



# Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis

Ajay Sharma<sup>1</sup> · Pramod Kumar Mishra<sup>1</sup>

Received: 10 May 2020 / Accepted: 13 April 2021 / Published online: 15 August 2021  
© Bharati Vidyapeeth's Institute of Computer Applications and Management 2021

**Abstract** Healthcare systems around the world are facing huge challenges in responding to trends of the rise of chronic diseases. The objective of our research study is the adaptation of Data Science and its approaches for prediction of various diseases in early stages. In this study we review latest proposed approaches with few limitations and their possible solutions for future work. This study also shows importance of finding significant features that improves results proposed by existing methodologies. This work aimed to build classification models such as Naïve Bayes, Logistic Regression, k-Nearest neighbor, Support vector machine, Decision tree, Random Forest, Artificial neural network, Adaboost, XGBoost and Gradient boosting. The experimental study chooses group of features by means of three feature selection approaches such as Correlation-based selection, Information Gain based selection and Sequential feature selection. Various Machine learning classifiers are applied on these feature subsets and based on their performance best feature subset is selected. Finally, ensemble based Max Voting Classifier is proposed on top of three best performing models. The proposed model produces an enhanced performance label with accuracy score of 99.41%.

**Keywords** Data science · Feature selection techniques · Machine learning · Breast cancer · Ensemble Learning

## 1 Introduction

Healthcare systems around the world are facing huge challenges in responding to the rising trends of chronic diseases, resources constraints, aging population and the growing focus of citizens on healthy living and prevention [1]. From 2012 to 2030 an assessment report declared nearly an economic loss of 3.6 trillion dollars will occur only due to four chronic disorders i.e., cancer, CRDs, CVDs and diabetes. Major Chronic Disease include Coronary Heart Disease [2], Chronic Kidney Disease [3], Parkinson's disease [4], Alzheimer's disease [5], Diabetes and Hypertension [6], Thyroid Disease [7], High Blood Pressure, Asthma, Cardiovascular Disease, Stoke, Peripheral Arterial Disease, Mental Health problems, and Dementia. Usually, Preventing and screening for disease before they start have been identified as the best ways to prevent the rise of chronic disease. However, in primary care most doctors lack the time, resources and tools to prevent chronic diseases. Traditional models of healthcare focus on one disease but prevent chronic diseases we need a compressive model. The easiest way to predict it in early stages is to analyse the already existing big healthcare data [8]. This is where data science and machine learning come in play to assist doctors to predict disease in early stages. Scientifically, we can add more quality to “Skill India and Make in India” By “Make India Healthy” [1]. Chronic diseases that are chief contributors to mortality and economic loss are Breast cancer, Heart disease, Cervical cancer and Diabetes. Diabetes is a chronic disease that occurs due to lack of insulin production in body. In a report published by IDF in 2013 a measure of 382 million people (8.3% of total) was affected by diabetes with 14 million more (184 million women and 198 million men) men than women. Mainly it is of three types: type-1 diabetes, type-2

✉ Ajay Sharma  
ajay.sharma17@bhu.ac.in

Pramod Kumar Mishra  
mishra@bhu.ac.in

<sup>1</sup> Department of Computer Science, Institute of Science,  
Banaras Hindu University, Varanasi 221005, India

diabetes and gestational diabetes [6]. Heart disease is also a main cause of mortality in India. Within last 26 years rate of mortality due to cardiac disease increased by 34% from 155.7 to 209.1 per lack population. In 2016 nearly 62.5 million premature mortality reported due to cardiac diseases [2]. *Cervical cancer* is the major chronic malignant disease and is the fourth most common disease in women's worldwide [9, 10]. There are several risk factors that develops cervical cancer are smoking, sexual transmitted disease and Human Papilloma virus (HPV) [11]. By identifying all those factors and developing a classification model we can easily predict that the case is malignant or benign [12]. *Breast Cancer* is one of the malignant tumor that accounts for 25% in women's globally according to American cancer society. It can be categorized into two types: Malignant (cancerous) and Benign(non-cancerous). Breast cancer is a collection of diseases in which cells present in breast tissue change their shape and divided abnormally, typically resulting in a lump or mass [13]. According to report published by WHO 1.2 million women's worldwide will be diagnosed breast cancer [14]. Early diagnosis and treatment of such cancerous cells is the only solution to cancer-free environment [15]. Data science is one of the emerging areas of research that incorporates artificial intelligence, machine learning, deep learning, statistics, optimizations and data mining [1]. Data Science approaches already have various applications in early classification of Human activity recognition, Industrial Process monitoring, Intelligent transportation, Quality monitoring, Medical Diagnosis and others. Healthcare sector produces huge volume of data and Data Science techniques supports to extract hidden knowledge that enable new opportunities and innovations to improve population health by addressing different perspectives [1]: (1) descriptive, to diagnose what happened; (2) diagnostic, to diagnose the reason why it happened (3) predictive, to diagnose what will happen and (4) prescriptive, to detect how we can make it happen [1, 16]. Data analytics technologies provide more effective tools [17] that helps to provide Home care, Lifestyle support, Precision medicine, better treatment of chronic disease by early detection, population health and better treatment of infectious diseases [18]. During last two decades researchers have proposed numerous novel ML techniques for predictive data analysis [1]. These useful techniques have been implemented in various data-intensive research areas like healthcare, biology, astronomy to mine hidden patterns [1]. In this article we review latest high-quality articles from major research databases for computer science like IEEEExplore, ACM Digital Library, GoogleScholar, ScienceDirect and SpringerLink on Breast Cancer, heart disease, cervical cancer, diabetes, and other chronic diseases. Also, we present some limitations of the existing work and

their probable solutions that can be extended to other related work.

The Whole structure of this article is described as follows, Sect. 2 discusses Review of latest high-quality work related to major chronic diseases with existing methodologies, datasets and limitations. In Sect. 3 we discuss Material and Methods that are used in this study. Section 4 discusses the results obtained after the implementation of materials and methods listed in Sect. 3. Finally, Sect. 5 ends with a conclusion and future enhancement.

## 2 Literature review

In the recent few years, several diagnostic techniques involving Data Science, Artificial Intelligence, Machine Learning and Deep Learning has been proposed by researchers for diagnosing diseases. In this study we review few latest articles and novel approaches developed in last few years that improves predictive and diagnostic power of existing healthcare systems and need to be further improved.

Early classification of mitral heart valve and aortic heart valve heart-valve disorder “based on linear discriminant analysis (LDA) and adaptive neuro-fuzzy inference system” [19] using dataset of Doppler heart sound (DHS) signals by ultrasound system proposed by Sengur [19] involves Pre-processing, normalizing and Filtering of DHS. Usage of Wavelet transform, short-time Fourier transform and wavelet entropy on DHS signals to extract waveform patterns along with ANFIS and linear discriminant analysis (LDA) for early classification of abnormal or normal heart valves. Although Healthcare sector is data rich but problem of class imbalance is always there but a new feature selection technique PSO-SVM given by Vijayashree and Sultana [20] for Feature Selection in Heart Disease early Classification. A novel function for selecting an optimal weight and fitness function PSO with SVM for selecting more relevant features. The study compares several feature selection methods such as PSO, gain ratio, CFS, filtered subset, Chi-squared, Consistency subset Relief and Info gain algorithms [20] when PSO-SVM is used for feature selection. SVM based classifiers outperforms other classifiers with accuracy effectively raised by 3.09%. For early and accurate classification of various medical disorders a New era of Hybrid Learning came that combines results of multiple algorithms to generate highly accurate classifiers. For “Early classification of Heart Disease a new Hybrid approach [2]” Using Ensemble Hybrid Machine Learning for assisting medical doctors. Process begins from a pre-processing phase followed by feature selection phase based on DT entropy, classification of modelling performance evaluation, and hyperparameter tuning the results with

improved accuracy. Numerous (ML) techniques are used namely, Logistic Regression, Decision Tree, Generalized Linear Model, Support Vector Machine, Gaussian Boosting. This Hybrid learning process involves Decision tree, Random forest and generalized linear model with 13 features extracted based on feature engineering. Basic building block of machine learning process involves data preprocessing, Feature selection, Feature Extraction and selecting best algorithm for classification based on problem. A new intelligent based system for “improved detection of heart disease Based on Random Search Algorithm (RSA) and Optimized Random Forest Model [21]” by JAVEED et al. This involves selection of relevant features and optimized RF with grid search for hyperparameter tuning for heart failure prediction based on Cleveland heart disease dataset. Contribution of Two new experiments: first Random forest and second RSA based RF method. Initially, the dataset is provided to random search algorithm for optimal feature selection. The feature extraction technique RSA extracts 7 most features while the past published work refers 13 features. RSA-RF is efficient that achieves an accuracy of 93.33% on 7 extracted features with improvement in accuracy by 3.3% when compared with other techniques.

Various approaches for Early classification of Time series and diseases have been proposed in recent times. “Early classification of cervical cancer grounded on Support Vector Machine by using various risk factors is [9] proposed by Wen Wu et al. determines relevant 10 risk factors and four target variables Cytology, Hinselmann, Biopsy and Schiller. The experimental study tried to reduce processing time more than other experimental studies by selecting most relevant features by using Principal Component Analysis (PCA) and RFE. Also, over-sampling technique is used with PCA and RFE. Major problem when we work with medical data is class imbalance. To handle this imbalance a new model “using feature reduction technique and SMOTE with random forest for the diagnosis of Cervical Cancer [10]” was designed. Along with random forest and SMOTE two Feature Reduction techniques PCA (principal component analysis) and RFE (Recursive Feature Elimination) is used. Tenfold cross validation is used for training, validation and testing the model. SMOTE-RF shows the rise in accuracy between 1.7 to 3.5% for respective 4 target variables. A new model for Diagnosis of automated cervical cytology and for the realization of squamous epithelial cell automatic detection system [11] was conducted on 500 cervical cells taken from Shenzhen Second People’s Hospital of pathology department. Using neural network model based on faster regions-CNN for classification purpose and cell detection. Model detects five target cells “low grade squamous intraepithelial, endocervical cell, Atypical squamous cells

of undetermined significance lesion, metaplastic squamous and high grade squamous intraepithelial lesion [22]. A new approach for the prediction of Cervical cancer Which is a gynaecological cancer [12]. Various algorithms Like “Recursive feature selection (RFE), Boruta algorithm, ctree () and Simulated annealing (SA)” are used to select right features for ML algorithms on UCI dataset. Efficient feature selection reduces features from 36 to 27 and four target variables are combined to make one target feature named ‘cancer’. ML techniques as KNN, C5.0, SVM, RPART and RF are used with tenfold cross validation whereas Random Forest and C5.0 perform significantly well with accuracy 100% and 99% respectively. Now a days Neuro Fuzzy System and various deep learning methods are used for purpose of early classification. Neuro Fuzzy System based diagnosis of cervical cancer using pap smear images [23] obtained from Leica Microsystems website containing 15 cervical cancer pap smear images. It uses clustering algorithm Fuzzy c-means for the segmentation of images. Finally shape theory was used for detection of segmented images under study to detect the abnormality in cells. Features extraction performed on segmented images extract these “nucleus-cytoplasm ratio (NCR), cytoplasm circularity (CC), nucleus area (NA), cytoplasm area (CA), nucleus circularity (NC) and maximum nucleus brightness (MNB)”. The neuro-fuzzy system is trained using Levenberg–Marquardt (LM) back-propagation technique to predict whether the cancer is malignant or benign. Also, for the prediction of Diabetes a Hybrid prediction system was Developed using C4.5 decision tree and classical kmeans clustering algorithm for assisting doctors to diagnose it in early stages efficiently [24]. Data-processing and feature extraction is performed for correct analysis of diabetes on PIMA Indian dataset. Findings of this proposed study achieves an accuracy of 92.38%. New hybrid approach for the diagnosis of diabetes mellitus based on Adaptive neuro fuzzy inference system with decision Tree for achieving higher [25] accuracy. It generates rules in crisp form by using decision tree classifier and fed into ANFIS as an input after application of Gaussian membership method. Following this, the optimization has been carried out using least square estimation, gradient descent approach and tenfold cross validation method. Also Model for the early diagnosis of type-II diabetes using decision tree approach along with particle swarm optimization with java implementation(J48) of decision tree(C45) with optimized parameters [26]. For the identification of optimized parameter set a self-adaptive Inertial weight with PSO is used and fitness function is standardized using J48 algorithm. Risk factors like “mean blood glucose (MBG), fasting plasma glucose (FPG), postprandial plasma glucose (PPG) and glycosylated hemoglobin (A1c)” are considered for evaluation. Finally,

Fisher's linear discriminant analysis has been used to test the efficiency of the predicted result. Also, a tree-based ML algorithm for diagnosis of diabetes using 8 different ML techniques as base classifiers in 5 different ensembles i.e., DECORATE, random subspace, boosting, bagging, rotation forest [27]. Base classifiers are classification and regression tree, decision tree (C4.5), functional tree, random tree, naïve Bayes tree, reduced error, pruning tree, best-first decision tree and logistic model tree. The exploratory study and performance of base classifiers and different ensembles are thoroughly benchmarked on three different datasets named PIDD, Tabriz dataset, RSMH [28] with AUC as characteristic metric. In Table 1, we listed some important references with source of datasets used, disease specified, algorithm used as well as limitations of study whereas performance measures are specified in italics.

The Challenge regarding dataset imbalance can be easily fixed by making use of *oversampling and under sampling techniques*. Under-sampling need sufficient quantity of data. It reduces the size of an abundant class to convert it into balanced one while Over-sampling handles insufficient quantity of data by increasing the size of rare samples. SMOTE (Synthetic Minority Over-Sampling Technique) is a powerful oversampling approach that is used to handle an imbalanced dataset and is introduced by [10] Chawla et al. Synthetically, it is used to increase the size of rare class by using K-nearest neighbors [32]. Also, to resolve an issue of selecting optimal features, reducing computational time and runtime storage space we can use Particle Swarm Optimization [33] (PSO) that improves predictive power of ML classifiers. PSO is an attractive approach for feature selection which is computationally inexpensive [34] and takes less memory and runtime by taking only few parameters.

### 3 Materials and methods

In Sect. 3, we are going to confer the materials and methods that have been used for finding results. This section is divided into four subsections, i.e., dataset description, proposed methodology, data preprocessing, Feature Selection Techniques, Algorithms used for comparison and performance metrics.

#### 3.1 Dataset description

The comparative and scientific analysis has been performed on publicly available *Wisconsin Breast Cancer* (WDBC) datasets acquired from UCI machine learning repository. The dataset was originally given by University of Wisconsin comprises of 569 samples (212-Malignant

and 357-benign) with each sample having 32 features. These features present the basic characteristics of the breast mass cell nuclei in the image. Generally, 30 real-valued input features present with id and diagnosis field. Field 2 named "diagnosis" involving 2 classes malignant and benign. By Using every cell nucleus 10 real valued variables are calculated. These variables are: radius, perimeter, area, texture, concavity, compactness, smoothness, concave points, symmetry, fractal dimension.

#### 3.2 Proposed methodology

Our proposed methodology for early classification is based on Max Voting Procedure. We used multiple models for predicting each data sample. Final prediction of voting classifier is based on majority vote that we get from the majority of the models [35]. After the data preprocessing we use several feature selection techniques listed in Table 2 and apply various algorithms shown in Table 3. In Table 3 results of best performing classification models are specified in bold. Finally, a best feature selection technique is selected and based on predictions of three top performing classifiers a new voting classifier is proposed. Workflow of the proposed methodology is shown in Fig. 1.

#### 3.3 Dataset preprocessing

Data preprocessing is a useful step that helps to remove noise, inconsistencies and redundancy to achieve high quality data which improves the performance. During the data pre-processing, all the missing values are filled by the mean of the corresponding feature. Formatting of the given dataset is ensured to be consistent. All the incorrect data types of the features changed to their required datatype. Data normalization is performed on the given dataset to make its range consistent. We used z-score normalization on all the features to restrict the range of values between 3 to  $-3$ . Z-score is calculated by the Eq. (1).

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

Generally, some of machine learning algorithms does not handle categorical data. Diagnosis feature in this dataset contains two categories M(malignant) and B(benign) is replaced numeric with M by 1 and N by 0. Also, feature ID is removed from the dataset as it is not necessarily required.

#### 3.4 Feature selection techniques

Feature selection is one of the useful stages in data modelling that helps to discard redundant, irrelevant and noisy

**Table 1** Literature review over medical data by various authors

References	Disease specified	Algorithm applied	Dataset used	Accuracy (%)	Limitations
Sengur [19]	Heart-valve disorder	LDA, ANFIS	Classes = 2, Samples = 215 <i>DHS signals</i>	$A = 95.03\%$ , $Se = 95.9\%$ , $Sp = 94\%$	Computation takes more time. No cross-validation
Vijayashree and Sultana [20]	Heart disease	PSO, SVM, MLP, RF, NB	Cleveland heart dataset (UCI), extracted features-13	$A = 88.22\%$	No cross-validation. Lack of Hybrid approach
Senthil Kumar Mohan et.al [2]	Heart disease	LR, NB, DT, GLM, DL, SVM, RF, GBT	UCI (Classes-2, records-303, features-76)	$A = 88.7\%$	Dataset is small. Training time is more
Javeed et al. [21]	Heart valve disorder	RSA and RF	Cleveland heart dataset (UCI), extracted features-7	$A = 93.33\%$	Training time is more
Wen Wu et al. [9]	Cervical cancer	SVM, SMOTE, PCA, RFE	UCI (Classes-2, records-858, features-32, 4 target variables)	$A = 93.97\%$	Lack of Optimization
Sherif fayz1 et al. [10]	Cervical cancer	RF, SMOTE, PCA, RFE, tenfold cross validation	UCI	$A = 97.60\%$	Computation takes more time
Nithya et al. [12]	Cervical cancer	KNN, C5.0, SVM, RFE, RF, RPART, Boruta,tenfold cross validation, Simulated annealing	UCI	$A = 100\%$	Class Imbalance
Meiquan et al. [11]	Cervical cytology	Faster regions-CNN	500 samples, 50 negative and 450 positives	$A = 78\%$	Imbalanced dataset. Less accuracy
Kar and Majumder [23]	Cervical cancer	NFS, Fuzzy c-means, Levenberg–Marquardt (LM), ANFIS	Classes = 2 Samples = 15 <i>Pap Smear image</i>	$A = 100\%$ $Se = 100\%$ $Sp = 100\%$	Data set is too small
Patil et al. [24]	Diabetes	Kmeans clustering and C4.5 decision tree, k-fold cross-validation	Pima Indian Diabetes	$A = 92.38\%$	Class imbalance. Lack of SMOTE, PSO
Chen et al. [25]	Diabetes Mellitus	Decision Tree—ANFIS <i>Ten-fold cross validation</i>	Classes = 2, Samples = 76, Attributes = 9 <i>Clinical data</i>	$A = 75.67\%$	Class imbalance. Less accuracy
Sarwar et al. [29]	Type-II diabetes	SVM, ANN, KNN, naïve bayes and Ensemble	Classes-2, Instances-400, Features-10,	$A = 98.60\%$	Data set is too small. No Cross validation
Tama and Rhee [27]	Diabetes	Tree-based ML algorithm	PIDD (768,8), Tabriz (435,11) dataset, RSMH (2536,13)	$A = 82.81\%$	Class imbalance. Less accuracy
Abdullah and Selvakumar [26]	Type-II diabetes	Decision tree, PSO, J48	Classes-2, instances-732, Features-23	$A = 97.21\%$	Class imbalance. Only one approach
Asri et al. [14]	Breast cancer	<i>C4.5, SVM, KNN, NB</i>	WBCD (UCI). Samples:569	93	Less accuracy. Lack of ensemble learning
Nematzadeh et al. [30]	Breast cancer	<i>ANN (tenfold)</i>	WPBC(UCI)	98.09%	Lack of Hybrid learning
Gayathri and Sumathi [31]	Breast cancer	<i>RVM</i>	WBCD (UCI). Samples:569	96%	Less Accuracy

data. The proposed study selects relevant features by using following feature selection techniques.

### 3.4.1 Correlation based feature selection (CFS)

This is one of a useful filter method approach to select features which is faster than wrapper-based approach and is suitably used for Higher-dimensional datasets. CFS

**Table 2** Summary of selected features using three feature selection techniques

Feature extraction technique	Number of features	Selected features
Correlation based feature selection	11	'texture_mean', 'concave points_mean', 'area_se', 'symmetry_se', 'radius_worst', 'concave points_worst', 'smoothness_worst', 'perimeter_worst', 'concavity_worst', 'concavity_mean', 'area_worst'
Sequential feature selection	7	'concavity_mean', 'texture_se', 'concave points_se', 'radius_worst', 'perimeter_worst', 'texture_worst', 'smoothness_worst'
Information gain (IG)	14	'concavity_mean', 'radius_mean', 'concave points_worst', 'perimeter_mean', 'radius_worst', 'area_mean', 'perimeter_worst', 'area_worst', 'area_se', 'concave points_mean', 'concavity_worst', 'radius_se', 'smoothness_se', 'texture_worst'

technique choose relevant features only on the basics of intrinsic characteristics of the data without implementation of any ML algorithms [36]. Often in a dataset some of the features maybe highly correlated to some other attributes. Such highly correlated features gives redundant information that does not contribute in performance improvement. CFS technique calculates the correlation to other features and excludes the highly correlated (similar) features [37]. Suppose the features p1 and p2 are highly correlated means both carry same type of information. The classification model including both p1 and p2 have same predictive power as dataset containing either p1 or p2. Similarly, the features that are highly correlated with the class label are retained. The most widely used pearson correlation coefficient is expressed by Eq. (2):

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (2)$$

### 3.4.2 Sequential feature selection (SFS)

It is a feature selection technique that reduces n-dimensional feature space to m-dimensional feature space with ( $m < n$ ) based upon a greedy search approach. It selects or discards features automatically one at a time on the basics of classifier performance until a new feature subset with best performance is obtained [38]. This wrapper approach is slower than correlation because it selects an optimal subset that increases the performance of classifier.

### 3.4.3 Information Gain (IG) attribute evaluation

IG is a useful feature selection technique that selects features by building a decision tree by using test attribute at every node of DT. This particular approach was introduced by J. R. Quinlan. [39] Consider node N that represent the partition D records from dataset. Finding IG helps in selecting an attribute for splitting at node N. The attribute having maximum IG is carefully chosen for splitting.

Higher the IG of an attribute minimizes the information required for classification of objects in partitions. Such an approach helps to improve the performance of decision tree-based algorithms. IG required for classifying objects in partition D is calculate by Eq. (3). Where  $p_i$  represents a probability of an object D belonging to class  $c_i$ .

$$\text{Info}(D) = - \sum_{i=1}^m p_i \log(p_i) \quad (3)$$

## 3.5 Algorithms used for study

A brief description of the algorithms selected for scientific study is given under:

### 3.5.1 Logistic regression

It is one of a useful supervised learning algorithm that can solve regression as well as classification problem. This statistical model is used to predict binary values (zero or one). It is widely used in healthcare sector and gives definite output values. Disadvantage of linear regression is capability of predicting only continuous variables but in case of categorical feature logistic regression is found useful [40]. This classification technique is based on the sigmoid or logistic function  $\left(\frac{1}{1+e^{-t}}\right)$ . The logistic regression is represented by Eq. (4).

$$LR = \frac{1}{1 + e^{-(Y=\beta_0+\beta_1X_1+\beta_2X_2+\dots+\beta_nX_n)}} \quad (4)$$

### 3.5.2 K-nearest neighbor

KNN is a non-parametric method of classification and can be used to solve both regression as well as classification problem. This algorithm responds to an input vector where the units are located near each other. It works only on the basis of the stored trained database without construction of

any general model so called a lazy learner. It categorises the new input vector on the basis of majority vote of k nearest neighbours irrespective of their labels assigned. In case we have M training vectors, the KNN technique computes K nearest neighbors to the test data, then categorise new input by taking majority vote of classes among k nearest neighbors. It is a distance-based function so scaling and normalization of features is useful in improving performance of K nearest neighbor classifier. The Euclidean distance is calculated by the Eq. (5).

$$\text{Euclidean} = \sqrt{\sum_{i=1}^m (x_i - y_i)^2} \tag{5}$$

### 3.5.3 Decision tree

It is a supervised machine learning classification algorithm that is constructed for training samples of dataset D on the basics of high entropy values. Construction of tree is very fast and simple by using recursive top-down DAC (divide and conquer) approach. Also, irrelevant samples on D are removed using tree pruning. Entropy is represented by Eq. (6).

$$\text{Entropy} = - \sum_{j=1}^m p_{ij} \log_2 p_{ij} \tag{6}$$

### 3.5.4 Random forest

It is a supervised ML learning classification ensemble technique that is based on many decision trees. This ensemble model constructs various decision trees that are incorporated to get the better performance. Mainly bagging or bootstrap aggregating is applied for tree learning. Consider a given data,  $X = \{x_1, x_2, \dots, x_n\}$  be input vectors and  $Y = \{y_1, y_2, \dots, y_n\}$  be a response variable with bagging b repeated from 1 to B. Finally, by averaging the predictions of all individual decision tree a new unseen sample  $x'$  are created as Eq. (7):

$$j = \frac{1}{B} \sum_{b=1}^B fb(x') \tag{7}$$

The standard deviation is used for measuring the uncertainty of their predictions on these trees specified by Eq. (8)

$$\sigma = \sqrt{\frac{\sum_{b=1}^B (fb(x') - f)^2}{B - 1}} \tag{8}$$

### 3.5.5 Support vector machine

It is a supervised ML technique that has been effectively used for classifying linear as well not linear problems. Because of its complexity, even it is highly accurate in higher dimensional spaces for classification as well as outlier detection. Let the dataset D having training samples data =  $\{x_i, y_i\}$  for all  $i = 1, 2, \dots, n$  where  $y_i$  and  $x_i \in R^n$  be a target item and an ith vector [41]. It represents an optimal hyperplane by  $f(x) = w^T x + b$  where w and b are dimensional coefficient and offset vector.

### 3.5.6 Neural networks

It is a supervised ML classification algorithm that is inspired by the functioning of human brain. It has three main components i.e., one input layer ( $x_i$ ), hidden layers and one output layer ( $y_i$ ). The number of hidden layers present in any ANN is atleast one [40]. The strength of neural network is dependent upon weight associated with it. Activation function plays a vital role in NN for giving final result by adding of a bias value [42]. Non-linearity in NN is achieved by this function and is represented by Eq. (9).

$$f\left(b + \sum_{i=1}^n x_i u_i\right) \tag{9}$$

## 3.6 Performance measures

These classification evaluation measures used are accuracy, Precision, Recall, specificity, F-measure [43]. These metrics are calculated by using the elements of confusion matrix that represents an information about predicted and actual values. The performance metrics are depicted by following Eqs ((10, 11, 12, 13, 14):

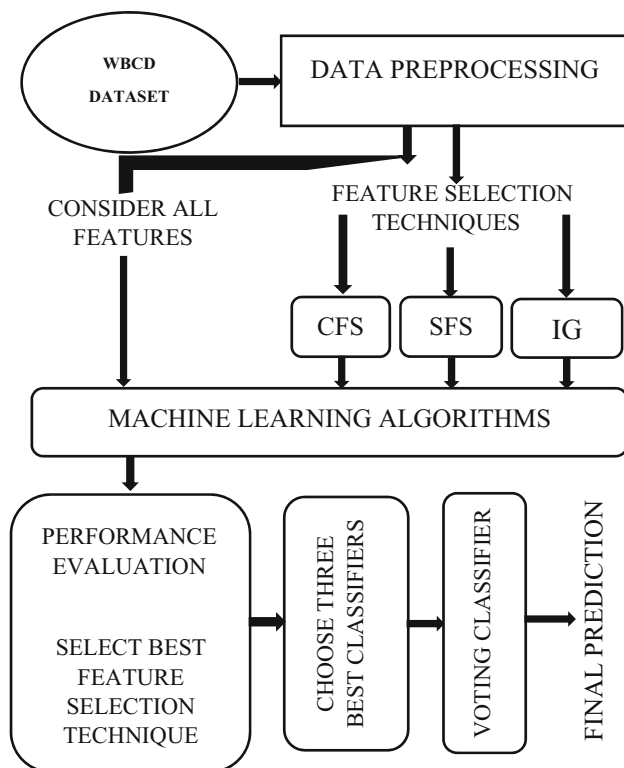
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$F1\text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{14}$$



**Fig. 1** Experiment proposed methodology

## 4 Results and discussion

In this section, we are going to discuss the prediction results of the breast cancer classification before and after using feature selection techniques using Logistic Regression (LR), Decision Tree (DT), Support vector machine (SVM), Artificial neural network (ANN), Adaboost and XGBoost. Based on these results best feature selection technique is decided and a voting classifier is proposed.

Experimental study implements several state-of-the-art classifiers on the features listed in Table 2 using corresponding feature selection technique. Table 3 compares the accuracy, precision, recall, F-measure of various classifiers by considering all the features along with features using CFS, SFS and Information Gain. Findings shows that Neural network achieved a higher accuracy with measure 98.83% and naive bayes achieved minimum accuracy of 95.90% using all the features. But due to the presence of some redundant features it takes more CPU time and memory for computation. Results shown in Table 3 insights that there is slightly increase in performance of some models but the overall CPU time and memory required is comparatively less than using all the features. Firstly, Correlation based feature selection (CFS) approach helps to select only 11 features having high correlation value with the class(diagnosis) and least correlation value with other features. Features such as perimeter, area and

radius show almost linear patterns that insights the presence of multicollinearity between these variables. Similarly, the relationship between compactness, concave\_points and concavity also display the presence of multicollinearity. Absence of some of these features may not affect the performance of the model. We select features with filter value 0.65 with the feature diagnosis representing class labels. Features such as ‘radius\_worst’, ‘perimeter\_worst’ and ‘area\_worst’ having correlation value of 0.78, 0.78 and 0.73 with the diagnosis are retained. Further, ‘concave points\_mean’, ‘concavity\_mean’, ‘concavity\_worst’ and ‘concave points\_worst’ features having 0.78, 0.7, 0.66 and 0.79 are also retained. ‘texture\_mean’, ‘symmetry\_se’ and ‘smoothness\_worst’ features are least correlated with all other features present are also retained. Machine learning classifiers are developed by considering these 11 features selected using a CFS technique and accuracy is computed for each modelling technique. Table 3 compares the accuracy, precision, recall, F-measure on features selected using CFS. The peak accuracy is achieved by support vector machine with measure 99.11%.

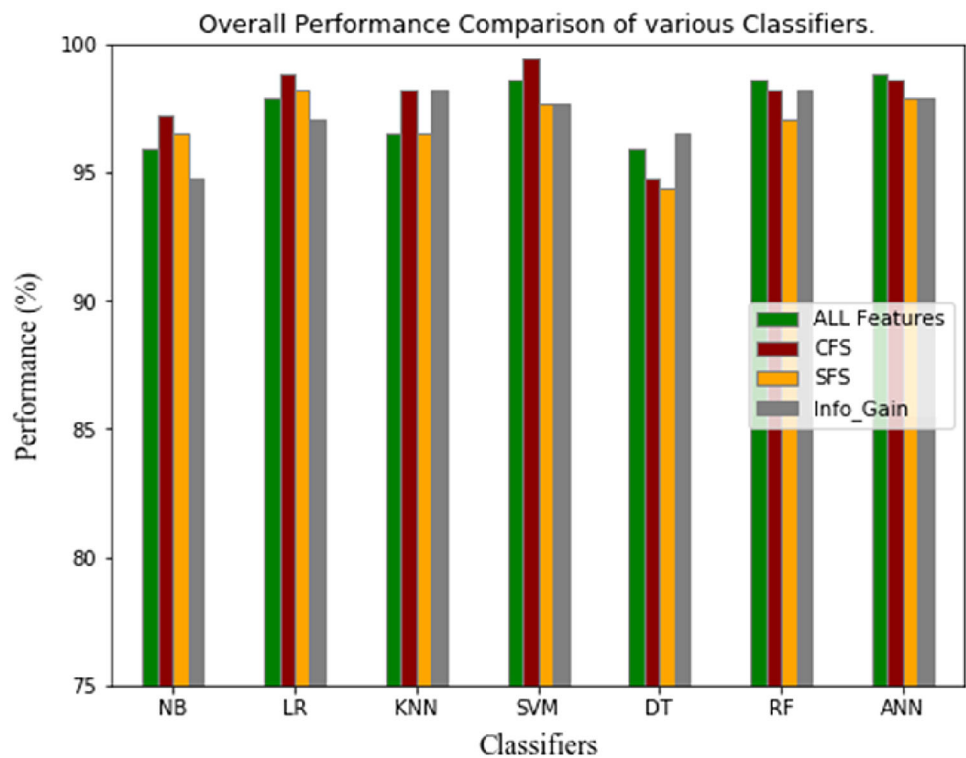
Secondly, based on sequential SFS approach used to select or discard one feature at a time until an optimal feature subset with best performance is achieved. Based on SFS we select 7 features named ‘concavity\_mean’, ‘texture\_se’, ‘concave points\_se’, ‘radius\_worst’, ‘perimeter\_worst’, ‘texture\_worst’ and ‘smoothness\_worst’ listed in Table 2. Logistic Regression outperforms all other ML classifiers with accuracy measure of 98.83%. Finally, an optimal feature set selection based on highest Information Gain we consider 14 features that are listed in Table 2. Performance measures for different ML classifiers based on IG are computed shown in Table 3. Based on results shown in Table 3 we have seen that the performance order of various models. Overall individual performance comparison of various ML Classifiers by using four feature subsets (All features, CFS, SFS, Info\_Gain) is depicted in “Fig. 2”. It depicts that ANN surpassed all other predictors with least error rate of 1.17% by using all features. The performance order based on CFS SVM > LR > NN > RF > KNN > DT. Overall Highest performance is achieved by SVM Classifier followed by Logistic Regression and neural network classifier as shown in Fig. 2 in case of CFS features. Finally, we make a voting based classifier by using SVM, LR and ANN that predicts results with 99.41% accuracy. In Table 4 we Compared results of our proposed model with various existing models where bold text signifies the accuracy and error of proposed classification model.



**Table 3** Performance comparison of classifiers using different feature subsets

	Models	Accuracy	Precision	Recall	F-Score
Classifiers performance with all features in feature set	Naïve Bayes	95.10	92.45	94.23	93.93
	Logistic regression	97.90	100	94.11	96.96
	K-Nearest neighbor	97.07	96.15	92.59	94.33
	Support vector machine	98.60	98.11	98.11	98.11
	Decision tree	95.90	98.43	91.30	94.73
	Random forest	98.24	96.29	100	98.11
	Neural network	98.83	100	96.61	98.27
	AdaBoost	97.66	98.38	95.31	96.82
	Gradient Boosting	93.70	92	90.19	91.08
	XGBoost	96.50	96	94.11	95.04
Classifiers performance with features selected based on CFS	Naïve Bayes	97.20	94.94	98.07	96.22
	Logistic regression	<b>98.83</b>	<b>98.21</b>	<b>98.21</b>	<b>98.21</b>
	K-Nearest neighbor	98.24	98.27	96.61	97.43
	Support vector machine	<b>99.11</b>	<b>100</b>	<b>98.43</b>	<b>99.21</b>
	Decision Tree	94.73	88.52	96.42	92.30
	Random forest	98.24	95.1	100	97.52
	Neural network	<b>98.60</b>	<b>98.07</b>	<b>98.07</b>	<b>98.07</b>
	AdaBoost	96.50	96	94.11	95.04
	Gradient boosting	93.70	93.75	88.23	90.90
	XGBoost	96.50	96	94.11	95.04
Classifiers performance with features selected based on SFS	Naïve Bayes	96.50	94.33	96.15	95.23
	Logistic regression	98.24	96.49	98.21	97.34
	K-Nearest neighbor	96.49	98.18	91.52	94.73
	Support vector machine	97.66	98.38	95.31	96.82
	Decision tree	94.40	90.74	94.23	92.45
	Random forest	97.07	95	96.61	95.79
	Neural network	97.90	98.21	96.49	97.34
	AdaBoost	94.40	90.56	94.11	92.30
	Gradient boosting	95.10	95.83	90.19	92.92
	XGBoost	95.80	95.91	92.15	94
Classifiers performance with features selected based on Information Gain (IG)	Naïve Bayes	94.73	92.72	91.07	91.89
	Logistic regression	97.07	93.22	98.21	95.65
	K-Nearest neighbor	98.24	96.66	98.30	97.47
	Support vector machine	97.66	94.82	98.21	96.49
	Decision tree	96.50	91.66	97.77	94.62
	Random forest	<b>98.24</b>	<b>95.16</b>	<b>100</b>	<b>97.52</b>
	Neural network	97.90	94.54	100	97.19
	AdaBoost	95.80	92.45	96.07	94.23
	Gradient boosting	93	93.61	86.27	89.79
	XGBoost	96.50	97.91	92.15	94.94

**Fig. 2** Overall performance comparison of various classifier



**Table 4** Comparison of proposed model with various models

Reference	Methodologies	Best model	Accuracy (%)	Classification error
Nematzadeh et al. [30]	SVM, DT, NB, ANN	ANN (tenfold)	98.09	1.91
Gayathri and Sumathi [31]	NB, SVM, NN, RVM	RVM	96	4
Asri et al. [14]	C4.5, SVM, KNN, NB	SVM	93	7
Ojha and Goel [44]	k-means, DT (c5.0) KNN, NB, SVM	DT (C5.0)	81.03	18.97
Proposed model	VOTING CLASSIFIER (ANN + SVM + LR)		<b>99.41</b>	<b>0.59</b>

## 5 Conclusion and future scope

In the current era, large volume of data is generated by using various sensors and machines in the healthcare domain named as Big Data. Early classification and analysis of this data is a challenge to predict diseases in early stages and is a hot topic of research. Machine learning and Deep learning evolution helps to do predictive analysis of such high-volume data. In our work we choose well established ML algorithms such as KNN, LR, DT, ANN, SVM, RF, Adaboost etc. based on various feature selection techniques such as CFS, SFS and Information Gain. Features selected using CFS achieves higher accuracy as comparative to other techniques. Finally, we build a voting classifier by combining results of best three models SVM, LR and ANN to classify the new test samples. Results shows that voting classifier can predict results with 99.41% accuracy.

In future, machine learning methodologies can be integrated with neuro-fuzzy and deep learning approaches for efficient diagnosis. A new hybrid models can be proposed by hybridization of Machine learning, CNN/Auto-encoder, neuro-fuzzy systems and swarm optimizations to integrate Optimum feature selection, Feature extraction and Classification for medical imaging data. Further, there are some innovative evolutionary optimization algorithms like Genetic Programming, Simulated annealing, Gradient Descent, Stochastic Optimization, Swarm optimization, SMOTE, Cuckoo search algorithm, ant colony optimization, Hunting search algorithm, Firefly algorithm, lyapunov exponents, Glow worm algorithm, Bat algorithm and wavelet transformations can be combined with Machine learning, Neuro-fuzzy systems, Deep learning and Artificial intelligence methodologies to develop hybrid models.

**Acknowledgements** The authors are highly thankful to the editor and reviewers for kind suggestions and critical comments for improving the quality of the paper.

## References

- Consoli S, Recupero DR, Petkovic M (2019) Data science for healthcare. Springer International Publishing, Berlin
- Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7:81542–81554
- Qin J, Chen L, Liu Y, Liu C, Feng C, Chen B (2019) A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access* 8:20991–21002
- Haq AU, Li JP, Memon MH, Malik A, Ahmad T, Ali A, Shahid M (2019) Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings. *IEEE Access* 7:37718–37734
- Sampath R, Saradha A (2015) Alzheimer's disease classification using hybrid neuro fuzzy Runge-Kutta (HNFRK) classifier. *Res J Appl Sci Eng Technol* 10(1):29–34
- Fitriyani NL, Syafrudin M, Alfian G, Rhee J (2019) Development of disease prediction model based on ensemble learning approach for diabetes and hypertension. *IEEE Access* 7:144777–144789
- Poudel P, Illanes A, Ataide EJ, Esmaili N, Balakrishnan S, Friebe M (2019) Thyroid ultrasound texture classification using autoregressive features in conjunction with machine learning approaches. *IEEE Access* 7:79354–79365
- Kour H, Manhas J, Sharma V (2020) Usage and implementation of neuro-fuzzy systems for classification and prediction in the diagnosis of different types of medical disorders: a decade review. *Artif Intell Rev* 53:4651–4706
- Wu W, Zhou H (2017) Data-driven diagnosis of cervical cancer with support vector machine-based approaches. *IEEE Access* 5:25189–25195
- Abdoh SF, Rizka MA, Maghraby FA (2018) Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access* 6:59475–59485
- Meiquan X et al. (2018) Cervical cytology intelligent diagnosis based on object detection technology. In: *Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, Amsterdam, The Netherlands (2018)
- Nithya B, Ilango V (2019) Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. *SN Appl Sci* 1(6):641
- Howell A, Sims AH, Ong KR, Harvie MN, Evans DGR, Clarke RB (2005) Mechanisms of disease: prediction and prevention of breast cancer: cellular and molecular interactions. *Nat Clin Pract Oncol* 2(12):635–646
- Asri H, Mousannif H, Al Moatassime H, Noel T (2016) Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Proced Comput Sci* 83:1064–1069
- Mohandas M, Deriche M, Aliyu SO (2018) Classifiers combination techniques: a comprehensive review. *IEEE Access* 6:19626–19639
- Jain D, Singh V (2018) Feature selection and classification systems for chronic disease prediction: a review. *Egypt Inform J* 19(3):179–189
- Mishra S, Tripathi AR (2019) Platforms oriented business and data analytics in digital ecosystem. *Int J Financ Eng* 6(04):1950036
- Ketu S, Mishra PK (2021) Empirical analysis of machine learning algorithms on imbalance electrocardiogram based arrhythmia dataset for heart disease detection. *Arabian Journal for Science and Engineering*, pp 1–23
- Sengur A (2008) An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases. *Expert Syst Appl* 35(1–2):214–222
- Vijayashree J, Sultana HP (2018) A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. *Program Comput Softw* 44(6):388–397
- Javeed A, Zhou S, Yongjian L, Qasim I, Noor A, Nour R (2019) An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. *IEEE Access* 7:180235–180243
- Mishra S, Tripathi AR (2020) IoT Platform Business Model for Innovative Management Systems. *Int J Financ Eng (IJFE)* 7(03):1–31
- Kar S, Majumder DD (2019) A novel approach of mathematical theory of shape and neuro-fuzzy based diagnostic analysis of cervical cancer. *Pathol Oncol Res* 25(2):777–790
- Patil BM, Joshi RC, Toshniwal D (2010) Hybrid prediction model for type-2 diabetic patients. *Expert Syst Appl* 37(12):8102–8108
- Chen T, Shang C, Su P, Antoniou G, Shen Q (2018) Effective diagnosis of diabetes with a decision tree-initialised neuro-fuzzy approach. *UK workshop on computational intelligence*. Springer, Cham, pp 227–239
- Abdullah AS, Selvakumar S (2019) Assessment of the risk factors for type II diabetes using an improved combination of particle swarm optimization and decision trees by evaluation with Fisher's linear discriminant analysis. *Soft Comput* 23(20):9995–10017
- Tama BA, Rhee KH (2019) Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. *Artif Intell Rev* 51(3):355–370
- Mishra S, Tripathi AR (2021) AI business model: an integrative business approach. *J Innov Entrepreneurship* 10(1):1–21
- Sarwar A, Ali M, Manhas J, Sharma V (2020) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *Int J Inf Technol* 12(2):419–428
- Nematzadeh Z, Ibrahim R, Selamat A (2015) Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. In: *2015 10th Asian Control Conference (ASCC)*, IEEE, pp 1–6
- Gayathri BM, Sumathi CP (2016) Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer. In: *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICIC)*, IEEE, pp 1–5
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 16:321–357
- Muthukaruppan S, Er MJ (2012) A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. *Expert Syst Appl* 39(14):11657–11665
- Xu R, Anagnostopoulos GC, Wunsch DC (2007) Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. *IEEE/ACM Trans Comput Biol Bioinf* 4(1):65–77
- Mishra S, Tripathi AR (2020) Literature review on business prototypes for digital platform. *J Innov Entrepreneurship* 9(1):1–19
- Ketu S, Mishra PK (2021) Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare. *Complex & Intelligent Systems*, pp 1–19

37. Mishra S (2018) Financial management and forecasting using business intelligence and big data analytic tools. *Int J Financ Eng* 5(02):1850011
38. Ketu S, Mishra PK (2021) Hybrid classification model for eye state detection using electroencephalogram signals. *Cognitive Neurodynamics* pp 1–18
39. Karegowda AG, Manjunath AS, Jayaram MA (2010) Comparative study of attribute selection using gain ratio and correlation-based feature selection. *Int J Inform Technol Knowl Manag* 2(2):271–277
40. Shailaja K, Seetharamulu B, Jabbar MA (2018) Machine learning in healthcare: a review. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, pp 910–914
41. Ketu S, Mishra PK (2021) Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection. *Appl Intell* 51(3):1492–1512
42. Ketu S, Mishra PK (2020) A hybrid deep learning model for COVID-19 prediction and current status of clinical trials worldwide. *Comput Mater Contin* 66(2)
43. Sharma A, Mishra PK (2020) State-of-the-art in performance metrics and future directions for data science algorithms. *J Sci Res* 64(2):221–238
44. Ojha U, Goel S (2017) A study on prediction of breast cancer recurrence using data mining techniques. In: 2017 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence, IEEE, pp 527–530