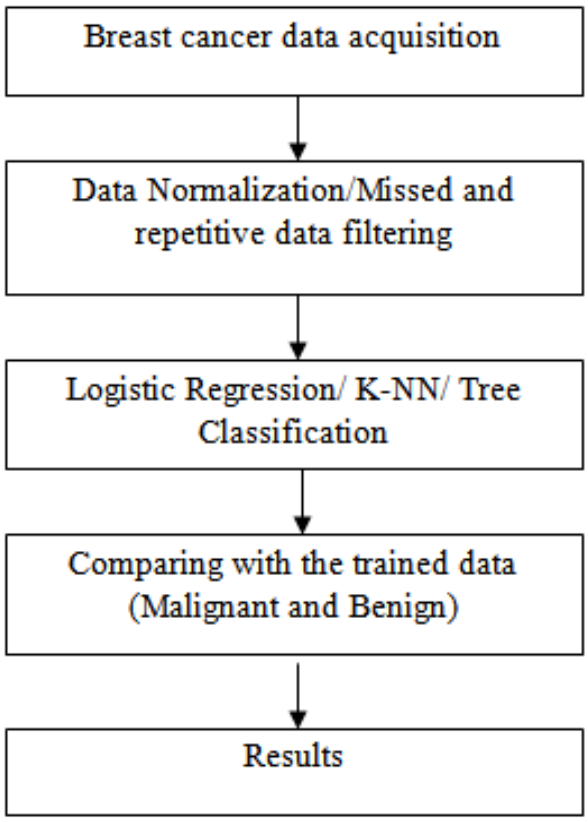

Fig. 2. Main Block Diagram.

D. Data Acquisition for Breast Cancer

Table I represents that Breast cancer dataset has been gathered from the database of UCI machine learning repository for the Benign and Malignant classification purpose. The data set has been used in many researches by using neural networks and machine learning based classification [13-14]. The data set comprised of eleven attributes including the patient ID. Column 1 describes the patient ID for each individual patient. Clump thickness has been mentioned in column 2. Clump is the bunch of close roots which have been grown with the tumor tissue. Uniformity of cell size has been represented in the column no. 3. Cell shape of tumor has been described in column no. 4. Column 5 shows the marginal adhesion for the tumor. Column no. 6 represents the single epithelial cell size. Epithelial cell is defined as the cell which protects the upper surface of the skin against germs and bacteria. Bare Nuclei has been demonstrated in the column no. 7. Bare nuclei are the cytology preparation which can be observed in the degeneration of cell. Bland chromatin has been defined in the column no. 8. Bland chromatin explains the pattern and texture of the Benign tumor. Usually in cancer cell the texture is found to be rough and harsh. Column no. 9 displays the normal nucleoli. Nucleoli depict the cell's response to the stress. Column no 10 displays the Mitosis attribute of breast tumor. Mitosis can be defined as the two daughter cells having same properties and number of chromosomes in parent cell like an ordinary tissue. Output results have been mentioned in the last column. Column 11 shows the output classes 1 and 2, 1 for Benign and 2 for Malignant [13-14].

TABLE I. BREAST CANCER DATA ACQUISITION [13-14]

Patient ID	CL Thick	Cell size U	Cell shape	Adhesion	Cell size	Bare Nuclei	Bland Chromatin	Nucleoli	Mitosis	Class
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1035283	4	2	1	1	2	1	2	1	1	2
1036172	1	1	1	1	1	1	3	1	1	2
1041801	2	1	1	1	2	1	2	1	1	2
1043999	1	1	1	1	2	3	3	1	1	2
1044572	8	7	5	10	7	9	5	5	4	4
1047630	7	4	6	4	6	1	4	3	1	4

II. IMPLEMENTATION OF CLASSIFIER ALGORITHMS FOR BREAST TUMOR IDENTIFICATION

A. Decision Tree Implementation

Decision tree is same as the tree in which the outputs are represented by leaves. Decision tree algorithm is able to classify and sort from roots to the leaf. For the information gain the entropy is estimated in information coding theory.

$$E = \sum \boldsymbol{i} = \mathbf{1} \text{ - } P \times \log_2(p_i) \quad (1)$$

Probabilistic classification can be positive or negative. Entropy for" t" training sample can be explained as:

$$(t) = -p+\log_2 p + -p-\log_2 p- \quad (2)$$

t = training of sample data.

$-p+\log_2 p$ = negative examples defined in data.

$+ -p-\log_2 p-$ = positive examples in data.

Fig. 3 graphically illustrated the ROC curve for the decision tree algorithm. ROC curve was plotted between true positive rate and false positive rate to assess the performance of the classifier. The area under the curve (AUC) defines the value up to which the classification model can classify. The values 0.40 and 0.92 were found to be false positive rate and true positive rate respectively in the graph. AUC was determined 0.70 that can be considered as a good classification model for the breast cancer prediction.

Fig. 4 represented decision tree confusion matrix. Usually confusion matrix is observed diagonally; all the values in diagonal show the true positive classes. Confusion matrix shows the performance evaluation of decision tree classifier. In Class "2- Benign" classification, true positive rate was found to be 92% and false positive rate was determined as 8%. For the class "4-Malignant" classification, 60% true positive rate was observed with 40% false positive rate.

B. K-Nearest Neighbors K-NN Classifier

Fig. 5 portrays the confusion matrix of K-NN algorithm. 88.2% accuracy was achieved by the K-NN for the classification of Benign and Malignant Tumor. The efficiency was estimated using true positive rate and false positive rate

values. Confusion matrix assessed the K-NN classifier and displayed that 92% true positive rate was achieved with 8% false positive rate for the classification of "class 1-Benign". In "Class-Malignant" classification, 80% true positive rate was achieved with 20% false positive rate.

Euclidean distance is estimated to determine the closest distance with the value of the K.

$$d = \sqrt{(x1 - xA1)^2 + (x2 - xA2)^2} \quad (3)$$

Fig. 6 shows ROC and area under the curve (AUC) for the K-NN classifier. The ROC curve has been plotted between true positive rate and false positive rate. Area under the curve (AUC) was found to be 0.86. It is slightly away from the 1. For good classification AUC must be close to 1.

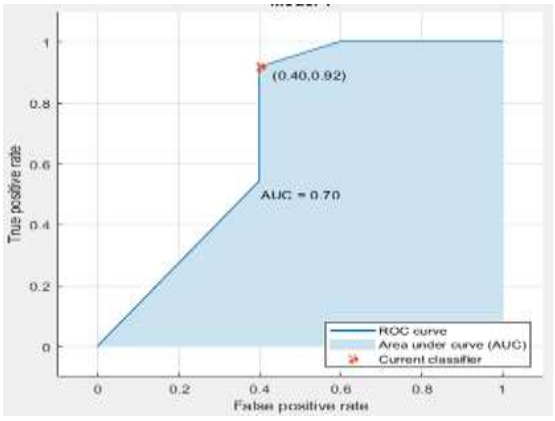

Fig. 3. Decision Tree ROC Curve.

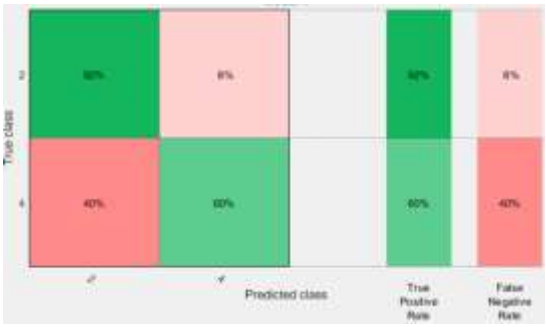

Fig. 4. Decision Tree Confusion Matrix.

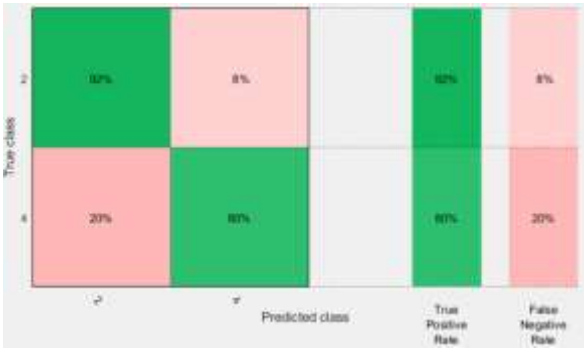

Fig. 5. K-NN Classifier Confusion Matrix.

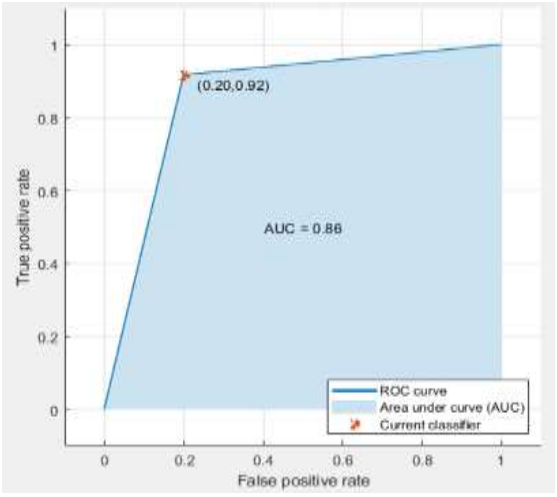

Fig. 6. K-NN ROC and AUC.

C. Logistic Regression

Logistic regression is defined as the regression model in which predictive model predict the output in binary. Outliers and missing values must be filtered out before processing the predictive regression analysis. Logistic function may be defined as:

$$\text{Logistic Regression}(x) = \frac{1}{1+\exp(-x)} \quad (4)$$

Fig. 7 elaborates the regression model for the logistic regression. It can be seen that all the attributes or the parameters related to the classification of classes have been entered to the logistic regression model for the prediction of tumor analysis. The weighted sum is converted into probability by logistic function.

$$o = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)}) \quad (5)$$

Where o is the predicted output, b0 is the bias and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

In Fig. 8, Logistic Regression confusion matrix portrayed that mean accuracy of 91.2% was achieved based on the true positive rate (TPR) and false positive rate (FPR). For the classification of Benign, confusion matrix showed that 100% true positive rate was achieved in classifying class 2 while 0% of false positive rate was achieved in the classification of class 1. It was also observed that 70% true positive rate was estimated in the classification of "class 2-Malignant" with the 30% false positive rate.

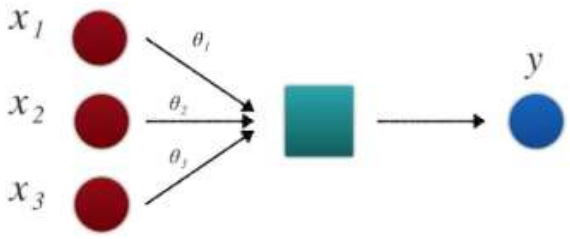

Fig. 7. Logistic Regression Sample Model.

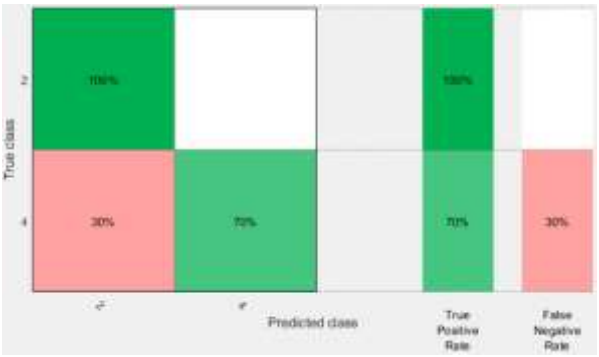

Fig. 8. Logistic Regression Confusion Matrix.

Fig. 9 illustrated graphically that area under the curve (AUC) for the logistic regression was found to be 0.89.

D. Fine Gaussian

In Fig. 10, Confusion matrix of fine Gaussian SVM elaborated that the algorithm performed very poor for the prediction of breast cancer as it classified all classes as class 1. Class 2 was not predicted at all therefore false negative rate was found to be 100% and false positive rate for class 2 was found to be 0%.

E. Comparative Study of Classifiers for the Prognosis of Breast Tumor

Fig. 11(a) shows the accuracy, prediction speed and training time for the Decision tree. Fig 11(b) explains the parametric analysis for the logistic regression. It can be observed from the parametric analysis that 91.2% accuracy has been achieved by logistic regression. Moreover, it can also be noticed that 88.2% accuracy was achieved in the trained predictive model for the breast tumor classification.

Table II showed that the Logistic Regression and K-NN performed better classification compared to the Decision tree and Fine Gaussian SVM in terms of Accuracy, prediction speed, training elapsed time, precision and area of under the curve. 91.2% accuracy was achieved by logistic regression for the benign and malignant classification.

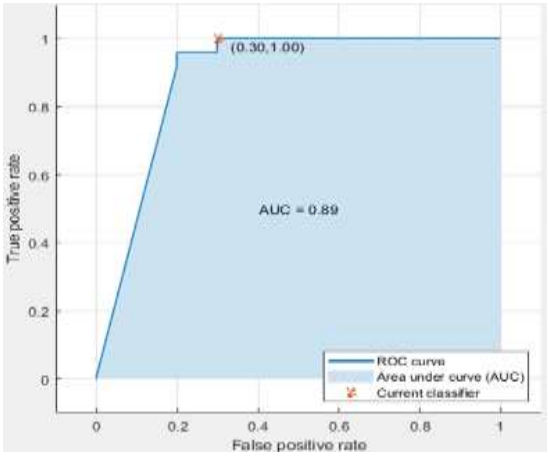

Fig. 9. ROC and AUC Curve for Logistic Regression.

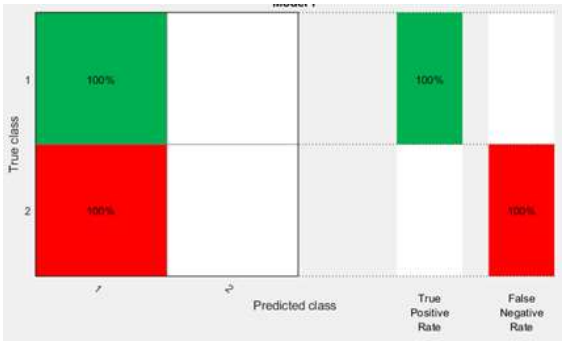

Fig. 10. Confusion Matrix of fine Gaussian SVM.

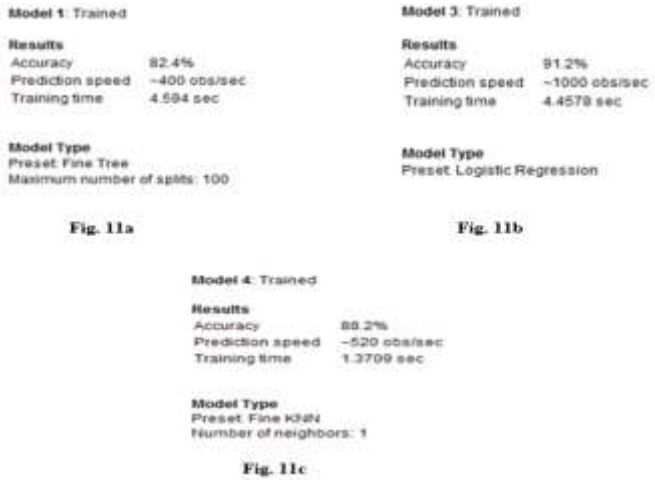

Fig. 11a Fig. 11b

Fig. 11c

Fig. 11. (a), (b), and (c) Displays the Trained Model for the Predictive Analysis of Breast Cancer.

TABLE II. TREE, K-NN AND LOGISTIC REGRESSION COMPARISON

Parameters	Decision Tree	K-NN	Logistic Regression	Fine Gaussian SVM
Accuracy	82.4%	88.2%	91.2%	59.8%
Prediction Speed	400 obs/sec	520 obs/sec	1000 Obs/sec	1700 obs/sec
Training time	4.594 seconds	1.3709 seconds	4.4578 seconds	0.8763 sec
Precision	0.82	0.89	0.90	0.59
AUC	0.70	0.86	0.89	0.60
Precision	0.83	0.89	0.93	0.55
F1-score	0.84	0.88	0.91	0.45
Recall	0.85	0.87	0.90	0.50

III. RECENT TRENDS FOR HEART DISEASE CLASSIFICATION

Cardiovascular disease is very common throughout the world even in United states of America in every thirty-four second one patient losses his or her life due to this silent disease [15]. Electrocardiography signals (P-wave, QRS complexes) and cardiac arrhythmias have also been processed and classified through convolutional neural network (CNN) to identify the heart disease [16]. Smoking and hypertension may also increase the chances of heart disease. Data mining

techniques were applied to the heart disease data (HDD) for the classification of heart disease [17]. Quadratic support vector machine and discriminant analysis have been performed in the MATLAB environment for the classification of heart disease [18]. Fuzzy based K-NN classifier was developed for the classification of pure cardiovascular disease. Training, testing and validation curve demonstrated that fuzzy based K-NN classifier worked better [19]. A wearable gadget was also fabricated for the real time data transmission. The health parameters including blood pressure, temperature and heart rate were optimized using particle swarm optimization for the optimal results in previous research [20].

Fig. 12 illustrated that cardiac arrhythmias and ECG signals classification were performed using, Fuzzy logic controller, MLP-PSO, Improved PSO(ImpSO) and Genetic algorithm [21-22]. The data was collected from PHYSIONET. Heart rate variability (HRV) has been used as a yardstick to measure the heart health. Heart rate variability signals have been processed for the classification of heart disease using artificial neural network (ANN) [23-26]. Cardiac arrhythmia can be categorized into following categories, Asystole, Bradycardia, Tachycardia, Ventricular Tachycardia and Ventricular flutter. Fuzzy logic was used to classify the heart disease as cardiac arrhythmia can be used for measuring the heart health [27]. Rs and QRS complexes were cleaned and classified using Pan and Tompkins algorithm [28]. Cardiac abnormalities have been observed for the heart rate classification [29]. ECG signals have been classified using fuzzy network for the heart disease classification [30]. Cardiac arrhythmias based on rate time series were forecasted using radial basis function [31]. 116 heart sounds were segmented using classification and regression trees [32]. An optimal architecture of multi-layer perceptron with the combination of particle swarm optimization has been designed for the prediction of cardiac arrhythmias [33]. A stand-alone system using DSK6713 was developed to measure the abnormalities in heart sound [34].

A. Problem Statement for Heart Disease Classification

Cardiovascular disease is very common throughout the world. Usually people ignore the early symptoms of the heart disease like people think the actual cardiac chest pain as a typical angina or non-cardiac chest pain. Ignorance of blood pressure, sugar and cholesterol serum may lead towards the heart disease. Sometimes gastric pain and non-cardiac chest pain occurred as a false alarm for the heart disease identification. The accurate and proper prediction of heart disease may be performed using machine learning based on the patient historical data set with respect to the age.

B. Implementation of Classifiers for Heart Disease Classification

Table III explained that dataset has been collected from the database of UCI machine learning repository for the heart disease classification purpose. The data set has been used in many researches by using neural networks and ensemble classification [35]. The data set contained thirteen attributes to predict the heart disease. Output results were represented as two classes 1 or 2. Class "1" confirms the absence of heart disease and heart is working fine while class "2" indicates the high risk of heart attack as the heart disease has been found and it needs urgent consultation or treatment with the doctor or heart specialist. The Age and sex of patients have been mentioned in the column 1 and 2, respectively. Column no. 3 defines the type of the chest pain (CP) as the chest pain has four types. The chest pain types include typical angina, atypical angina, non-angina pain and asymptomatic. A heart pain or chest trouble caused by the muscles of heart due to the less oxygen in the blood is usually referred as typical angina pain. Atypical angina pain is a symbol of the problem that is not related to the heart actually. Non-angina pain is also acknowledged as non-cardiac chest pain (NCCP) that has the same feel like heart pain but that doesn't describe the heart disease. Asymptomatic shows that there was no heart disease detected. Column no. 4 shows the blood pressure of the patients in the rest condition. The attribute no. 5 describes the cholesterol serum level value in mg/dl. The sixth column explains that blood sugar (FBS) was measured in fasting of patients to identify the sugar greater than 120 mg/dl. The electrocardiographic results of patients in rest have been observed in the column no. 7 having the values of 0, 1 and 2. The maximum peak heart rate of patient was measured in the column no. 8. The column no. 9 shows the data of those patients who got induced angina pain due to the exercise. In column no. 10 old peak which is related to the ST depression achieved by exercise at rest position. Slope of the peak exercise has been mentioned in the column no. 11 (up, flat, down). Number of the main vessels (0-3) has been recorded in the column no. 12 that has been colored by the fluoroscopy. Thallium is the stress scintigraphy which elaborates that the heart rate is normal, fixed defect or reversible defect. The predicted results have been described in the 14th column having two classes. Class "1" shows that there was no heart disease identified and the heart is working properly. Class "2" indicates that the presence of heart disease has been confirmed therefore emergency consultation or treatment will be needed to cure it.

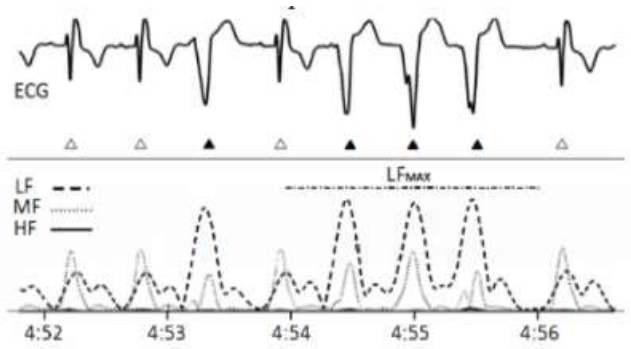

Fig. 12. Electrocardiogram Signals [21].

TABLE III. DATA SET FOR HEART DISEASE CLASSIFICATION [19]

Age	Sex	CP	RBP	SCHL	FBS	RECR	MHR	EIA	OP	SP	MV	TH	Output
67	0	3	15	564	0	2	60	0	1	62	0	7	1
57	1	3	24	261	0	0	41	0	0	31	0	7	2
67	1	2	28	263	0	0	5	1	0	22	1	7	1
74	0	4	20	269	0	2	21	1	0	21	1	3	1
65	1	2	20	77	0	0	40	0	0	41	0	7	1
56	1	4	30	256	0	2	42	1	0	62	1	6	2
59	1	3	10	239	0	2	42	1	1	22	1	7	2
60	1	4	40	293	0	2	70	0	1	22	2	7	2
63	0	4	50	407	0	2	54	0	4	42	3	7	2
59	1	4	35	234	0	0	61	0	0	52	0	7	1
53	1	4	42	226	0	2	11	1	0	22	0	7	1
44	1	4	40	235	0	2	80	0	0	40	0	3	1
61	1	3	34	234	0	0	45	0	2	62	2	3	2
57	0	1	28	303	0	2	59	0	0	59	1	3	1
71	0	4	12	49	0	0	25	0	1	62	0	3	1
46	1	4	40	311	0	0	20	1	1	82	2	7	2
53	1	4	40	203	0	2	55	1	3	13	0	7	2
64	1	4	10	211	0	2	44	1	1	82	0	3	1
40	1	4	40	99	0	0	78	1	1	41	0	7	1

C. Support Vector Machine (SVM)

Generally, Support Vector Machine (SVM) classifiers are applied to resolve complicated engineering problems of the real world; it has been observed in many classification applications that SVM performed better classification. A support vector mechanism produces a hyperplane or a series of hyperplanes in a high or infinite dimension area that can be used for classification, regression or detection of outliers. In the proposed research two hyper planes have been designed for the two classes.

H1 and H2 are the planes:

$$H1: w \bullet x_i + b = 1 \quad (6)$$

$$H2: w \bullet x_i + b = 2 \quad (7)$$

The plane H0 is the median in between, where $w \bullet xi + b = 0$

$$w^T x + b \geq 0 \text{ for } di = 1 \quad (8)$$

$$w^T x + b \geq 1 \text{ for } di = 2 \quad (9)$$

For the maximization of the margin, ||w|| can be minimized. Having the condition that there will be no data points between H1 and H2.

Non-Linear SVMs also used to separate the classes linearly by using the quadratic equation.

$$(x-a)(x-b) = x^2 - (a+b)x + ab \quad (10)$$

Optimization issue of the weight values can be resolved by using the following equations foe SVMs:

For the maximization;

$$\frac{1}{\|w\|} \quad (11)$$

Min. $| w^T x + b |= 0$ for n = 1,2,3……n

For the minimization;

$$\frac{1}{2} W^T.W \quad (12)$$

$y_n = | w^T x + b |= 0$ for n = 1,2,3……n

The preprocessed data was made ready for analysis. The data was filtered and the missing values were recovered. The data was given in the form of numbers and fractions and can be used for training.

Fig. 13 represents the proposed model data points of parameters. Proposed predictive model was trained with different algorithms and cores to verify performance. The method of teaching the algorithm also makes a big difference. If the wrong training data type is provided, the algorithm cannot achieve useful results.

Fig. 14 shows the Support Vector Machine confusion matrix. Commonly confusion matrix is seen diagonally; all the values in diagonal show the true positive classes. The confusion matrix shows the performance of the classifier. The confusion matrix of SVM elaborates that classification has been performed for the two classes. According to the confusion matrix 89% true positive rate with the 11% false positive rate have been predicted while classifying class 1 for the absence of heart disease. It can be easily observed that

68% true positive rate has been achieved with 32% false positive rate for the classification of class 2. 80.4% accuracy for SVM algorithm was experienced in the classification of heart disease. The training time for the proposed algorithm was found to be 0.75427 with the prediction speed of 5900 observations per second.

Fig. 15 illustrates SVM receiver operating characteristics (ROC). To evaluate the multi-class classifier performance, it must be visualized for the analysis. The area under the curve (AUC) determines the degree up to which scale it can classify. Receiver operating characteristics (ROC) is usually measured as the probability of classification. The graph of ROC is plotted between true positive rate and false positive rate. The area under the curve (AUC) was determined as 0.88 and it may be acknowledged as a competent classifier due to the closeness of AUC to 1.

D. K-NN Classifier

Fig. 16 shows the confusion matrix of K-NN algorithm. 76.1% accuracy has been achieved by the K-NN algorithm for the classification of heart disease. The efficiency was calculated using true positive rate and false positive rate. Confusion matrix evaluated the K-NN algorithm and showed that 85% class 1 was classified as true positive rate and false positive rate was achieved as 15% for class 1. For the class 2 classification using K-NN, 62% positive rate was achieved with the 38% of false positive rate.

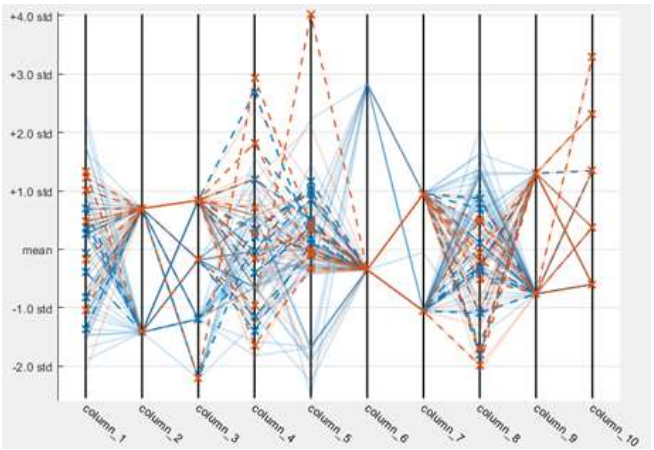

Fig. 13. Prediction Model.

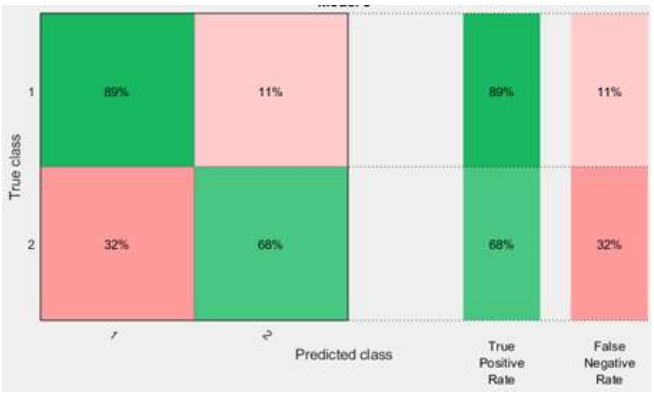

Fig. 14. SVM Confusion Matrix.

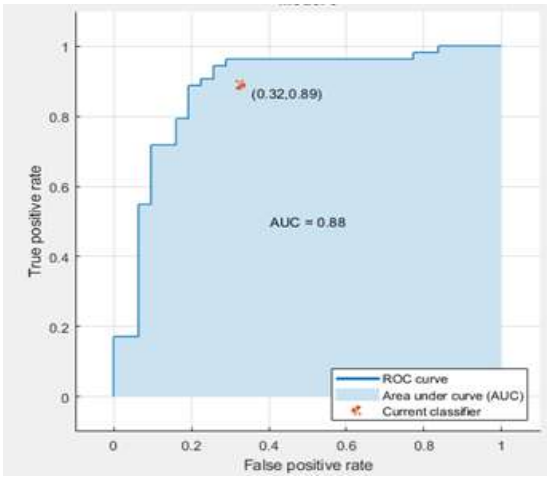

Fig. 15. Support Vector Machine ROC Curve.

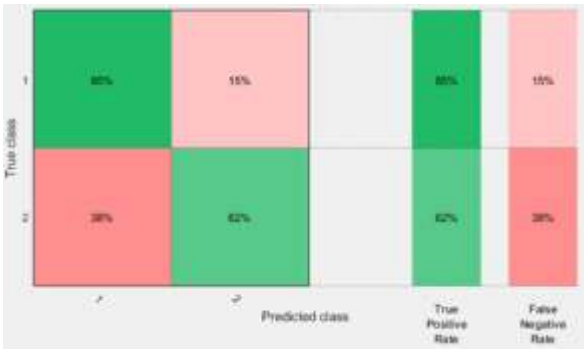

Fig. 16. K-NN Classifier Confusion Matrix.

Usually Euclidean distance is calculated to find out the closest distance with the value of the K.

$$d = \sqrt{(x1 - xA1)^2 + (x2 - xA2)^2} \quad (13)$$

Fig. 17 shows ROC and area under the curve (AUC) for the K-NN classifier. The ROC curve has been plotted between true positive rate and false positive rate. Area under the curve (AUC) was found to be 0.80. It is slightly away from the 1. For good classification AUC must be close to 1.

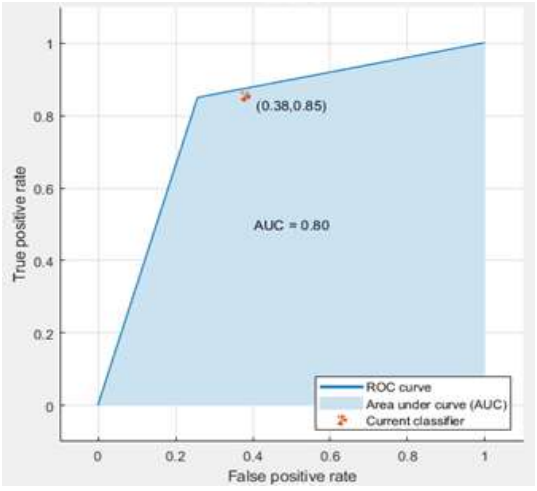

Fig. 17. K-NN ROC and AUC.

E. Linear Discriminant Analyss (LDA)

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are the two most popular classifiers which are based on probabilistic method. For each class predictions can be easily computed using the following mathematical formulation of Baye's rule.

In Fig. 18, linear discriminant analysis (LDA) confusion matrix demonstrated that overall accuracy of 79.3% was achieved based on the true positive rate (TPR) and false positive rate (FPR). For the classification of heart disease, confusion matrix showed that 87% true positive rate was achieved in classifying class 1 while 13% of false positive rate was achieved in the classification of class 1.

Fig. 19 demonstrated that area under the curve (AUC) for the linear discriminant analysis was found to be 0.85.

F. Fine Gaussian SVM

Fig. 20 demonstrated the confusion matrix for fine Gaussian SVM. Confusion matrix of fine Gaussian SVM elaborated that the algorithm performed very poor as it classified all classes as class 1. Class 2 was not predicted at all therefore false negative rate was found to be 100% and false positive rate for class 2 was found to be 0%.

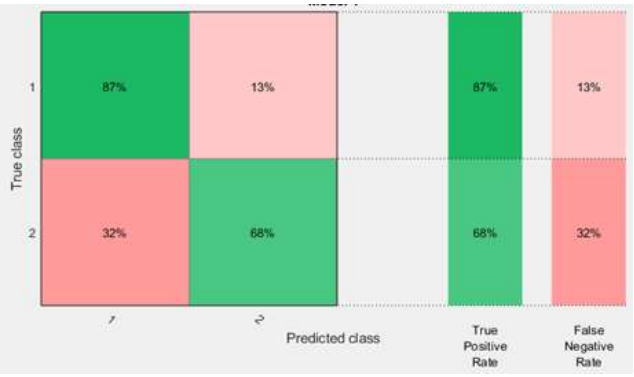

Fig. 18. LDA Confusion Matrix.

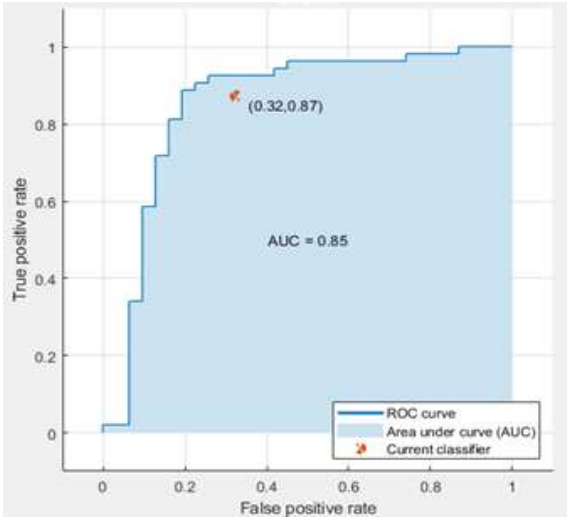

Fig. 19. ROC and AUC Curve for LDA.

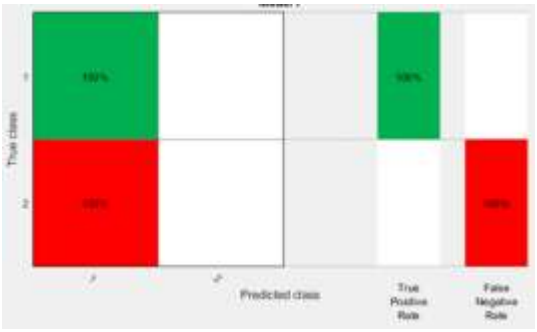

Fig. 20. Confusion Matrix of Fine Gaussian SVM.

G. Performance Comparison of Classifiers for Heart Disease Classification

Table IV proved that the SVM performed better classification compared to the K-NN and LDA in terms of Accuracy, prediction speed, training elapsed time, precision and area of under the curve. 80.4% accuracy was achieved by SVM for the heart disease classification which was greater than the accuracies of K-NN and LDA.

TABLE IV. SVM, KNN AND LDA COMPARISON

Parameters	SVM	K-NN	LDA	Fine Gaussian SVM
Accuracy	80.4%	76.1%	79.3%	59.8%
Prediction Speed	5900 obs/sec	1900 obs/sec	2100 obs/sec	6700 obs/sec
Training time	0.742347 seconds	1.2347 seconds	1.3924 seconds	0.8763 sec
Precision	0.89	0.80	0.83	0.59
AUC	0.88	0.80	0.85	0.60

IV. RESULTS AND CONCLUSION

Comparative study of classifiers was performed to determine the better classifier for the breast cancer prediction. It has been proved from the results that Logistic regression gained highest accuracy of 91.2%. K-NN also performed better with the accuracy of 88.25. Research study shows that logistic regression may be adopted on the real time data set of the patients to reduce the false alarm rate in the prediction of breast cancer tumors. Moreover, simulated results on real time data showed that SVM performed classification rapidly in very less time of 0.74237 seconds compared to the K-NN and LDA for heart disease classification. Prediction was observed 5900 observations per second which is higher than the LDA and K-NN classification algorithms. Accuracies and area under the curve of SVM were found to be 80.4% and 0.88 respectively. SVM proved to be a better and robust classifier for the heart disease classification.

REFERENCES

[1] E. Magnin, D. Vray and A. Brémond, "Early detection of breast cancer using computer assisted diagnosis," 1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Paris, 1992, pp. 849-850.

[2] R. D. Ghongade and D. G. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM

algorithm," 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), Kolkata, 2017, pp. 1-6. doi: 10.1109/IEMENTECH.2017.8076982.

[3] P. Král and L. Lenc, "LBP features for breast cancer detection," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016, pp. 2643-2647. doi: 10.1109/ICIP.2016.7532838.

[4] B. Hela, M. Hela, H. Kamel, B. Sana and M. Najla, "Breast cancer detection: A review on mammograms analysis techniques," 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13), Hammamet, 2013, pp. 1-6. doi: 10.1109/SSD.2013.6563999.

[5] R. Sangeetha and K. S. Murthy, "A novel approach for detection of breast cancer at an early stage using digital image processing techniques," 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2017, pp. 1-4. doi: 10.1109/ICISC.2017.8068625.

[6] V. Krishnaiah, M. Srinivas, G. Narsimha and N. S. Chandra, "Diagnosis of heart disease patients using fuzzy classification technique," International Conference on Computing and Communication Technologies, Hyderabad, 2014, pp. 1-7.

[7] T. Botterill, T. Lotz, A. Kashif and J. G. Chase, "Reconstructing 3-D Skin Surface Motion for the DIET Breast Cancer Screening System," in IEEE Transactions on Medical Imaging, vol. 33, no. 5, pp. 1109-1118, May 2014. doi: 10.1109/TMI.2014.2304959.

[8] Q. Li et al., "Direct Extraction of Tumor Response Based on Ensemble Empirical Mode Decomposition for Image Reconstruction of Early Breast Cancer Detection by UWB," in IEEE Transactions on Biomedical Circuits and Systems, vol. 9, no. 5, pp. 710-724, Oct. 2015.doi: 10.1109/TBCAS.2015.2481940.

[9] S. C. Hagness, A. Taflove and J. E. Bridges, "Two-dimensional FDTD analysis of a pulsed microwave confocal system for breast cancer detection: fixed-focus and antenna-array sensors," in IEEE Transactions on Biomedical Engineering, vol. 45, no. 12, pp. 1470-1479, Dec. 1998. doi: 10.1109/10.730440.

[10] T. Kao et al., "Regional Admittivity Spectra with Tomosynthesis Images for Breast Cancer Detection: Preliminary Patient Study," in IEEE Transactions on Medical Imaging, vol. 27, no. 12, pp. 1762-1768, Dec. 2008. doi: 10.1109/TMI.2008.926049.

[11] D. A. Woten, J. Lusth and M. El-Shenawee, "Interpreting Artificial Neural Networks for Microwave Detection of Breast Cancer," in IEEE Microwave and Wireless Components Letters, vol. 17, no. 12, pp. 825-827, Dec. 2007. doi: 10.1109/LMWC.2007.910466.

[12] P. M. Meaney, M. W. Fanning, D. Li, S. P. Poplack, and K. D. Paulsen,"A clinical prototype for active microwave imaging of the breast,"IEEE Trans. Microw. Theory Tech., vol. 48, no. 11, pp. 1841–1853, Nov. 2000doi: 10.1109/MCS.2009.932223.

[13] E. J. Bond, X. Li, S. C. Hagness, and B. D. Van Veen, "Microwaveimaging via space-time beamforming for early detection of breastcancer," IEEE Trans. Antennas Propag., vol. 51, no. 8, pp. 1690–705, Aug. 2003.

[14] Wolberg, W.H., & Mangasarian, O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193--9196.

[15] Zhang, J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470--479). Aberdeen, Scotland: Morgan.

[16] Talha Khan, Muhammad Alam, Kushsairy Kadir, Zeeshan Shahid and M. S. Mazliham," Artificial Intelligence based Prediction of Seizures for Epileptic Patients: IoT Based Cost Effective Solution ", 2019 IEEE the 7th International Conference on Information and Communication Technology (IEEE-ICOICT), 24-26 July2019, Kuala Lumpur, Malaysia.

[17] Talha Khan, Muhammad Alam, Kushsairy Kadir, Sheroz Khan, M.S Mazliham, Faraz Shaikh, Syed Faiz Ahmed, Zeeshan Shahid" An Implementation of Electroencephalogram signals acquisition to control manipulator through Brain Computer Interface ", 2nd IEEE International Conference on Innovative research and development 2019 (IEEE-ICIRD), 24-26 July 2019, Kuala Lumpur, Malaysia.

[18] Talha Ahmed Khan, Muhammad Alam, Kushsairy Kadir, Zeeshan Shahid, MS. Mazliham, "False Alarm Reduction For The Cardiac Arryhthmias: AI Based Comparative Analysis", Journal of Engineering and Technology", Universiti Kuala lumpur Journal of Engineering and Technology, Vol. 5 (2017).

[19] K. D. Kochanek, J. Xu, S. L. Murphy, A. M. Miniño, and H.-C. Kung, "Deaths: final data for 2009," National Vital StatisticsReports, vol.60, no.3, pp.1–116,2011.

[20] N. Gawande and A. Barhatte, "Heart diseases classification using convolutional neural network," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2017, pp. 17-20. doi: 10.1109/CESYS.2017.8321264.

[21] G. Meena, P. S. Chauhan and R. R. Choudhary, "Empirical Study on Classification of Heart Disease Dataset-its Prediction and Mining," 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, 2017, pp. 1041-1043. [4] S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2017, pp. 1-4. doi: 10.1109/EBBT.2017.7956761.

[22] V. Krishnaiah, M. Srinivas, G. Narsimha and N. S. Chandra, "Diagnosis of heart disease patients using fuzzy classification technique," International Conference on Computing and Communication Technologies, Hyderabad, 2014, pp. 1-7.

[23] Talha Ahmed Khan, Muhammad Alam , M. Junaid Tahir, Kushsairy Kadir , Zeeshan Shahid, M.S Mazliham, "Optimized health parameters using PSO: a cost effective RFID based wearable gadget with less false alarm rate" Indonesian Journal of Electrical Engineering and Computer Science, Vol. 15, No. 1, July 2019, pp. 230~239, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v15.i1.pp230-239.

[24] Y. Özbay, B. Karlik, "A recognition of ECG arrhythmias using artificial neural networks," Proceedings of the 23rd Annual Conference, IEEE/EMBS, Istanbul, Turkey, 2001.

[25] A. Kampouraki, G. Manis, C. Nikou, "Heartbeat time series classification with support vector machines," IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 4, 2009.

[26] G. Evensen, "The ensemble Kalman filter for combined state and parameter estimation," in IEEE Control Systems, vol. 29, no. 3, pp. 83-104, June 2009. doi: 10.1109/MCS.2009.932223.

[27] F. Plesinger, P. Klimes, J. Halamek and P. Jurak, "False alarms in intensive care unit monitors: Detection of life-threatening arrhythmias using elementary algebra, descriptive statistics and fuzzy logic,"2015 Computing in Cardiology Conference (CinC), Nice, 2015, pp. 281-284.2009. doi: 10.1109/MCS.2009.932223.

[28] J. Pan, W.J. Tompkins, "A real-time QRS detection algorithm," IEEE Transactions on Biomedical Engineering, vol. 32, no. 3, 1985.

[29] R. Acharya, A. Kumar, P. S. Bhat, C.M. Lim, S.S. Iyengar, N. Kannathal, S.M. Krishnan, "Classification of cardiac abnormalities using heart rate signals," Med. Biol. Eng. Comput., vol. 42, pp. 288–293, 2004.

[30] M. M. Engin, "ECG beat classification using neuro-fuzzy network," Elsevier, Pattern Recognition Letters, vol. 25, pp. 1715–1722, 2004.

[31] J. P. Kelwade and S. S. Salankar, "Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series,"2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI), Kolkata, 2016, pp. 454-458.

[32] A. M. Amiri and G. Armano, "Early diagnosis of heart disease using classification and regression trees," The 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, 2013, pp. 1-4.

[33] J. P. Kelwade and S. S. Salankar, "An optimal structure of multilayer perceptron using particle swarm optimization for the prediction of cardiac arrhythmias,"2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2016, pp. 426-430.doi: 10.1109/ICRITO.2016.7784993.

[34] P. Majety and V. Umamaheshwari, "An electronic system to recognize heart diseases based on heart sounds: A stochasatic alogorithm implemented on DSK6713," 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2016, pp. 1617-1621.

[35] Gavin Brown. Diversity in Neural Network Ensembles. The University of Birmingham. 2004.