IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received December 30, 2020, accepted January 27, 2021, date of publication February 1, 2021, date of current version February 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3055806

Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques

AMIN UL HAQ[1], JIAN PING LI[1], ABDUS SABOOR[1], JALALUDDIN KHAN[1], SAMAD WALI[2], SULTAN AHMAD[3], (Member, IEEE), AMJAD ALI[4], GHUFRAN AHMAD KHAN[5], AND WANG ZHOU[6]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Department of Mathematics, Namal Institute, Mianwali 42250, Pakistan
[3]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[4]Department of Computer Science and Software Technology, University of Swat, Mingora 19130, Pakistan
[5]School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[6]School of Computer and Software Engineering, Xihua University, Chengdu 610039, China

Corresponding authors: Amin Ul Haq (khan.amin50@yahoo.com) and Jian Ping Li (jpli2222@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61370073, in part by the National High Technology Research and Development Program of China under Grant 2007AA01Z423, and in part by the Project of Science and Technology Department of Sichuan Province under Grant H04010601W00614016.

ABSTRACT Breast cancer is one the most critical disease and suffered many people around the world. The efficient and correct detection of breast cancer is still needed to ensure this medical issue although the researchers around the world are proposed different diagnostic methods for detection of this disease, however these existing methods still needed further improvement to correct and efficient detection of this disease. In this study, we proposed a new breast cancer identification method by using machine learning algorithms and clinical data. In the proposed method supervised (Relief algorithm) and unsupervised (Autoencoder, PCA algorithms) techniques have been used for related features selection from data set and then these selected features have been used for training and testing of classifier support vector machine for accurate and on time detection of breast cancer. Additionally, in the proposed approach k fold cross validation method has been used for model validation and best hyperparameters selection. The model performance evaluation metrics have been used for model performance evaluation. The BC data sets have been used for testing of the proposed method. The analysis of experimental results has been demonstrated that the features selected by Relief algorithm are more related for accurate detection of Breast cancer instead of features selected by Auotencoder and PCA algorithms. The proposed method has been attained high results in terms of accuracy on selected feature selected by Relief algorithm and achieved 99.91% accuracy. We have been employed McNemar's statistical test for performance comparison of our different models. Further, the proposed method performance has been compared with baseline methods in the literature and the proposed method performance is high as compared to base line methods. Due to the high performance of the proposed method (Relief-Support vector machine) we highly recommended it for the diagnosis of breast cancer. In addition, the proposed method can be easily incorporated into the healthcare system for reliable diagnosis of Breast cancer.

INDEX TERMS Machine learning algorithms, breast cancer detection, accuracy, feature selection, clinical data.

I. INTRODUCTION

Breast Cancer (BC) is a dangerous disease and suffered many women across the world [1]. In 2018 there were 2 million fresh cases reported. The 5^{th} big reason of females death is BC comparatively to cancers in terms of all types. The malignant tumor of BC which produced inside breast cells. A group of splitting cells that form a lump or mass of extra tissue which is called Tumors and these tumors can be whichever cancerous (malignant) or non-cancerous (benign). In [2] different countries with the advanced developed medical technology accumulate the 5-year survival rate of first stages BC

The associate editor coordinating the review of this manuscript and approving it for publication was Bohui Wang.

is (80-90%), and decreasing up to 24% for identification of BC at the first stages. In order to recognize, the BC different invoice approaches have been used. Biopsy approach [3], tissues of breast are use for detection of cancer, and highly accurate results achieved. However, the process of biopsy is painful for the patient. Similarly, BC detection technique is [4] mammogram. In this method of diagnosis 2-Dimensional projection image is design from breast. However, this method is not reliable for detection of breast cancer. Magnetic Reasoning imaging (MRI) is used for BC detection [5]. These invoice methods are not effective for BC detection [6].

In order to handle these difficulties in invasive based methods for detection of BC, a non-invasive based methods, such as machine learning (ML) methods are highly suitable for detection of breast cancer. Thus, the early stage recognition of breast cancer is necessary for proper treatment and recovery. To diagnosis the BC, different methods have been proposed however, all these methods have some major limitation's to detect the BC in its early stages. Thus, the intelligent analysis of clinical data including machine learning methods which are effective approaches for the detection of BC. However, there are various factors to analyze for diagnosis of BC and this complicates the job of the clinical doctors. The medical data and expert decision system to detect the BC are the most important factors in the diagnosis of BC. The review of the literature of the proposed breast cancer techniques are important for understanding the significance of our method. All these prior proposed methods used different methods to diagnosis the BC. Though, all these approaches have a low prediction accuracy and more execution time. The prediction accuracy of the BC identification technique needs more enhancements for efficient and accurate detection at early stages for better treatment and recovery. Thus, the key problems in these current methods are low accuracy and high computation time and these might be due to the use of non-suitable features in the data set. To tackle these issues new approaches are required to detect BC properly. The improvement in prediction accuracy of the ML model is a big challenge and research gap.

From the literature, we reached on the conclusion that BC diagnosis methods need further improvement that detect the BC effectively at initial stage for proper treatment and recovery of patient possible. In order to tackle the early stage detection of BC, in this research study, we have been proposed ML based identification method for breast cancer. In the proposed method three feature selection methods such as Relief, Autocoder and principal components analysis(PCA) have been used for appropriate features selection. The machine learning algorithms required suitable data for training and testing. The performance of machine learning model can be improved if balanced dataset is use for training and testing of the model. Additionally, the model performance can be increased by employing appropriate and related features from the data. Hence, data balancing and feature selection is significantly important for model better performance. To increase the predictive capability of ML models data pre-processing is necessary for data standardization and normalization. Various Preprocessing techniques, such as removal of missing feature value instances from the dataset, Standard Scalar, Min-Max Scalar are necessary for data preprocessing. The feature extraction and selection techniques are also improve model performance. In [7] described various methods for various kinds of feature selection, such as feature selection for High dimensional small instances size data, Large scale data, and secure feature selection. They also discussed some important topics for feature selection have emerged, such as stable feature selection, multi-view feature selection, distributed feature selection, multi-label feature selection, online feature selection, and adversarial feature selection. Due to these reasons we used pre-processing and feature selection techniques in the proposed method. The classifier SVM has been used for classification of BC and healthy people. The classification performance of SVM is more high and for problems of classification are mostly used [6], [8], [9]. Due to high performance and very efficient SVM, This paper is utilizing SVM approach over clinical data sets we consider it in this work. Two breast cancer data sets have been used for testing of the proposed system. Further K-fold cross validation method has been applied for validation of the proposed method and performance evaluation metrics have been used for model performance evaluation. McNemar's statistical test has been employed for proposed models performance comparison. In addition, the proposed work performance has been compared with existing state of the art methods.

This work has the following major contributions.

- Important features have been selected by using supervised learning (Relief algorithm) and unsupervised learning (Autoencoder and PCA algorithms) for effective identification of BC.
- Identified weak features in the data sets that have low impact in detection of BC.
- Relief integration with SVM is suitable method for identification of breast cancer.
- The proposed method has been checked on two breast cancer data sets.

The rest of the paper is organized as follow: The literature review has been presented in section2. Materials and methods have been discussed in detail in Section 3. The carried experiments and results analysis are reported with briefly comparison in the section 4. In section 5 conclusion and future work have been reported.

II. LITERATURE REVIEW

To identify breast cancer different machine learning methods have been proposed by various researchers. In this work we have been discussed some of the state of the art breast cancer diagnosis methods. The main purpose of literature review to identify the problems in existing methods and provide a reliable solution. Azar *et al.* [4] for identification of BC proposed a method. ML algorithms, Radial-Basis-Function (RBF), Probabilistic-Neural-Network (PNN) and Multi-Layer-Perception (MLP) have been used

TABLE 1. Summary of the baseline methods in the literature.

Ref	Model	Feature selection	Data set	Evaluation metrics	Accuracy%
[4]	RBF, PNN, MLP	-	-	Sensitivity, specificity, accuracy and ROC	97.80
[16]	SVM	k-means	WDBC	AUC, accuracy	97.38
[18]	PSO-KDE	PSO	WBCD	Sensitivity, specificity, accuracy	98.45
[20]	DBN	-	-	accuracy	99.70
[21]	SVM	mRMR and chi square	WBC, WDBC	C, AUC	99.70
[22]	HBSVM-C	-	WBC	Accuarcy	99.1
[25]	SVM	RFS	WDBC	Sensitivity, specificity, accuracy	99.00
[24]	RBL-RBFNN	-	WBC, BCD, BCP, and WBCD	Accuracy	97.4, 98.4, 97.7, 97.0
[23]	ML and BOADICEA	-	-	AU-ROC	-
[12]	SRMGP	-	WBC	Accuracy	99.00
[13]	ML	-	WBC	Accuracy	98.8
[14]	Fuzzy GA system	-	WBCD	Accuracy	97.36
[15]	SVM	F1-measure	WBCD	Sensitivity, specificity, accuracy, Positive predictive value, Negative predictive value, ROC	99.51

for identification of BC. Theses classifiers obtained high accuracy. In [10], the authors proposed a BC prediction system by using Genetic Algorithm for FS, and Rotation-Forest for identification of BC. The 99% accuracy obtained by Rotation-Forest on selected features. In [11] recommended a BC diagnosis method (GAMOO-NN). The performance of the proposed method is good in term of accuracy. In [12] the authors, designed a system for the analysis of BC utilizing Symbolic-Regression of Multigene-Genetic-Programming (SRMGP). The ten-folds validation has been used and achieved 99% accuracy. In [13] proposed a technique to diagnosis breast cancer and achieved 98.8% accuracy. Another study [14] authors suggested a method based on Fuzzy GA and attained accuracy 97.36%. Similarly, in [15] designed a BC method using the F1-measure procedure for feature selection and SVM for classification of BC. Zheng *et al.* [16] designed diagnosis method of BC using K-means and SVM. The K-mean has been used for feature extraction and SVM for classification. In [17] author, suggested an smart method for BC diagnosis. Fuzzy rough set was used for an instance selection, and FS by consistency. Fuzzy-Rough-Nearest-Neighbor Algorithm (FRNNA) was to detect BC. In [18] considered a system used Particle-Swarm-Optimization (PSO) combined with non-parametric kernel density for BC diagnosis. In [19] considered a BC identification method using Mixture Ensemble (ME) of Conventional Neural Networks (CNN). In [20] authors, recommended a system of BC by applying Deep-Belief-Networks (DBN) and attained 99.70% accuracy. In another study [21] authors proposed an integrated intelligent BC identification and in the proposed method they have been used FS selection algorithm for suitable features selection. Classifier SVM has been used for classification of malignant and benign subjects. Hold out method has been used for model validation and also used performance evaluation metrics for model performance evaluation. The proposed method achieved high performance in terms of accuracy. Osman *et al.* [22] proposed a breast tumour diagnosis method by employing hybrid SVM and two step clustering approach(HBSVM-C). To increase the accuracy of the predictive system of breast cancer diagnosis they employed hybrid approach. The proposed system has been tested on WBC data set and the predictive accuracy of the proposed method reached to 99.1%. Ming *et al.* [23] proposed a breast cancer diagnosis method by incorporating machine learning and BOADICEA model. The proposed method has been achieved performance in terms of AU-ROC 88.9%. Osman *et al.* [24] developed an effective of ensemble boosting learning approach for diagnosis of breast cancer virtual screening employing radial based function neural network models (RBFNN). They adapted 10 fold cross validation technique for best model selection and hyperparameters tuning. The proposed has been evaluated on breast cancer data sets. The proposed RBFNN method obtained 97.4%, 98.4%, 97.7% and 97.0% for the accuracy's on datsets WBC, BCD, BCP, and WBCD respectively.

The proposed methods in literature have been summarized in Table 1. In Table 1, we reported the proposed models, feature selection techniques, data sets, performance evaluation metrics and accuracy of these proposed methods for better understanding the existing literature of Breast cancer.

TABLE 2. Data sets description.

Repository	Name	Instances	Attributes	Developer	Class Distribution
UCI [1]	Wisconsin Diagnostic Breast Cancer (WBC) original	699	11	Wolberg et al.(University of Wisconsin)	444 benign and 239 malignant subjects
UCI [2]	Wisconsin Diagnostic Breast Cancer (WDBC)	569	32	Wolberg et al.(University of Wisconsin)	355 benign and 214 malignant subjects

TABLE 3. Data set WBC feature information.

Label	Feature Name	Code
F1	Simple code subject	S.No
F2	Clump Thickness	CT
F3	Uniformity of cell size	UCS
F4	Uniformity of cell shape	UCSH
F5	Marginal Adhesion	MA
F6	Single Epithelial Cell Size	SECS
F7	Bare Nuclei	BN
F8	Bland Chromatin	BC
F9	Normal Nucleoli	NN
F10	Mitoses	M

According to Table 1 the prediction accuracy of BC detection techniques need further improvement for efficient and accurate detection at early stages for better treatment and recovery. Thus, the major issues in these previous methods are low accuracy and high computation time and these might be due the use of irrelevant features in dataset. In order to tackle these problems new methods are needed to detect BC correctly. The improvement in prediction accuracy is a big challenge and research gap.

III. MATERIALS AND METHOD

The materials and method used in this research work are as follows.

A. DATA SET AND PRE-PROCESSING

In this study two breast cancer data sets have been used for our experimental work. Breast Cancer Wisconsin (Original) WBC dataset and Breast Cancer Wisconsin (Diagnostic)(WDBC) Data Set were designed by Wolberg *et al.* at University of Wisconsin and available on UCI data repository [26]. In Table 2 the data sets used in this work have been described. Further, the details features of both datasets have been given in Table 3 and 4 respectively.

The WBC dataset has samples sized of 699 and 11 attributes in which one is the code of instance, real values attributes are 9. Target output has two classes to demonstrated the malignant and benign subjects. The class distribution is 458 benign and 241 malignant subjects. 16 missing values instance have been removed, and thus remaining instances for two classes, are 444 benign and 239 malignant. Similarly WDBC data have 569 instance and 32 attributes with one output class label. There is no missing values instances in this data set. The two classes distribution of WDBC data are 355 benign and 214 malignant. The classes distribution of both data sets have been shown Figure 1.

B. PROPOSED FEATURE SELECTION ALGORITHMS

Feature selection (FS) is necessary step in machine learning process and due to appropriate feature selection the machine learning (ML) model performance increases and computational time of model decrease [27]. Feature selection process has great implication on classification results of the model [28]. The selection of suitable feature selection algorithms is a complicated process for selection of more appropriate feature from data set. In the literature different feature selection algorithms have been proposed for appropriate feature selection such as Genetic Algorithm [10], PCA [29], PSO [18], FRNNA [17], k-mean [16], Chi square [21], Mrmr [21]. In order to tackle the problem of feature selection in this study, we proposed two algorithms Supervised (Relief) algorithm and Unsupervised (Auto-Encoder and PCA) algorithms for appropriate feature selection because to date, researchers have studied the two types of feature selection algorithms separately. Supervised feature selection determines feature relevance by evaluating features correlation with the class, and without labels, unsupervised feature selection exploits data variance and separability to evaluate feature relevance [30]. The theoretical and mathematical background knowledge of these algorithms have been presented in below subsections.

1) SUPERVISED FEATURE SELECTION ALGORITHM

We have been used supervised learning based FS algorithm Relief for feature selection.

- Relief

Relief (RF) is supervised learning feature selection algorithm which uses filter mechanism for feature selection from data set. The theoretical and mathematical knowledge of RF algorithm has been presented for better understanding

TABLE 4. Data set WDBC feature information's.

Label	Feature Name	Code
F1	Id number	Integer
F2	Radius mean	Mean of distances from the center to points on the perimeter cell
F3	Texture mean	The standard deviation of gray-scale values
F4	perimeter mean	Perimeter of cell
F5	Area mean	Area of cell
F6	Smoothness mean	local variation in radius lengths
F7	Compactness mean	$\frac{Perimeter^2}{area - 1.0}$
F8	Concavity mean	The severity of concave portions of the contour
F9	Concave points mean	Number of concave portions of the contour
F10	Symmetry mean	Symmetry
F11	Fractal dimension mean	Coastline approximation, - 1
F12	Radius severity	-
F13	Texture severity	-
F14	Perimeter severity	-
F15	Area severity	-
F16	Smoothness severity	-
F17	Compactness severity	-
F18	Concavity severity	-
F19	Concave points severity	-
F20	Symmetry severity	-
F21	Fractal dimension severity	-
F22	Radius worst	-
F23	Texture worst	-
F24	Perimeter worst	-
F25	Area worst	-
F26	Smoothness worst	-
F27	Compactness worst	-
F28	Concavity worst	-
F29	Concave points worst	-
F30	Symmetry worst	-
F31	Fractal dimension worst	-

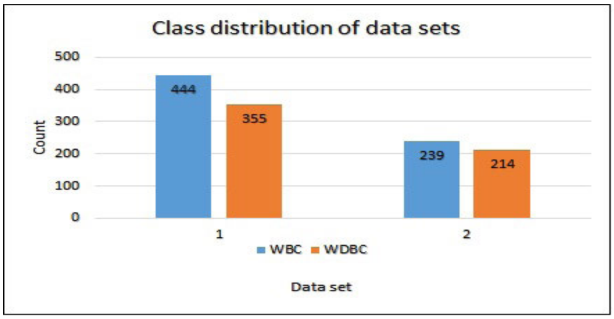


FIGURE 1. Class distribution of data sets WBC and WDBC.

of the algorithm. Relief is functionally distance based filter FS algorithm which ranks features that differentiate classes based on how to create organize feature that can separate classes. Relief algorithm was designed by Kira and Rendell [31], which is two class filter feature normalization to [0, 1] algorithm. Initially each feature is assigned a zero weight. A Dimensional training examples R is selected randomly. The Euclidean distance is computed for remaining samples. Represent the nearest hit in the same class H, while the nearest miss in a distinguish class M. The suitable feature $R[A]$ would be able to isolate class values, it have a short distance to H and a high distance to M. Therefore, $W[A]$ is adjusted to reward high variables and penalize non appropriate ones. The last selection of variables is made by choosing those large $W[A]$. Different diff function would be utilized for discrete such as diff $(x, y) = 0$ if x and y have the equal class, 1 otherwise and feature values continue. E.g. diff $(x, y) = (x - y)^2$. The relief algorithm have two major advantages one that its computationally less expensive and second it more suitable big data set. The pseudo code of the supervised filter based Relief algorithm is given in Algorithm 1 and illustrated in Figure 2.

Algorithm 1 Feature Selection Relief Algorithm Pseudo Code

```
Input: $S$: Training data with labels Feature, Parameters
    required $m$: Training instance out all instances applied to
    updated wight $W$
Output: $W$: Feature weight
n ← Total instances used for training
d ← Features used
$W[A] \leftarrow 0.0$;
for $k \leftarrow 1$ To $m$ do
    Select randomly 'Target' instance $R_k$
    Compute hit of nearest $H$ and miss of nearest $M$
    for $A \leftarrow 1$ To $a$ do
        $W[A] \leftarrow W[A] - Diff(A, R_k, H)/m + diff(A, R_k, M)/m$
    end for
end for
Return $W$;    ▷ Features weight vector that compute features quality
```

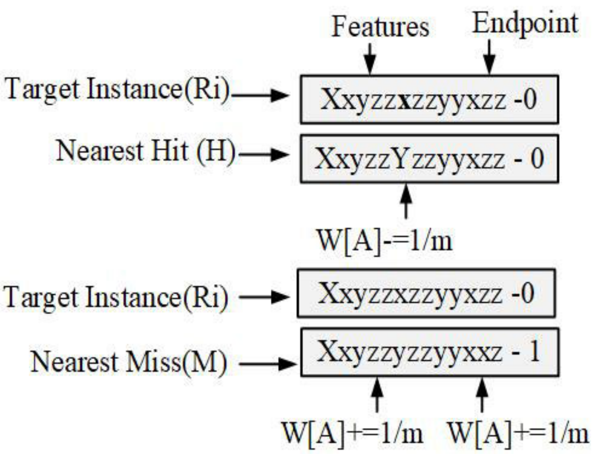


FIGURE 2. Feature selection process by supervised filter relief algorithm. [32].

2) UNSUPERVISED FEATURE SELECTION ALGORITHMS

We have been used two unsupervised FS algorithms i.e., Autoencoder and PCA for feature selection.

- Autoencoder based Feature Selection

The auto encoder is unsupervised learning model for extraction of useful feature from the original data set. The Generic diagram of Autoencoder has been shown in Figure 3. Let us consider that $X = \{x_1, x_2, \ldots, x_n\}^T \in R^{n \times d}$, is unlabeled sample matrix, where n is unlabeled samples with dimension (Features) d. In unsupervised selection of feature process to select subset $H(h \leq d)$ from X with unlabeled data that have more informative and discriminative features. The Auto encoder [33] is specific type of feedforward neural network(FFNN) which accept a features set as a input and generate output after applying different transforms. We consider a two fully connected layers autoencoder network as proposed in [33]. The simple autoencoder network with a h-dimension hidden layer consist of two parts such as an encoder function $f(X) = \sigma_1(XW^{(1)})$, and a decoder that perform the function of reconstruction $\hat{X} = g(f(x)) = \sigma_2(f(X)W^{(2)})$, Where σ_1, σ_2 are activation functions of the hidden layer and output layer. The activation functions such as sigmoid, ReLU, tanh and it can linear or non-linear ones are use with hidden and output layer. While weight parameters are $\Theta = \{W^{(1)}, W^{(2)}\}$ and W_{ij}^{l} represents the connection parameter between i-th neuron in the l-th layer and j-th neuron in the (l+1)-th layer.

The autoencoder overall function can be written as g(f(X)). The autoencoder learning process the loss function is represented in equation 1.

$$\tau(\Theta) = \frac{1}{2n}||X - g(f(X))||_F^2 \tag{1}$$

In this equation 1 n is samples, and $||.||_F$ is Frobenius norm for matrices. After the optimization of equation 1 the autoencoder compresses matrix X as reduced dimensional data $f(X)$. The output of decoded matrix given as $\hat{X}$ =g(f(X)).

- Sigmoid activation function:
  In our proposed autoencoder based feature selection algorithm, we use sigmoid activation function. Sigmoid is one of the activation function that mostly used for non-linear activation function. It output values exist in range of between 0 and 1. Thus, anything exists between 0 and 1 it is easy for probability detection. Since sigmoid is the good selection for binary classification problems. Mathematically sigmoid function can be written in equation 2 and graphically shown in Figure 4.

$$y = f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

- Optimizer Stochastic gradient descent (SGD):
  SGD is the mostly popular optimizer for machine learning and deep learning models. In the proposed autoencoder feature selection algorithm SGD has been used for optimization purpose. The architecture of the autoencoder for WBC and WDBC data sets have been shown in Figure 12 and 12 appendix section.

The following is the pseudo code of the Unsupervised autoencoder based feature selection Algorithm 2.

- Principal components analysis (PCA) based Feature Selection

The PCA [34] is a feature extraction and dimensions reduction algorithm. PCA constructs appropriate features by linearly transforming correlated features into a small number of uncorrelated features also called as principal components [35]. The constructed principal components are necessarily linear combinations of the actual data capturing most of the variance in the data. PCA have two major advantages to dimensionality reduction in clinical data machine learning studies. First, PCA is easily implemented and computationally fast. Secondly, un-supervised techniques does not require corresponding categorical labels to extract relevant features.

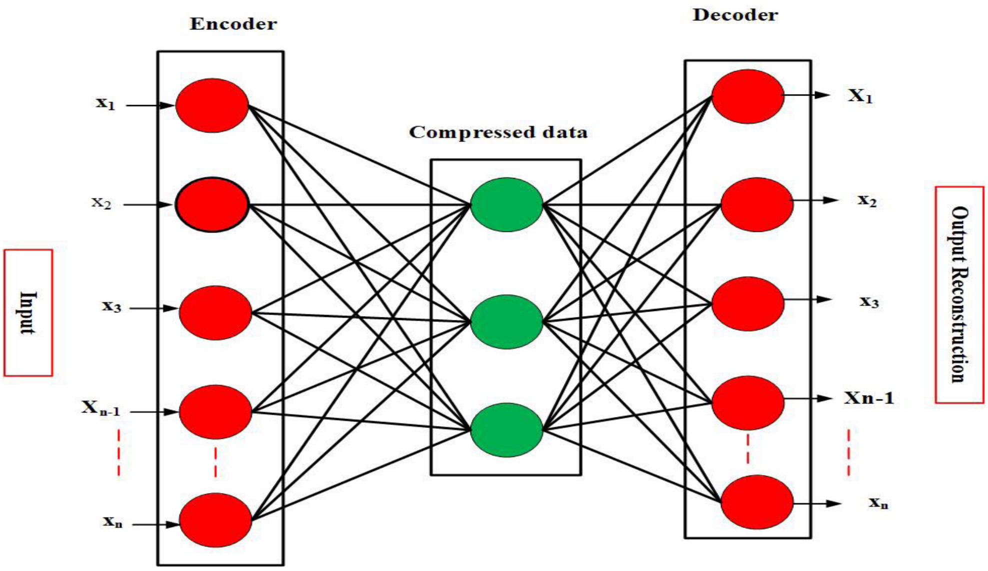


FIGURE 3. Feature selection process by Unsupervised Autoencoder.

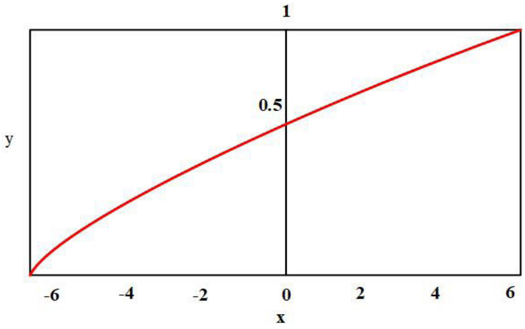


FIGURE 4. Sigmoid activation function.

```
Algorithm 2 Unsupervised Autoencoder Based Feature Selection
1: Begin
2: Input original unlabeled data as input to autoencoder which is unlabeled sample matrix, i.e $X = \{x_1, x_2, \ldots, x_n\}^T \in R^{n\times d}$;
3: Encoder function performed the encoding of features i.e $f(X) = \sigma_1(XW^{(1)})$;
4: Produced reduced features set after serious of transforms;
5: Decoder that perform the function of reconstruction of feature i.e $\hat{X} = g(f(x)) = \sigma_2(f(X)W^{(2)})$;
6: The output of decoder is equal to original features set;
7: End
```

C. CLASSIFICATION USING SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised classification algorithm [36], [37]. Because of the good results of SVM in classification, it is mostly used for various classification applications [6], [8], [9]. In the case of binary classification, the instances are divided by a hyperplane $w^Tx + b = 0$, where w and d-Dimensional coefficient vector, that is common for the surface hyperplane and b, are offset from the origin, x is data set values. The SVM receives w and b results. The w can solve in the linear case by adding Lagrangian multipliers. The w solution can be expressed as $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, where n is the number of vectors supported, y_i is the target output labels to x. The value of w and b is computed, as in Equation 3 the linear discriminating function can be written in Equation 3.

$$g(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b) \tag{3}$$

The nonlinear scenario can be written as in Equation 4 for kernel trick and decision function.

$$g(x) = sgn(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b) \tag{4}$$

D. CROSS VALIDATION METHOD

K-fold validation method has been used for the training and testing of the proposed method. In k folds validation we use k=5, in which k-1 using for training of the model and k-4 for validation of the model. The average values of 5 folds validation computed. The 5-folds CV method performance for our model is good because the numbers of instances in both data sets are small. So instead of 10 folds CV method we incorporated 5 folds method.

E. PERFORMANCE EVALUATION METRICS

The performance evaluation metrics [25], [38], [39] are use for performance evaluation of the model such as accuracy, specificity, sensitivity, F1-score, MCC, ROC and AUC. These metrics are described mathematically in equation 5-10 respectively. Where TP (true positive), TN (true negative), FP (false positive), FN (false negative).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100 \tag{5}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100 \tag{6}$$

$$Specificity = \frac{TN}{(TN + FP)} \times 100 \tag{7}$$

$$Precision = \frac{TP}{(TP + FP)} \times 100 \tag{8}$$

$$F1 - score = 2\frac{Precision \times Recall}{(Precision + Recall)} \times 100 \tag{9}$$

$$MCC = \frac{T_1}{\sqrt{T_2 \times T_3 \times T_4 \times T_5}} \times 100 \tag{10}$$

Here MCC is Matthews correlation coefficient, $T_1 = (TP \times TN - FP \times FN)$, $T_2 = (TP + FP)$, $T_3 = (TP + FN)$, $T_4 = (TN + FP)$, and $T_5 = (TN + FN)$

ROC-AUC: AUC illustrates the ROC of the classifier and high value of AUC represent high performance results of the classifier.

F. MC-NEMAR'S STATISTICAL TEST

The statistical tests are important for performance comparison of machine learning models. Thus, we employed McNemar's test [21] to compare the proposed method performance and other methods of breast cancer. To employ McNemar's test, the instances of dataset S have been divided into a training set R and testing set T. We train models with training data and test on test dataset. For each sample x $\in$ T of the test set we compute how it get classified by two models. The test is used to a 2 × 2 contingency Table, that tabulate the output of two tests on a sample of n subjects. The total number of samples in the test set are n expressed mathematically as n $= \mathrm{n}_{00} + \mathrm{n}_{01} + \mathrm{n}_{10} + \mathrm{n}_{11}$. Hypothesis of two tails under the null hypothesis, the two models should have equal accuracy which expressed as mathematically H_0: $\mathrm{n}_{01} = \mathrm{n}_{10}$. While he alternate hypothesis, the two models have accuracy different which can be expressed mathematically as H_1: $\mathrm{n}_{01} \neq \mathrm{n}_{10}$. In equation 11 McNemar's test computed.

$$P - value = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}} \tag{11}$$

The significance selection level, the test statistic or p-value illustrated as, the test statistic is chi-square distribution with freedom of degree 1. In addition the confidence level and α are complement of each other. The significant level is alpha, if alpha value is small then high confidence level and the significance of the model will high. While if the alpha value is large then confidence level will be small and the model is less significant. Mathematically we write it as bellow: If p > α: then H_0 is failed to reject, the models are not difference, If p $\leq \alpha$: then H_0 is rejected and alternate H_1 is accepted the models have performance different when trained on the specific training datsset R.

G. PROPOSED CLASSIFICATION METHOD

The proposed method has been designed to identify the Breast cancer. In this method, the classifier SVM has been used for prediction of BC. The Relief, PCA and autoencoder algorithms have been used for features selection that classifier effectively classifies breast cancer and healthy people. Additionally, the k-fold cross-validation method has been used for best hyper-parameters and for predictive model selection. Performance measuring metrics have been used for model performance evaluation. The Breast cancer clinical data sets have been used for testing of the proposed method. McNemar's statistical test has been incorporated for proposed models comparison. The following is the pseudo code of the proposed method which is given in Algorithm 3. The flow chart of the proposed method has been shown in Figure 5.

Algorithm 3 Proposed BC Identification Method

```
1: Begin
2: Pre-processing of clinical BC data sets using min-max scalar;
3: Supervised based Relief algorithm and unsupervised based autoencoder and PCA algorithms have been used for appropriate feature selection;
4: Training the classifier with k-1 instances of the data set; Validate with k-5 instances of the data set;
5: Train model with k-1sub-groups with initial hyper parameters values (C, $\gamma$);
6: Validate the model on test set of 5 folds and obtained the best hyper parameters; Repeat steps 4 and 5
7: Calculated model average classification results of 5 folds CV;
8: Performance of best model on testing set;
9: End
```

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments have been conducted in this section to check the classification performance of the proposed method. The "Wisconsin Diagnostic Breast Cancer (WBC) original" and Wisconsin Breast Cancer Diagnostic (WDBC) data sets have been used for testing of the proposed method. The k-fold were k = 5 has been used for validation of the proposed method. The classifier SVM performance have been evaluated on full features set. Supervised learning based FS algorithm Relief and unsupervised learning autoencoder and PCA algorithms have been used for feature selection and on these selected features the classifier SVM performance has been evaluated. In addition, the classifier has been trained with essential hyper parameters values. Furthermore, Performance evaluation metrics have been used to check the performance of classifier such as accuracy, specificity, sensitivity, MCC, ROC-AUC. Before applying to classifier, all the features were standardized and normalized. Additionally, McNemar's statistical test has been incorporated for the proposed models comparison. All the experimental results have been reported in different tables and graphically have been demonstrated with various graphics.

For experimental setup computer configure with Intel(R) Core^{TM} i3-2400 CPU @3.10 GHz PC with window XP 10 has been used. Different machine leaning libraries

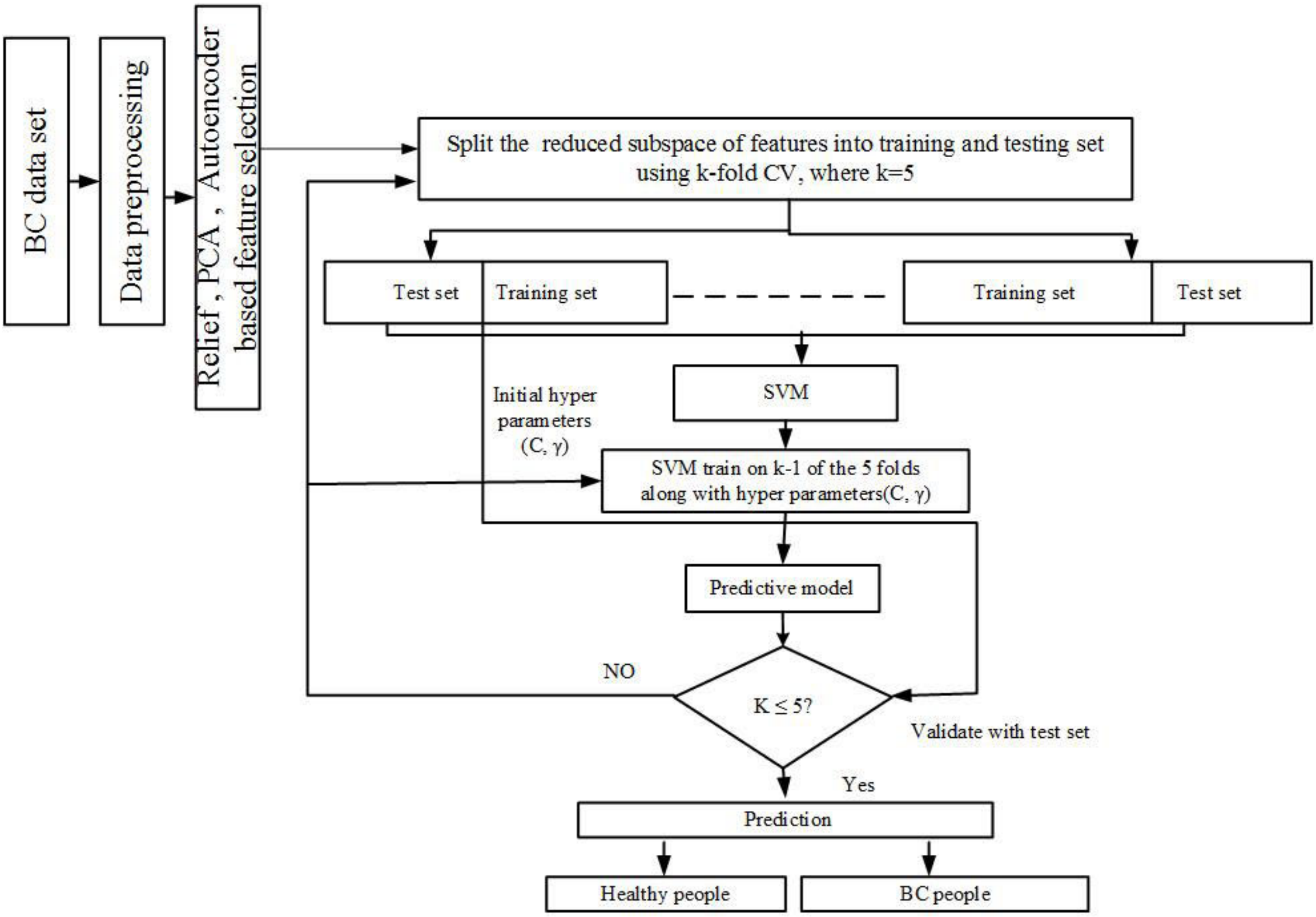

BC data set
Data preprocessing
Relief , PCA , Autoencoder based feature selection
Split the reduced subspace of features into training and testing set using k-fold CV, where k=5
Test set
Training set
Training set
Test set
SVM
Initial hyper parameters (C, γ)
SVM train on k-1 of the 5 folds along with hyper parameters(C, γ)
Predictive model
NO
K ≤ 5?
Validate with test set
Yes
Prediction
Healthy people
BC people

FIGURE 5. Flow chart of the proposed BC method.

have been configured on python programming language for simulation.

Furthermore, all required hyper-parameters of the concern models have been reported with related values in different experimental subsections of this section.

A. RESULT OF SUPERVISED RELIEF FEATURE SELECTION ALGORITHM

For important feature selection from WBC dataset Relief algorithm has been used. Relief algorithm assign weight of all the features and selected those features whose weight value is high, it means high weighted features selected by relief and low weight features are removed from the data set. The features selected by relief algorithm have been given in Table 5. According to relief Algorithm 1, these are important features from the data set and these features have great contribution in detection of breast cancer. Similarly from Wisconsin Diagnostic Breast Cancer (WDBC) data set features selected by Relief have been reported in Table 6.

B. RESULT OF UNSUPERVISED AUTOENCODER AND PCA FEATURE SELECTION ALGORITHMS

The unsupervised based autoencoder feature selection algorithm has been selected important features from WBC dataset which have been reported in the Table 5. According to Autoencoder FS algorithm these features have significant contribution in the detection of breast cancer. On other hand feature selected by Autoencode from WDBC data set have been reported in the Table 6. Similarly feature selected by PCA from WBC data set also reported in the Table 5 and from Wisconsin Diagnostic Breast Cancer (WDBC) data set features selected by PCA have been reported in the Table 6.

C. CLASSIFICATION PERFORMANCE OF CLASSIFIER SVM ON FULL AND ON SELECTED FEATURE SELECTED FROM WBC DATA SET BY RELIEF ALGORITHM

The classification performance of SVM has been checked on full and on selected features set by relief for prediction of breast cancer. The SVM different kernels, such as RBF and Linear with hyper parameters values of $C = 1$ and $\gamma = 0.002$ have been used in these experiments for prediction of breast cancer. The classification of SVM on full features set and on selected features set have been tabulated in Table 7. Thus, according to Table 7, SVM linear performance on full features have been achieved 97.22% accuracy, 95% specificity, 89% sensitivity, 97% F1-measure, 98% AUC and 0.037 seconds processing time. On other hand SVM linear with hyper-parameter $C = 1$ and $\gamma = 0.002$ trained and tested

TABLE 5. Features selected from Wisconsin Diagnostic Breast Cancer (WBC) original data set by Relief, PCA and Autoencoder algorithms.

FS algorithm	Feature Name	Feature Code
Relief	Clump Thickness Uniformity of Cell Size Uniformity of Cell Shape Marginal Adhesion Single Epithelial Cell Size Bare Nuclei Bland Chromatin Normal Nucleoli Mitoses	CT UCS UCSH MA SECS BN BC NN M
PCA	Clump Thickness Uniformity of Cell Size Uniformity of Cell Shape Single Epithelial Cell Size Bland Chromatin Normal Nucleoli	CT UCS UCSH SECS BC NN
Autoencoder	Uniformity of Cell Size Uniformity of Cell Shape Marginal Adhesion Single Epithelial Cell Size Bland Chromatin Normal Nucleoli Mitoses	UCS UCSH MA SECS BC NN M

with selected features set and obtained 99.91% accuracy, 99% specificity, 100% sensitivity, 88% MCC, 99% F1-measure, 99% AUC and 0.002 second was model processing.

While the classification of SVM RBF with hyper parameters C = 1 and $\gamma = 0.002$ on full feature set also reported in Table 7. According to Table 7 the SVM (RBF) obtained 97.22% accuracy, 88% specificity,99% sensitivity, 97% MCC, 98% F1-score and 0.048 seconds was processing of the model. Similarly on selected feature set with same hypermeters SVM(RBF) achieved 99.75% accuracy, 89% specificity, 87% sensitivity, 98% MCC, 99% F1-measure and 0.023 seconds was processing time of the model. Table 7 demonstrated the performance of SVM(Linear) has high as compared to SVM(RBF) on selected features set. The high performance of SVM linear on selected features might be due to the data set in linear. The SVM linear obtained 99.91% accuracy on selected features set. The high performance due

TABLE 6. Features selected from Wisconsin Diagnostic Breast Cancer (WDBC) data set by Relief, PCA and Autoencoder algorithms.

FS algorithm	Lebel	Feature Name
Relief	F3	Texture mean
	F5	Area mean
	F6	Smoothness mean
	F7	Compactness mean
	F8	Concavity mean
	F9	Concave points mean
	F11	Fractal dimension mean
	F12	Radius severity
	F13	Texture severity
	F15	Area severity
	F17	Compactness severity
	F19	Concave points severity
	F20	Symmetry severity
	F22	Radius worst
	F23	Texture worst
	F24	Perimeter worst
	F25	Area worst
	F26	Smoothness worst
	F29	Concave points worst
	F30	Symmetry worst
PCA	F3	Texture mean
	F4	perimeter mean
	F5	Area mean
	F8	Concavity mean
	F9	Concave points mean
	F11	Fractal dimension mean
	F12	Radius severity
	F14	Perimeter severity
	F16	Smoothness severity
	F17	Compactness severity
	F19	Concave points severity
	F21	Fractal dimension severity
	F25	Area worst
	F26	Smoothness worst
	F27	Compactness worst
	F28	Concavity worst
	F29	Concave points worst
	F31	Fractal dimension worst
Autoencoder	F3	Texture mean
	F4	perimeter mean
	F5	Area mean
	F9	Concave points mean
	F11	Fractal dimension mean
	F12	Radius severity
	F14	Perimeter severity
	F16	Smoothness severity
	F17	Compactness severity
	F19	Concave points severity
	F21	Fractal dimension severity
	F26	Smoothness worst
	F27	Compactness worst
	F29	Concave points worst
	F31	Fractal dimension worst

TABLE 7. Classification results on full and on selected features set from WBC data set by Relief.

Classifier	Parameters	Feature set	Performance evaluation metrics						
	(C, γ)		Acc	Sp	Sn	MCC	F1-score	ROC-AUC	Time(s)
SVM(Linear)	1, 0.002	Full	97.22	95	89	87	97	98	0.037
	1, 0.002	Selected	99.91	99	100	88	99	99	0.002
SVM(RBF)	1, 0.002	Full	97.22	88	99	98	97	98	0.048
	1, 0.002	Selected	99.75	96	89	87	98	99	0.023

TABLE 8. Classification results on full and on selected features set from WBC data set by Autoencoder.

Classifier	Parameters	Feature set	Performance evaluation metrics						
	(C, γ)		Acc	Sp	Sn	MCC	F1-score	ROC-AUC	Time(s)
SVM(Linear)	1, 0.002	Full	97.22	95	89	87	97	98	0.037
	1, 0.002	Selected	99.01	98	87	88	89.00	99	0.001
SVM(RBF)	1, 0.002	Full	97.22	88	99	98	97	98	0.048
	1, 0.002	Selected	98.75	98	79	81	96	99	0.001

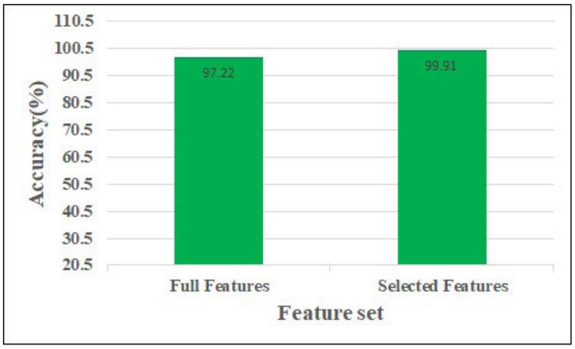


FIGURE 6. Classification accuracy of SVM on full and on selected features from WBC data set by Relief.

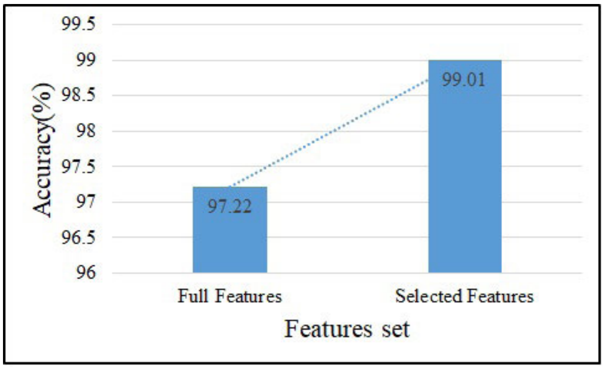


FIGURE 7. Classification accuracy of SVM with Autoencoder on WBC data set.

to the most related features selection by relief FS algorithm. The classification performance of SVM on full and selected features set has been shown in Figure 6.

D. CLASSIFICATION PERFORMANCE OF CLASSIFIER SVM ON FULL AND ON SELECTED FEATURE SELECTED FROM WBC DATA SET BY AUTOENCODER ALGORITHM

The classification performance of SVM has been checked on full and on selected features set by autoencoder for prediction of breast cancer. The SVM different kernels, such as RBF and Linear with hyper parameters values of C = 1 and γ = 0.002 have been used in these experiments for prediction of breast cancer. The classification of SVM on full features set and on selected features set by autoencoder FS algorithm have been tabulated in Table 8. Thus, according to Table 8, SVM linear performance on full features achieved 97.22% accuracy, 95% specificity, 89% sensitivity, 97% F1-measure, 98% AUC and 0.037 seconds processing time. On other hand SVM linear with hyper-parameter C = 1 and γ = 0.002 trained and tested with selected features set selected by autoencoder FS algorithm and obtained 99.01% accuracy, 98% specificity, 87% sensitivity, 89% MCC, 99% F1-measure, 98% AUC and 0.001 second was model processing.

While the classification of SVM RBF with hyper parameters C = 1 and γ = 0.002 on full feature set also reported in Table 7. According to Table 8 the SVM (RBF) obtained 97.22% accuracy, 88% specificity, 99% sensitivity, 97% MCC, 98% F1-score and 0.048 seconds was processing of the model. Similarly on selected feature set selected by autoencoder FS algorithm with same hyper-meters SVM(RBF) achieved 98.75% accuracy, 79% specificity, 81% sensitivity, 96% MCC, 99% F1-measure and 0.001 seconds was processing time of the model. Table 8 demonstrated the performance of SVM(Linear) has high as compared to SVM(RBF) on selected features set. The high performance of SVM linear on selected features might be due to the data set in linear. The SVM linear obtained 99.01% accuracy on selected features set. The high performance due to the most related features selection by autoencoder FS algorithm. The classification performance of SVM on full and selected features set have been shown in Figure 7.

TABLE 9. Classification results on full and on selected features set from WBC data set by PCA.

Classifier	Parameters	Feature set	Performance evaluation metrics						
	(C, γ)		Acc	Sp	Sn	MCC	F1-score	ROC-AUC	Time(s)
SVM(Linear)	1, 0.002	Full	89.00	99	80	83	89	90	0.027
	1, 0.002	Selected	98.33	98	88	98	80	97	0.011
SVM(RBF)	1, 0.002	Full	98.00	98	97	98	97	98	0.038
	1, 0.002	Selected	98.01	99	79	87	86	99	0.011

E. CLASSIFICATION PERFORMANCE OF CLASSIFIER SVM ON FULL AND ON SELECTED FEATURE SELECTED FROM WBC DATA SET BY PCA ALGORITHM

The classification performance of SVM has been checked on full and on selected features set by PCA for prediction of breast cancer. The SVM different kernels, such as RBF and Linear with hyper parameters values of C = 1 and $\gamma = 0.002$ have been used in these experiments for prediction of breast cancer. The classification of SVM on full features set and on selected features set by PCA FS algorithm have been tabulated in Table 9. Thus, according to Table 9, SVM linear performance on full features achieved 89% accuracy, 99% specificity, 80% sensitivity, 83% F1-measure, 90% AUC and 0.037 seconds processing time. On other hand SVM linear with hyper-parameter C = 1 and $\gamma = 0.002$ trained and tested with selected features set selected by PCA FS algorithm and obtained 98.44% accuracy, 98% specificity, 88% sensitivity, 98% MCC, 80% F1-measure, 97% AUC and 0.011 second was model processing.

While the classification of SVM RBF with hyper parameters C = 1 and $\gamma = 0.002$ on full feature set also reported in Table 7. According to Table 9 the SVM (RBF) obtained 98% accuracy, 98% specificity, 97% sensitivity, 98% MCC, 97% F1-score,98% AUC and 0.038 seconds was processing of the model. Similarly on selected feature set selected by PCA FS algorithm with same hypermeters SVM(RBF) achieved 98.01% accuracy, 99% specificity, 87% sensitivity, 86% MCC, 99% F1-measure and 0.011 seconds was processing time of the model. Table 9 demonstrated the performance of SVM(Linear) has high as compared to SVM(RBF) on selected features set. The high performance of SVM linear on selected features might be due to the data set in linear. The SVM linear obtained 98.45% accuracy on selected features set. The high performance due to the most related features selection by PCA FS algorithm.

The classification performance of SVM on features selected by Relief algorithm comparatively high to the features selected by autoencoder and PCA FS algorithm. The classification performance of SVM on Relief based selected features from WBC data set are 99.91% accuracy, 89% specificity, 87% sensitivity, 98% MCC, 99% F1-measure and 0.023 seconds is processing time of the model, while the performance of SVM on features selected by autoencoder from WBC data set are 99.01% accuracy, 98% specificity, 87% sensitivity, 89% MCC, 99% F1-measure, 98% AUC and 0.001 second is model processing. On other hand SVM obtained 98.45% accuracy on selected features from WBC data set by PCA FS algorithm. Thus, the breast cancer diagnosis system based of Relief and SVM is more suitable for accurate and efficient of detection of BC when using WBC data set. The classification performance of SVM on features selected from WBC data set by Relief, PCA and Autoencoder algorithms has been shown in Figure 8. Thus, we reached on the conclusion that the performance of Relief-SVM model on WBC data set is high as compared to WDBC data set and we recommend it for detection of breast cancer.

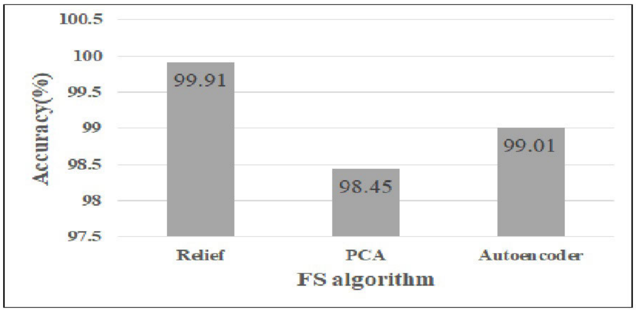


FIGURE 8. SVM performance on features selected from WBC data by Relief, PCA and Autoencode algorithms.

F. CLASSIFICATION PERFORMANCE OF CLASSIFIER SVM ON FULL AND ON SELECTED FEATURE SELECTED FROM WDBC DATA SET BY RELIEF, AUTENCODER AND PCA FS ALGORITHMS

In this section, we have been performed experiments for checking the classification performance of SVM using WDBC data set. The performance of model has been checked on full and on selected features sets selected by Relief, Autoencoder and PCA FS algorithms for prediction of breast cancer. The SVM different kernels, such as RBF and Linear with hyper parameters values of C = 1 and $\gamma = 0.003$ have been used in these experiments for effectively trained the classifier. The classification of SVM on full features set and on selected features sets selected by Relief, Autoencoder and PCA FS algorithms have been tabulated in Table 10. Thus, according to Table 10, SVM performance in terms of accuracy with Relief based features selection was 96.48%, while the accuracy of SVM linear with Autoencoder based features selection was 91.12%. Similarly PCA based selected features the SVM linear achieved 97.45% accuracy which is very high as compared to full features sets and other FS algorithms such as Relief and Autoencoder. The Accuracy of these three models have been shown in Figure 9 for better understanding the performance of these models.

TABLE 10. Classification results on full and on selected features set from WDBC data set by Relief, Autencoder, and PCA FS algorithms.

FS Algorithm	Classifier	Parameters	Feature set	Performance evaluation metrics						
		(C, γ)		Acc	Sp	Sn	MCC	F1-score	AUC	Time(s)
Relief	Linear	1, 0.003	Full	94.00	98	83	86	89	93	0.099
		1, 0.003	Selected	96.48	98	100	98	89	96	0.291
Relief	RBF	1, 0.003	Full	90.70	98	97	98	91	98	0.038
		1, 0.003	Selected	92.01	99	79	97	87	81	0.011
Auto	Linear	1, 0.003	Full	89.00	99	88	89	88	91	0.027
		1, 0.003	Selected	91.12	99	98	97	89	96	0.011
Auto	RBF	1, 0.003	Full	88.00	98	91	97	95	94	0.032
		1, 0.003	Selected	90.11	99	100	81	81	90	0.041
PCA	Linear	1, 0.003	Full	92.10	99	87	82	91	92	0.027
		1, 0.003	Selected	97.45	100	98	90	88	93	0.051
PCA	RBF	1, 0.003	Full	90.50	98	90	90	91	90	0.088
		1, 0.003	Selected	95.12	99	100	88	86	94	0.061

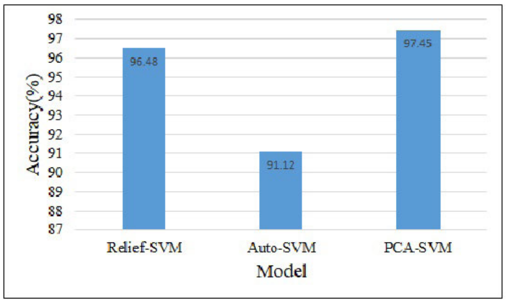


FIGURE 9. SVM performance on features selected from WDBC data by Relief, PCA and Autoencode algorithms.

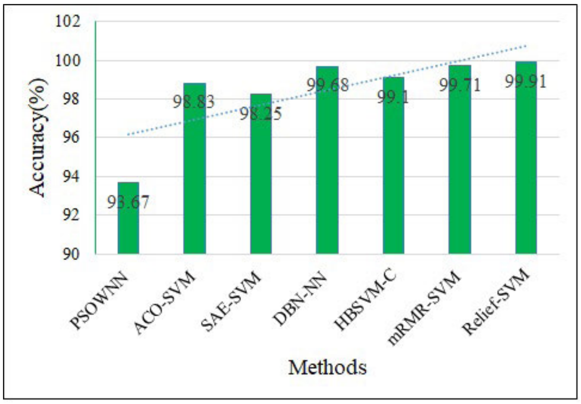


FIGURE 10. Performance comparison of our method with baseline methods.

G. MCNEMAR'S STATISTICAL TEST FOR THE MODELS PERFORMANCE COMPARISON

McNemar's test is employed, which is a well-known statistical test to compare our resulted performance among the machine learning models. We set the hypothesis for our experiments as H_0: $n_{01} = n_{10}$, if models performance are same accuracy. otherwise H_1: $n_{01} \neq n_{10}$, the alternate hypothesis, the two model accuracy are different. To test the null and alternate hypothesis p-value is computed for all models employing McNemar's test. For all experiments the value of alpha is 0.5, and confidence level is 95%. Hence on the basis of p-value and alpha, We consider accept or reject the null hypothesis on criteria as If $p - value > \alpha$: then H_0 fail to reject, the model's are same performance. If p-value $<= \alpha$: then H_0 is rejected and alternate H_1 is accepted. These models performance are different when trained on the particular training set R. The experimental results of p-value are calculated for all employed models and reported in Table 11 with level significant is 0.5.

TABLE 11. Models comparison by using McNemar's statistical test.

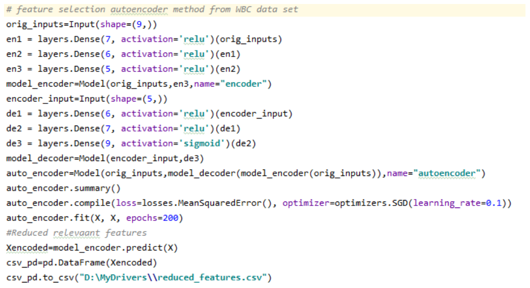
Data set	Method	Accuracy(%)	p-value
WBC	Relief-SVM	99.91	0.4
	Autoencoder-SVM	99.01	0.6
	PCA-SVM	98.33	0.7
WDBC	Relief-SVM	96.48	1.8
	Autoencoder-SVM	91.12	1.11
	PCA-SVM	97.45	1.0

TABLE 12. Performance comparison of proposed method with existing methods.

Reference	Method	Accuracy(%)
[40]	PSOWNN	93.67
[41]	ACO-SVM	98.83
[42]	SAE-SVM	98.3
[22]	HBSVM-C	99.1
[20]	DBN-NN	99.7
[21]	mRMR-SVM	99.71
Proposed method	Relief-SVM	99.91

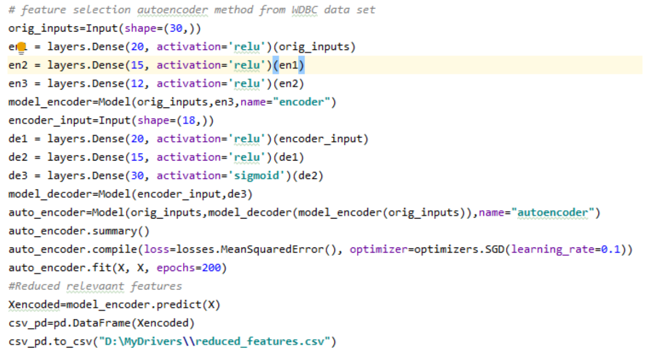
H. COMPARISON OF PROPOSED METHOD WITH BASELINE METHODS

The performance of proposed (Relief-SVM) in terms of accuracy have been compared with state of the art method in the Table 12 and graphically shown in the Figure 12 for better understanding. According to Table 12 and Figure 12 the proposed method has been achieved high accuracy 99.91% as compared to existing state of the art methods. The high performance of proposed method due to appropriate features selection of Relief FS algorithm and SVM predictive



```
# feature selection autoencoder method from WBC data set
orig_inputs=Input(shape=(9,))
en1 = layers.Dense(7, activation='relu')(orig_inputs)
en2 = layers.Dense(6, activation='relu')(en1)
en3 = layers.Dense(5, activation='relu')(en2)
model_encoder=Model(orig_inputs,en3,name="encoder")
encoder_input=Input(shape=(5,))
de1 = layers.Dense(6, activation='relu')(encoder_input)
de2 = layers.Dense(7, activation='relu')(de1)
de3 = layers.Dense(9, activation='sigmoid')(de2)
model_decoder=Model(encoder_input,de3)
auto_encoder=Model(orig_inputs,model_decoder(model_encoder(orig_inputs)),name="autoencoder")
auto_encoder.summary()
auto_encoder.compile(loss=losses.MeanSquaredError(), optimizer=optimizers.SGD(learning_rate=0.1))
auto_encoder.fit(X, X, epochs=200)
#Reduced relevaant features
Xencoded=model_encoder.predict(X)
csv_pd=pd.DataFrame(Xencoded)
csv_pd.to_csv("D:\MyDrivers\\reduced_features.csv")
```

FIGURE 11. Autoencoder FS algorithm architecture for wbc data set.



```
# feature selection autoencoder method from WDBC data set
orig_inputs=Input(shape=(30,))
en1 = layers.Dense(20, activation='relu')(orig_inputs)
en2 = layers.Dense(15, activation='relu')(en1)
en3 = layers.Dense(12, activation='relu')(en2)
model_encoder=Model(orig_inputs,en3,name="encoder")
encoder_input=Input(shape=(18,))
de1 = layers.Dense(20, activation='relu')(encoder_input)
de2 = layers.Dense(15, activation='relu')(de1)
de3 = layers.Dense(30, activation='sigmoid')(de2)
model_decoder=Model(encoder_input,de3)
auto_encoder=Model(orig_inputs,model_decoder(model_encoder(orig_inputs)),name="autoencoder")
auto_encoder.summary()
auto_encoder.compile(loss=losses.MeanSquaredError(), optimizer=optimizers.SGD(learning_rate=0.1))
auto_encoder.fit(X, X, epochs=200)
#Reduced relevaant features
Xencoded=model_encoder.predict(X)
csv_pd=pd.DataFrame(Xencoded)
csv_pd.to_csv("D:\MyDrivers\\reduced_features.csv")
```

FIGURE 12. Autoencoder FS algorithm architecture for wdbc data set.

model. The proposed method can be easily incorporated in e-healthcare systems for effective identification of BC.

V. CONCLUSION

Breast cancer is one of the highly dangerous disease among the females around the world. The efficient and correct detection of BC is big medical issue and many researchers proposed different diagnostic methods for detection of this disease, however these existing methods still needed further improvement to correct and efficient detection of this disease. In this study, we proposed a new BC identification method by using machine learning algorithms and clinical data. In the proposed method supervised (Relief) algorithm and unsupervised (Autoencoder and PCA) algorithms have been used for related features selection from data set and then these selected features have been used for the training and testing of the classifier SVM for accurate and on time detection of BC. Additionally in the proposed method k folds cross validation method has been used for model validation and best hyper parameters selection. The model performance evaluation metrics have been used for model performance evaluation. The BC data sets have been used for testing of the proposed method. The experimental results are demonstrated that the features selection take a deep significant in accurate and on time detection of BC. The proposed method has achieved high results in term of accuracy and achieved 99.91% accuracy on the feature selection by Relief FS algorithm. Further, the performance of SVM on features selected by autoencoder and PCA have low performance as compared to the performance of SVM on features selected by Relief algorithm. Thus, the proposed method Relief-Support vector machine is highly recommended for diagnosis of BC. The performance of the proposed method is high as compared to existing state of the art method in terms of accuracy. Additionally, we employed McNemar's statistical test for performance comparison of our models. The novelty of the proposed study, is to designed a BC diagnosis method using machine learning classification and feature selection techniques. Firstly, a suitable FS algorithm have been used for important features selection and classifier SVM achieved high accuracy on these selection features. Secondly, the weak features have been successfully separated from the data sets that have low impact on prediction of BC. Thirdly, the WBC data set is more suitable and classifier SVM achieved high performance as compared to WDBC data set. Lastly, the BC detection method based on Relief SVM is more suitable for the detection of BC. Further, the proposed method could be easily incorporated in healthcare system for diagnosis of BC. In future, we will use other features selection algorithms along with other data sets of BC for further improvement in BC detection. Additionally, deep learning models will also apply for detection BC.

TABLE 13. Mathematical symbols and notations are used in this paper.

Symbol	Description
H	Data set
S	Subset
F	Feature set
n	Number of instances in dataset
X	Input features in dataset
Y	Predicted output classes label
b	Bais is offset value from the origin
w	d-dimensional coefficient vector
i	i is ith sample in data set
x_i	ith instance of dataset sample X
x^{-x}	exponential function
y_i	Target labels to x
R	Training set
sgn	sgn is significant margin of hyperplane
T	Test set
t	Finite set
P-value	Test probability value
α	Degree of freedom
f	Feature in dataset
F_i	ith feature in dataset
ϕ	Empty sect
p	probability
H_0	Null hypothesis
H_1	Alternate hypothesis

APPENDIX

The mathematical Notations used in paper are given in Table 13.

ACKNOWLEDGMENT

(All authors contributed equally to this work.)

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AVAILABILITY OF DATA AND MATERIAL

The data set used in this study available on UCI machine learning repository.

REFERENCES

[1] *American Institute for Breast Cancer Research*. Accessed: Sep. 21, 2020. [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics

[2] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-nearest neighbors," in *Proc. IEEE Region Humanitarian Technol. Conf. (R-HTC)*, Dec. 2017, pp. 226–229.

[3] A. M. Ahmad, G. M. Khan, S. A. Mahmud, and J. F. Miller, "Breast cancer detection using Cartesian genetic programming evolved artificial neural networks," in *Proc. 14th Int. Conf. Genetic Evol. Comput. Conf. (GECCO)*, New York, NY, USA, 2012, pp. 1031–1038.

[4] A. T. Azar and S. A. El-Said, "Probabilistic neural network for breast cancer classification," *Neural Comput. Appl.*, vol. 23, no. 6, pp. 1737–1751, Nov. 2013.

[5] E. Warner, H. Messersmith, P. Causer, A. Eisen, R. Shumak, and D. Plewes, "Systematic review: Using magnetic resonance imaging to screen women at high risk for breast cancer," *Ann. Internal Med.*, vol. 148, no. 9, pp. 671–679, 2008.

[6] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Inf. Syst.*, vol. 2018, pp. 1–21, Dec. 2018.

[7] Y. Li, T. Li, and H. Liu, "Recent advances in feature selection and its applications," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 551–577, Dec. 2017.

[8] A. U. Haq, J. Li, M. H. Memon, J. Khan, S. U. Din, I. Ahad, R. Sun, and Z. Lai, "Comparative analysis of the classification performance of machine learning classifiers and deep neural network classifier for prediction of parkinson disease," in *Proc. 15th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP)*, Dec. 2018, pp. 101–106.

[9] A. U. Haq, J. P. Li, M. H. Memon, J. Khan, A. Malik, T. Ahmad, A. Ali, S. Nazir, I. Ahad, and M. Shahid, "Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings," *IEEE Access*, vol. 7, pp. 37718–37734, 2019.

[10] E. Alickovic and A. Subasi, "Breast cancer diagnosis using GA feature selection and rotation forest," *Neural Comput. Appl.*, vol. 28, no. 4, pp. 753–763, Apr. 2017.

[11] F. Ahmad, N. A. M. Isa, Z. Hussain, and S. N. Sulaiman, "A genetic algorithm-based multi-objective optimization of an artificial neural network classifier for breast cancer diagnosis," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1427–1435, Oct. 2013.

[12] M. K. Hasan, M. M. Islam, and M. M. A. Hashem, "Mathematical model development to detect breast cancer using multigene genetic programming," in *Proc. 5th Int. Conf. Informat., Electron. Vis. (ICIEV)*, May 2016, pp. 574–579.

[13] A. A. Albrecht, G. Lappas, S. A. Vinterbo, C. Wong, and L. Ohno-Machado, "Two applications of the LSA machine," in *Proc. 9th Int. Conf. Neural Inf. Process. (ICONIP)*, vol. 1, 2002, pp. 184–189.

[14] C. A. Peña-Reyes and M. Sipper, "A fuzzy-genetic approach to breast cancer diagnosis," *Artif. Intell. Med.*, vol. 17, no. 2, pp. 131–155, Oct. 1999.

[15] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, Mar. 2009.

[16] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1476–1482, Mar. 2014.

[17] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6844–6852, Nov. 2015.

[18] R. Sheikhpour, M. A. Sarram, and R. Sheikhpour, "Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer," *Appl. Soft Comput.*, vol. 40, pp. 113–131, Mar. 2016.

[19] R. Rasti, M. Teshnehlab, and S. L. Phung, "Breast cancer diagnosis in DCE-MRI using mixture ensemble of convolutional neural networks," *Pattern Recognit.*, vol. 72, pp. 381–390, Dec. 2017.

[20] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Syst. Appl.*, vol. 46, pp. 139–144, Mar. 2016.

[21] A. Ul Haq, J. Li, M. H. Memon, J. Khan, and S. Ud Din, "A novel integrated diagnosis method for breast cancer detection," *J. Intell. Fuzzy Syst.*, vol. 38, no. 2, pp. 2383–2398, 2020.

[22] A. Hamza, "An enhanced breast cancer diagnosis scheme based on two-step-SVM technique," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 158–165, 2017.

[23] C. Ming, V. Viassolo, N. Probst-Hensch, I. D. Dinov, P. O. Chappuis, and M. C. Katapodi, "Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: Impact on screening recommendations," *Brit. J. Cancer*, vol. 2020, pp. 1–8, Jun. 2020.

[24] A. H. Osman and H. M. A. Aljahdali, "An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model," *IEEE Access*, vol. 8, pp. 39165–39174, 2020.

[25] M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, and W. Zhou, "Breast cancer detection in the IOT health environment using modified recursive feature selection," *Wireless Commun. Mobile Comput.*, vol. 2019, pp. 1–19, Nov. 2019.

[26] W. Wolberg, N. Street, and O. Mangasarian, "Wisconsin diagnostic breast cancer (WDBC)," UCI Mach. Learn. Repository, Sacramento, CA, USA, Tech. Rep., 1995. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[27] P. Meesad, P. Boonrawd, and V. Nuipian, "A chi-square-test for word importance differentiation in text classification," in *Proc. Int. Conf. Inf. Electron. Eng.*, 2011, pp. 110–114.

[28] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Nashville, TN, USA, 1997, vol. 97, nos. 412–420, p. 35.

[29] G. N. Ramadevi, K. U. Rani, and D. Lavanya, "Importance of feature extraction for classification of breast cancer datasets, a study," *Int. J. Sci. Innov. Math. Res.*, vol. 3, no. 2, pp. 368–763, 2015.

[30] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 1151–1157.

[31] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proc. Mach. Learn.* Amsterdam, The Netherlands: Elsevier, 1992, pp. 249–256.

[32] R. J. Urbanowicz, M. Meeker, W. L. Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Informat.*, vol. 85, pp. 189–203, Sep. 2018.

[33] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length, and Helmholtz free energy," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 3–10.

[34] B. Mwangi, T. S. Tian, and J. C. Soares, "A review of feature reduction techniques in neuroimaging," *Neuroinformatics*, vol. 12, no. 2, pp. 229–244, Apr. 2014.

[35] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, Apr. 2016, Art. no. 20150202.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[37] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 9014–9022, Jul. 2011.

[38] A. U. Haq, J. Li, Z. Ali, M. H. Memon, M. Abbas, and S. Nazir, "Recognition of the Parkinson's disease using a hybrid feature selection approach," *J. Intell. Fuzzy Syst.*, vol. 39, pp. 1–21, Jan. 2020.

[39] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data," *Sensors*, vol. 20, no. 9, p. 2649, May 2020.

[40] J. Dheeba, N. A. Singh, and S. T. Selvi, "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach," *J. Biomed. Informat.*, vol. 49, pp. 45–52, Jun. 2014.

[41] Y. Prasad, K. K. Biswas, and C. K. Jain, "SVM classifier based feature selection using GA, ACO and PSO for siRNA design," in *Proc. Int. Conf. Swarm Intell.* Berlin, Germany: Springer, 2010, pp. 307–314.

[42] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "Breast cancer diagnosis using an unsupervised feature extraction algorithm based on deep learning," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9428–9433.

AMIN UL HAQ is currently working as a Postdoctoral Scientific Research Fellow with the University of Electronic Science and Technology of China (UESTC), China. He has a Vast Academic, Technical, and Professional Experience, in Pakistan. He is also a Lecturer with Agricultural University Peshawar Pakistan. His research interests include machine learning, deep learning, medical big data, the IoT, e-health and telemedicine, concerned Technologies, and Algorithms. He is also associated with the Wavelets Active Media Technology and Big Data Laboratory, as a Postdoctoral Scientific Research Fellow. He has been published high-level research articles in good journals. He is an invited Reviewer of numerous world-leading high-impact journals (reviewed more than 40 journal articles to date).

JIAN PING LI is currently the Chairman of the Computer Science and Engineering College and the Model Software College, University of Electronic Science and Technology of China. He is also the Director of the International Centre for Wavelet Analysis and its Applications. He is also the Chief Editor of the International Progress on Wavelet Active Media Technology and Information Processing. He is also the Associate Editor of the *International Journal of Wavelet* and *Multimedia and Information Processing*. He received the National Science and Technology Award Evaluation Committee, the National Natural Science Foundation Committee of China, The ministry of public security of the People's Republic of China, such as a Technical Adviser and a dozen academic and social positions.

ABDUS SABOOR is currently pursuing the M.S. degree from the School of Computer Science and Engineering, UESTC, China. He is also a Lecturer with Government University Peshawar Pakistan. His research interests include machine learning, medical big data, the IoT, e-health and telemedicine, concerned technologies, and algorithms.

JALALUDDIN KHAN received the M.S. degree in computer science from Aligarh Muslim University Aligarh, India, and the Ph.D. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China. He has an Impressive Academic, a Research, and a Professional Experience from the Kingdom of Saudi Arabia. He was a Lecturer with the Deanship of Skills Development and a Researcher, Center of Excellence in Information Assurance (COEIA), King Saud University, Riyadh, Saudi Arabia. His research interests include the IoT, security and privacy, e-health and telemedicine, machine learning, and medical big data concerned technologies. He is currently focusing the IoT security with medical data. He is also accompanying with the Wavelets Active Media Technology and the Big Data Laboratory under supervision with Prof. J. P. Li and with a collaborated way with other researchers in UESTC. He has authored some research articles.

SAMAD WALI received the B.Sc. degree in mathematics from the Forman Christian College Lahore, Pakistan, the M.Sc. degree in applied mathematics from The Islamia University of Bahawalpur, Pakistan, and the Ph.D. degree in computational mathematics from Nankai University, Tianjin, China, in June 2018. From 2018 to 2020, he was Postdoctoral Research Fellow in image processing and medical image analysis with the School of Communication and Information Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu. He is currently an Assistant Professor with Namal Institute Mainwali Pakistan. His current research interests include image processing, variational methods, and numerical solution to partial differential equations.

SULTAN AHMAD (Member, IEEE) received the master's degree (Hons.) in computer science and applications from the Prestigious Aligarh Muslim University, India. He graduated in Computer Science and Applications in 2002 from Patna University, India. He is currently working as a Lecturer with the Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia. He has a unique blend of education and experience. He has more than 12 years of teaching and research experience. His research and teaching interests include cloud computing, big data, machine learning, and the Internet of Things. He has presented his research papers in many national and international conferences and published research articles in many peer-reviewed reputed journals. He is a member of IACSIT and Computer Society of India.

AMJAD ALI received the Ph.D. degree in real time systems from Gyeongsang National University, South Korea. He is currently an Assistant Professor and the Chairman of the Department of Computer Science and Software Technology, University of Swat. He published several research articles in international journals and conferences.

GHUFRAN AHMAD KHAN received the M.S. degree in computer science from Aligarh Muslim University, Aligarh, India. He is currently pursuing the Ph.D. degree from the School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China. He has huge Academic and Technical Experience in India. He has authored some research articles. His research interests include machine learning, data mining, rough set theory, and deep learning.

WANG ZHOU received the Ph.D. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China. He is currently a Vice Professor with the School of Computer Science and Engineering, Xihua University. His current research interests include artificial intelligence, recommender algorithm, and data mining.