International Journal on Emerging Technologies **11**(4): 193-201(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Automated Decision Support System for Breast Cancer Prediction

Madhu Kumari[1], Vijendra Singh[2] and Prachi Ahlawat[3]
[1]Research Scholar, Department of CSE, Northcap University, Gurugram (Haryana), India.
[2]Associate Professor, Department of Informatics, School of Computer Science,
University of Petroleum and Energy Studies, Dehradun-248007 (Uttarakhand) India.
[3]Associate Professor, Department of CSE, Northcap University, Gurugram (Haryana), India.

(Corresponding author: Madhu Kumari)
(Received 11 May 2020, Revised 19 June 2020, Accepted 01 July 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Breast cancer is invasive and the second leading cause of mortality among women. Medical flaws during the diagnosis due to unintended human error are the most damaging and expensive kinds of investigative errors even with the contemporary diagnostic procedures. Automated decision support system to assist the medical expert is required during decision making that will help to avoid such unwanted events. But, selecting the most predictive feature and achieving the best classification accuracy is the major challenge for the systems that are designed to automatically predict cancer and gives decision support. In this study, an automated decision support system for predicting breast cancer has been proposed. It assists medical expert in the process of decision making. It predicts the breast cancer at the developing stage by learning from the most predictive set of features of the Wisconsin breast cancer dataset (WBCD). The performance of the proposed method is evaluated using different measures such as classification accuracy, precision, recall, f1-score and confusion matrix. The classification accuracy achieved in this study is 99.28%, which is evidently good when compared with state-of-art methods.
The proposed model is intended to assist the medical experts by providing quick, precise and reliable recommendations that can directly be applied in the process of crucial decision-making and contribute towards the improvement of quality of life.

Keywords: Breast cancer, decision support system, feature selection, KNN, Knowledge-mining, WBCD.

Abbreviations: LR, Linear regression; SVM, support vector machine; KNN, K-nearest neighbor.

I. INTRODUCTION

The prevalence of chronic diseases such as breast cancer is increasing gradually with the urbanization and digitization of our modern society. Breast cancer is the most invasive life-threatening disease among females. Deficient diagnosis and high cost of treatment contribute to high mortality. Cancer is the noticeable lump in which cells begin to rattle uncontrollably and can be fatal. These lumps are called tumors which can be benign (non-cancerous) or malignant (cancerous). The uncontrolled growth of lump is identified as cancerous (malignant tumor) which also invades to its nearby tissues and destroy healthy body cells. Breast cancer is the leading cause of mortality among females. The etiologies of uncontrollable growth of body cell is still unidentified and therefore prevention is challenging. The most successful approach to increase the cancer patient survival, is to identify it at an early stage. It helps medical expert to make the optimal decision about treatment strategies. Early detection is the most desirable but often difficult to achieve as the symptoms are usually not prominent in the developing phase. Early detection involves a precise and consistent detection system that analyze available healthcare data and allows general practitioner to differentiate benign breast tumors from malignant ones, devoid of surgical biopsy.

The survival rates of breast cancer have increased with an increased emphasis on diagnostic techniques and more effective treatment strategies [1]. To detect malignant tumor advanced screening and lab test are performed on patients. Medical expert evaluates the data gathered from various screening practices and gives decision about the presence or absence of malignancy. The decision solely depends upon the experience of medical expert and became the major source of medical flaws. Therefore, an automated decision support system is required. An automated decision support system will assist medical expert by providing suitable suggestion during the crucial process of decision making. This additional automated decision support will minimize the probability of occurrence of medical flaws due to fatigued or inexperience medical experts. But, selecting most appropriate features and accuracy improvement is the main challenge for the design of such automated decision support systems.

In the context of the automated decision support, statistical data learning and data mining techniques are the complement to the researches in the field of healthcare and biotechnologies. The exploitation of machine learning and big data technique in medical science provided evidence that it can greatly reduce the healthcare cost, mortality rate and improves the treatment by assisting oncologist in the process of decision-making. A Data-driven statistical study is becoming a prevalent complement to many scientific areas like medicine, genomics, and biotechnology. Certainly, analysis of available patient's medical

information and judgment of doctors is the most significant feature in diagnosis. Moreover, automated decision support system using data mining and machine learning techniques for classification, prediction and diagnosis also help medical expert in a great deal. The objective of this study is to develop an automated decision support systems that will provide a reliable and quick recommendation about the sample under consideration during the screening. The proposed model has two phase. In the first phase, most predictive features are extracted by applying advanced feature selection techniques. In the second phase, three classifiers are trained and evaluated on the basis of accuracy achieved.

Many knowledge engineering methods have been used in modern years for malignancy prediction and prognosis [2, 3]. Automated methods by utilizing the power of data science and machine learning can discover hidden knowledge and statistical regularities from the available heap of historical data to make predictions on new data. A range of machine learning techniques was used by the researchers for predicting susceptibility [4-6], diagnosis [7-13], recurrence [14-20] and survivability [21-27] of breast cancer in women.

This study will contribute towards minimizing the probability of medical flaws by providing automated decision support during the process of decision making to the medical expert. Section II recapitulates the present state-of-art methods reported in the literature. The detailed framework of the proposed model is given in section III. Performance of the proposed model is evaluated and compared with other state-of-art methods in section IV. Sections V conclude the research finding and outline the limitations and suggest further research directions.

II. RELATED WORK

A substantial amount of work has been done by the researchers to explore the potential of healthcare data analysis, and most of them turn up with good classification accuracy. A classification modal "LS-SVM" was proposed by Polat *et al*., to detect the breast cancer in women and by using cross-validation he attained the accuracy of 98.53% [8]. For breast cancer detection, Akay proposed a new method using SVM and attained the classification accuracy of 99.02% [9]. A breast cancer detection model with the accuracy of 98.71 % was designed by Yeh *et al*., by using machine learning techniques with swarm optimization [28]. For the detection of hepatitis disease, Kaya and Uyar proposed a hybrid framework by utilizing back propagation neural network and rough set theory. UCI hepatitis disease dataset was selected to experiment. Rough set theory was used to form 20 reduct with 3 to 7 features and ignoring records with the missing value from each reduct. The selected reducts are then trained using back propagation neural network and the classification accuracy of 98.6% was achieved [29]. A new method for classification of breast cancer dataset was given by Marcano-Cede *et al*, utilizing ANN over biological metaplasticity property proves to perform well with the classification accuracy of 99.26% [30]. Chen *et. al* achieved the classification accuracy of 89.2% on WBCD by using Rough Set algorithm to select the most predictive in combination with SVM classifier for training [31]. Setiono proposed a method that improves the performance of WBCD classification by achieving the accuracy of 98.10%. Before applying classifier, the dataset is first pre-processed. Attributes having missing values are simply ignored and a Neural Network classifier is trained on most predictive features selected by one hidden layer of neural networks. This technique reduced the training time and enhances the accuracy of the model [32]. Nahato *et al.,* used rough set in combination with backpropagation neural network for mining clinical dataset to uncover hidden knowledge. The classifier reported promising accuracy of 97.3%, 98.6%, and 90.4% for hepatitis, WBCD and Statlog heart disease datasets [33].

An automatic diagnostic model was proposed by Karabatak *et al*., for detecting breast cancer. The dimensionality of the WBCD is reduced by using association rules and then the model is trained with neural network classifier. 3- fold CV is used in the testing phase, which results in the classification accuracy of 95.6% [34].

Another hybrid method for breast cancer diagnosis was developed by Seral *et al.,* In this approach, the size of the training dataset was reduced by using AIS artificial intelligence algorithm. The Fuzzy weighing technique was adopted to consolidate the effect of using the distance-based algorithm. The above-proposed model was then trained with classifier knn. The fuzzy-AIS-Knn model with 10 fold cross-validation achieved a very high accuracy of 99.14% [35].

A hybrid hepatitis diagnostic system was developed by Sartakhti *et al*., [36]. The dataset used for the study was from the UCI repository. They eliminated the missing values from the original dataset and reduced the size to 80 samples. The entire dataset is divided into two classes having labels "Die" and "live." There were 67 samples of "Die" classes and the remaining 13 samples were belongs to "Live" class. Standard normalization techniques were used to standardize the dataset before fitting it to the classifier. They used SVM classifier for training the model along with simulated annealing. With the 10 fold cross-validation they achieved the accuracy of 96.25%. Liu *et al.,* designed a fully automated breast cancer diagnostic system by detecting true mass segmentation boundaries [37]. Patil *et al.,* proposed a hybrid model for analyzing available clinical datasets and predicting disease outcome. K means clustering and C4.5 decision-tree classifier along with k fold cross validation when applied to eight datasets from UCI repository shows remarkable improvement in the classification accuracy as compared other studies present in the literature [38].

Different classification model has been designed to analyze available clinical data to identify the malignancy of the tumor [39-43].

In addition to the above-listed research efforts, there are other studies related to using available healthcare data for prediction in medical domains [43-49]. These studies are just a small representative of the existing large number of research in utilizing data mining techniques to various medical domains for prediction and pattern recognition purposes.

III. MATERIALS AND METHOD

By utilizing data mining and machine learning techniques, an automated breast cancer decision support system is proposed that assist medical expert to identify malignant masses. Further, on the basis of learning acquire in the training phase predicts the malignancy when exposed to a new data sample and support medical expert in making the optimal decision. The architecture of the proposed automated system contains the following different phases:

— Selecting Data set
— Data Understanding and Pre-processing
— Handling missing values.
— Normalization
— Feature selection Model Building (Training) using SVM, Linear Regression and KNN.
— Testing and comparing the trained model on the basis of accuracy achieved.
— Using classifier with the highest accuracy for prediction.
— Providing decision support to medical support.

A. Selecting data set

Collecting the most appropriate data from the precise domain that is meaningful, informative and facilitate learning during analysis is a crucial task. In this study, Breast cancer dataset was obtained from the UCI repository having ten attributes that are used to predict that the breast tumor is malignant or benign. Based on past information stored in the dataset, the classifiers are trained to identify cancerous tumor.

All of the attributes are of a numeric-valued continuous data type. The attribute for the class label is a dichotomous variable (i.e., the binary response variable) within the dataset follows each tuple of the dataset. WBCD (Original) from UCI repository has 699 instances and 16 instances among them contain missing values. 35.0% of samples of the total dataset are malignant and evenly distributed within the class. The detailed descriptions of all the 11 attributes are shown in Table 1.

Fig. 1 explains the detailed framework of the proposed work which is explained in detail in succeeding sections.

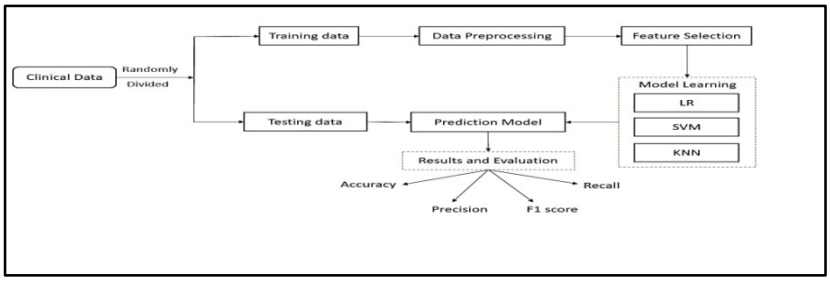

Fig. 1. Proposed System Framework.

Table 1: WBCD in-depth detail.

	id_no.	clum_thickness	unif_ cell_size	unif_ cell_shape	marge_ adhesion	single_epith_ cell_size	bare_nucloi	bland_ chromatin	norm_ necleoli	mitosis	class
count	699.00	699.00	699.00	699.00	699.00	699.00	699.00	699.00	699.00	699.00	699.00
mean	0.071	4.417	3.314	3.207	2.806	3.216	3.45	3.437	2.866	1.589	2.689
std	6.17	2.281	3.051	2.971	2.855	2.214	3.63	2.438	3.053	1.715	0.951
min	6.163	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	2.00
25%	8.706	2.00	1.00	1.00	1.00	2.00	1.00	2.00	1.00	1.00	2.00
50%	1.171	4.00	1.00	1.00	1.00	2.00	1.00	3.00	1.00	1.00	2.00
75%	1.238	6.00	5.00	5.00	4.00	4.00	5.00	5.00	4.00	1.00	4.00
max	1.345	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	4.00

B. Data Pre-processing

Pre-processing is the most crucial and important step, that significantly affects the performance of the classifier. It is important to pre-process the dataset under consideration before training it on a classifier to enhance the ability to learn the unseen patterns in the dataset. Fig. 2 shows the framework for preprocessing of WBCD. Ignoring the sample code number and class attribute, remaining 9 features provide specific information appertain to the incidence of malignant tumor. The sample space was examined for unidentified values, inconsistency and flawed data. The dataset was scanned for missing values that were represented by"?".Such values can significantly influence the analyses derived from the data. Replace all those null value initially with "nan". Each missing value is calculated and imputed by using regression model. All available information in the dataset is utilized to predict the missing values of the specific feature. Imputing missing value with regression preserve its correlation with other attributes of dataset. To scale the entire dataset into one standard range, the dataset is normalized using the min-max normalization technique.

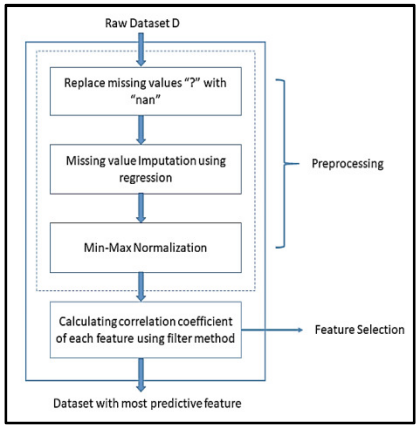

Fig. 2. Preprocessing of WBCD dataset.

```
//Pseudocode for filling missing values
// Dataset: d, Features: f, index: i , number of //instances
in feature: n
begin
seti=0
repeat for i=1 to n
if f[i] is null
predict value "v" by fitting to regression
impute f[i] with value "v"
end of if
end of for
end
```

```
//Pseudocode for min-max normalization
// D: Dataset
begin
define dataset_minmax(D)  // Find the min and max values
for each column
minmax=list()
i=1
repeat while i in range len(D[0])
        column_val=[row[i]for each row in D]
        min_val=min(column_val)
        max_val=max(column_val)
        append to minmax([min_val],[max_val])
end of while loop
return minmax
end of dataset_minmax
define normalize(D, minmax)
//Normalize dataset columns to the range 0-1
row=1
repeat for each row in D
        repeat while j<=len(row)
                row[j]=(row[j]-
minmax[j][0])/(minmax[j][1]-minmax[j][0])
        end of while
end of for
end
```

Distributions of data within the dataset helps in identifying the most predictive and significant feature that can participate in learning process. Fig. 2 shows distribution of all the features with reference to the class variable.

C. Feature selection

Building the prediction system with most predictive features improves the predictive power and reduces the complexity of the overall prediction system. Appropriate features can be selected from the data set either by evaluating the significance of the features under consideration using learning algorithms or by evaluating the significance of feature based on score obtained by applying some statistical test.

To select the most predictive features we have used the filter method. A mathematical function is used to find out the association between independent variable and dependent variable. The features are selected depending on their correlation coefficient values. Highly correlated feature with the class variable are said to be most predictive one and included in the final feature set. Pearson's linear correlation: Consider a dataset D having feature set F.

$$F= \{x_1, x_2, x_3 \ldots\ldots\ldots\ldots x_n\} \quad (1)$$

and classes C with values c, where X, C are treated as random variables, Pearson's linear correlation coefficient is defined as

$$r= \frac{\sum_{i=1}^{n}(x_i-x)(c_i-c)}{\sqrt{[\sum_{i=1}^{n}(x_i-\bar{x})^2][\sum_{i=1}^{n}(c_i-\bar{c})^2]}} \quad (2)$$

Where x_i and c_i the i^{th} value of x and c respectively. % of (r) = ±1 if X and Y are linearly dependent and zero if they are completely uncorrelated.

Table 2 shows the correlation coefficient values between independent features and dependent class variable. To select the most predictive features we have used the filter method. A mathematical function is used to find out the association between independent variable and dependent. D 4 describes the association between WBCD variable. On the basis of correlation coefficient score, six (bare_nucloi, unif_cell_shape, unif_cell_size, bland_chromatin, clum_thickness, norm_necleoli) features are selected for model learning.

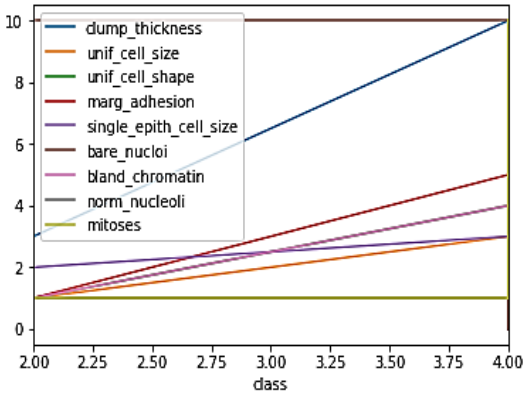

Fig. 3. Joint data distribution of all the features with respect to class.

Table 2: Correlation between independent variables and dependent variable.

	id_no.	clum_ thickness	unif_ cell_size	unif_ cell_shape	marge_ adhesion	single_epith_ cell_size	bare_nucloi	bland_ chromatin	norm_ necleoli	mitosis	class
id_no	1.00	-0.05	-0.04	-0.04	-0.06	-0.04	-0.08	-0.06	-0.05	-0.03	-0.08
clum_thickness	-0.05	1.00	0.64	0.65	0.48	0.52	0.51	0.55	0.53	0.37	0.71
unif_cell_size	-0.04	0.64	1.00	0.90	0.70	0.75	0.61	0.75	0.72	0.45	0.81
unif_cell_shape	-0.04	0.65	0.90	1.00	0.68	0.71	0.63	0.73	0.71	0.43	0.81
marge_adhesion	-0.06	0.48	0.70	0.68	1.00	0.59	0.59	0.66	0.60	0.41	0.69
single_epith_cell_size	-0.04	0.52	0.75	0.71	0.59	1.00	0.52	0.61	0.62	0.47	0.68
bare_nucloi	-0.08	0.51	0.61	0.63	0.59	0.52	1.00	0.62	0.52	0.28	0.84
bland_chromatin	-0.06	0.55	0.75	0.73	0.66	0.61	0.62	1.00	0.66	0.34	0.75
norm_necleoli	-0.05	0.53	0.72	0.71	0.60	0.62	0.52	0.66	1.00	0.42	0.71
mitosis	-0.03	0.35	0.45	0.43	0.41	0.47	0.28	0.34	0.42	1.00	0.42
class	-0.08	0.71	0.81	0.81	0.69	0.68	0.72	0.75	0.71	0.42	1.00

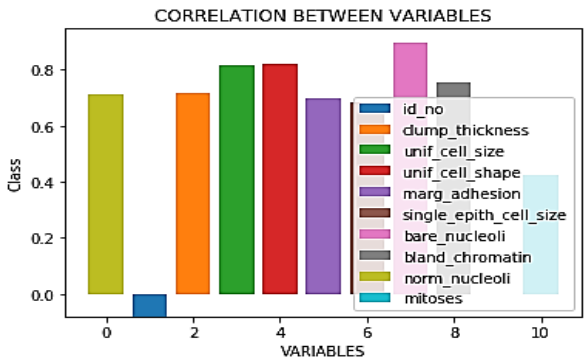

Fig. 4. Association among WBCD variables.

D. Training and Classification

Classification is a supervised learning technique to classify the input data into a particular class by learning through classifiers. Classifiers learn through the attributes having specific decision properties about the class labels. It is done by, initially, training the known sample data using various classifiers and then predicting trained samples. The objective for this study is to train the model with the three most widely used classifier i.e. LR, SVM and KNN, and then make the predictions for new sample on the basis of classifier which performs best among them. Classification is done by training the known sample dataset with the classifier and the performance of the model is measured in terms of accuracy achieved while predicting the unknown sample based on above learning.

(i) LR: Linear regression is a simple supervised learning approach to project the association between explanatory variables and dependent variable by fitting to a linear equation on experiential data. Mathematically, linear regression can be modelled as:

$$y=\beta_0 + \beta_1 x_1 + e \quad (3)$$

Where, *y:* response variable

β_0 and β_1 : model coefficients.

These unknown constant values represent the intercept and slop, and are learned during training phase.

e:error associated with the given value of $\beta_0$and β_1.. After fitting the model in the training phase, prediction is made on the basis of following conditions:

$$y=\beta_0 + \beta_1 . x \quad (4)$$

Where y is the predicted value on the basis of x

Error eis given by:

$$e_i = y_i + \hat{y}_1 \quad (5)$$

ith residual.

(ii) SVM: SVM is a linear or nonlinear, statistical, supervised learning algorithm that was introduced by Vapnik *et al.* [34] and has been applied in a large variety of application. Classification is done by projecting the input data points into n-dimensional vector space and finding the best hyper-plane that maximize the margin between the two classes. The performance of SVM is greatly influenced by the choice of parameters like kernel, C and gamma. Kernels are the function which convert low dimensional space into high dimensional space and makes the classification easy. Kernel control the non-linearity Gamma is the constant value for misclassification tolerance. Linear separability is achieved by projecting low dimensional space with the use of margin maximization and kernel function. Separating hyper is defined as:

$$f(x)=b+ w.x \quad (6)$$

Where w is the *weight vector* and b is the *bias*. The value of the parameter "b" and "w" are scaled infinite number of times to find the best separating hyper plane that maximize the margin between the sample data points.

OSH (Optimal separating hyper plane) is defined as:

$$b+ w.x=1 \quad (7)$$

Where x is the training sample points which are also called support vectors and are closest to the hyper plane, also called canonical hyper plane.

The distance among the training data points (x) and a hyper plane is given as:

$$D_{\text{support vector}} = \frac{|b+w.x|}{\|w\|} = \frac{1}{\|w\|} \quad (8)$$

It is a canonical hyper plane. Margin is twice the distance between the hyper plane and the support

$$M = \frac{2}{\|w\|} \quad (9)$$

Best hyper plane is one that has the maximum separating margins between both the classes and this can be achieved by maximizing M and minimizing L(w) liable to some constraints. Formally,

$$\min_{w,b} \mathrm{L(w)} = \frac{1}{2}\|w\| \quad (10)$$

Subject to $y_i(w, x^i + b) \geq 1 \ \forall \ i$

Where x_i are the training samples and y_i is the training data sample labels.

(iii) KNN: KNN is a supervised classifier which learns from the labelled data samples. It is lazy algorithm as it does not perform any generalization about the sample data points and all the computations are pending until the classification. Given n training vectors, KNN works by identifying the K nearest neighbours. KNN map the training data set into multi-dimensional feature space and partition them into different regions according to the classes of the training dataset. The key element of this algorithm is the notion of similarity distance. There are different ways to calculate this similarity measure for identifying the nearest neighbours such as Euclidean distance, Manhattan Distance, Minkowski distance for continuous variable and hamming distance for categorical variables. Here we have used a Euclidean distance measure to calculate the similarity between the data points. Similarity

$$\text{Distance} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (11)$$

Where x_i and y_i are two data points in the feature space. A point in the feature space is assigned the most frequent category among the k nearest training data. In the training phase, all the feature vectors are stored according to the regions of the training data set. During the classification phase, similarity distance between the test document and each feature vectored are calculated and then allotted the test document to the class of its majority k neighbors.

Algorithm 1: KNN Pseudo code

```
//X :training dataset, C : Class labels of X, y: unknown samples, d: Distance Matrix
K Neighbors Classifier(X, C, y)
for(i=0; i<=m;i++)
        Compute distance d (Xi, y)
 Calculate set Z containing indices for the k smallest distances d(Xi, y).
Return class with {Ci where i∈ Z}
```

In this study, to get the optimal value of K, the entire data set is divided into the training set and test set. A small portion of the training data is ripped off to form the validation set. This validation set was used to evaluate the different possible values of K. The validation model were then trained with initial k value 1 and iterated 50 times with different k values ranging from 1 to 50. The k-value which gives the highest accuracy were used in the final model learning process and that is 5. Therefore, the final model was trained with k=5.K fold CV is statistical technique that minimizes the biasness allied with random training data samples. It involves the splitting of actual sample space into mutually exclusive k identical size subsamples where any one subsample is used for validating the model and remaining k1 subsamples are utilized as a training data. With this approach each k^{th} sample is used just once for iteratively validating the system. The value of k is set to 10 for this study. The accuracy of the model with cross validation is then estimated as the average accuracy of k individual iteration accuracy as:

$$\text{Cross validation accuracy} = \sum_{i=1}^{k} a_i \quad (12)$$

IV. RESULTS AND DISCUSSION

The experiments are performed on Intel(R) Core(TM) i7-8565U CPU @ 1.80 GHz with 8 GB RAM. Scientific Python Development Environment (SPYDER) is used with Python 3.7.3. It is an open-source Integrated Development Environment for writing and executing python codes.

Dataset is divided into training and testing sets in 80%-20% ratio. Classifier training phase is validated by using 10 fold cross validation method.

The performance of the proposed model is compared with the existing work done by other researchers in the literature and evaluated in terms of classification accuracy. Additionally, compared to the state-of-art methods, the proposed model is observed to perform excellent with classification accuracy of 99.28% via 10-fold cross validation (CV) analysis.

KNN is a versatile algorithm that performs extremely well in many situations. Real world data is very in-deterministic and do not obey any typical theoretical assumptions made and KNN is the classifier which do not require much knowledge about the data distribution. The statistical model of KNN is determined from the samples under consideration. This non parametric behavior of KNN makes it low bias classifier. Decision making in healthcare domain, more precisely disease prediction is very crucial. Ignoring, discarding and missing out any valuable information during learning process can leads to devastating results.

For such problem KNN turns out to be effective classifier as it does not perform any generalization on the data points.

All the training data are required at the testing time. Most of the important calculations are pending till the testing phase.

Only the feature vector and class labels are stored in the training phase. Predictions are done in the testing phase on the basis of the feature similarity which is measured by calculating the distance between two data points.

In this study, we have used Euclidean distance as it treats each feature as equally significant. The performance of KNN is greatly affected by the choice of value of K. Higher value of K will bring more bias whereas a too low value makes it more sensitive to noise. No explicit training is required for learning KNN classifier. It is been observed from the experiment that KNN outperform other classifiers used in this study in terms of classification accuracy.

Table 3: Performance of classifier considered in the study.

Classifier	Precision (%)	Recall (%)	F1score (%)	Accuracy (%)
LR	49.69	37.00	41.51	79.84
SVM	65.45	61.92	63.59	86.10
KNN	99.10	99.41	99.25	99.28

Performance of all the three classifier used here is evaluated and document in Table 3. Fig. 5 shows ROC curve for the three classifier.

A lot of research had been done by the researcher for the detection and timely prediction of breast cancer to improve the patient's survival rate.

Various statistical techniques are used by the different authors in order to attain high classification accuracy of the model. Proposed model under this study turns out to achieve best performance in terms of classification accuracy as compared to the work reported in the present literature. Performance measurement of previous study by different researchers on the same dataset by using different approaches is given in Table 4. EHR capture structured controlled vocabulary encoded with lab results, general information about patient and medication list etc. whereas a wealth of information is present in unstructured clinical data. Unstructured data is computationally complex due to its high dimensionality. Most of the clinical data remains unexplored which limits the potential healthcare data analysis. These unstructured clinical data are the rich source of information for learning specific patterns and findings based on NPL algorithms and statistical learning method.

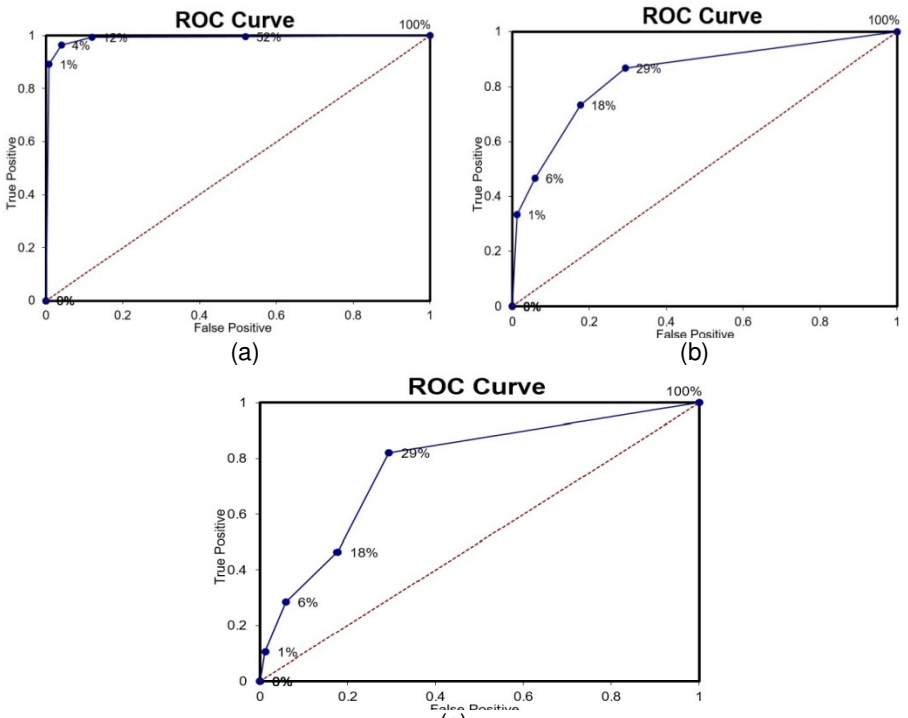

Fig. 5. ROC graphs for (a) Optimized KNN, (b) SVM, (c) LR.

Table 4: Evaluating performance of this study with the previous work present in the literature.

Dataset	Authored By	Approach	Accuracy achieved	This study Accuracy
WBCD (Original)	Marcano-Cedeno [30]	Metaplasticity Neural Network	99.26%	99.28%
	Chen HL *et al.*, [31]	Rough set theory and SVM	89.20%	
	Setiono [32]	Feature selection and neural network	98.10%	
	Nahato KB *et al.*, [33]	Rough set and back propagation network	98.60%	
	Karabatak *et al.*, [34]	Association rule and Neural Network	95.6%	
	Sahan *et al.*, [35]	Fuzzy, AIS and KNN	99.14%	
	Ed-daoudy *et al.*, [42]	Association Rules and SVM	98.00%	

V. CONCLUSION

Automated decision support systems in medical domain are the intelligent tool that assist medical experts in making optimal decision in the case of ambiguity and inadequate information which reduces the overall cost of treatment. WBCD dataset is classified by using different classifiers to conduct the experiments under this study. It is evident from the work that KNN classifier performs best when used with most predictive variables and attain the accuracy of 99.26%. The proposed model is intended to assist the medical experts by providing quick, precise and reliable recommendations that can directly be applied in process of crucial decision-making and contribute towards the improvement of quality of life.

VI. FUTURE SCOPE

In this study we have considered only the structured data but 80% of the available healthcare data is unstructured. The prediction of certain medical illness with the analysis of unstructured healthcare dataset will be the focus of future work.

Conflict of Interest. No.

REFERENCES

[1]. Jemal, A., Murray, T., Ward, E., Samuels, A., Tiwari, R. C., Ghafoor, A., & Thun, M. J. (2005). Cancer statistics, 2005. *CA: a cancer journal for clinicians*, *55*(1), 10-30.

[2]. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 117693510600200030.

[3]. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.

[4]. Listgarten, J., Damaraju, S., Poulin, B., Cook, L., Dufour, J., Driga, A., & Zanke, B. (2004). Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. *Clinical cancer research*, *10*(8), 2725-2737.

[5]. Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital signal processing*, *17*(4), 694-701.
[6]. Ayer, T., Chhatwal, J., Alagoz, O., Kahn Jr, C. E., Woods, R. W., & Burnside, E. S. (2010). Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. *Radiographics*, *30*(1), 13-22.
[7]. Wang, X. H., Zheng, B., Good, W. F., King, J. L., & Chang, Y. H. (1999). Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics*, *54*(2), 115-126.
[8]. Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital signal processing*, *17*(4), 694-701.
[9]. Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert systems with applications*, *36*(2), 3240-3247.
[10]. Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial intelligence in Medicine*, *25*(3), 265-281.
[11]. Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014-9022.
[12]. Liu, H. X., Zhang, R. S., Luan, F., Yao, X. J., Liu, M. C., Hu, Z. D., & Fan, B. T. (2003). Diagnosing breast cancer based on support vector machines. *Journal of Chemical Information and Computer Sciences*, *43*(3), 900-907.
[13]. Kiyan, T., & Yildirim, T. (2004). Breast cancer diagnosis using statistical neural networks. *Istanbul University-Journal of Electrical & Electronics Engineering*, *4*(2), 1149-1153.
[14]. Dai, H., van't Veer, L., Lamb, J., He, Y. D., Mao, M., Fine, B. M., & Stoughton, R. (2005). A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer research*, *65*(10), 4059-4066.
[15]. Jerez-Aragonés, J. M., Gómez-Ruiz, J. A., Ramos-Jiménez, G., Muñoz-Pérez, J., & Alba-Conejo, E. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial intelligence in medicine*, *27*(1), 45-63.
[16]. De Laurentiis, M., De Placido, S., Bianco, A. R., Clark, G. M., &Ravdin, P. M. (1999). A prognostic model that makes quantitative estimates of probability of relapse for breast cancer patients. *Clinical Cancer Research*, *5*(12), 4133-4139.
[17]. Marchevsky, A. M., Shah, S., & Patel, S. (1999). Reasoning with uncertainty in pathology: artificial neural networks and logistic regression as tools for prediction of lymph node status in breast cancer patients. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology, Inc*, *12*(5), 505-513.
[18]. Kate, R. J., & Nadig, R. (2017). Stage-specific predictive models for breast cancer survivability. *International journal of medical informatics*, *97*, 304-311.
[19]. Mariani, L., Coradini, D., Biganzoli, E., Boracchi, P., Marubini, E., Pilotti, S., & Rilke, F. (1997). Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear Cox regression model and its artificial neural network extension. *Breast cancer research and treatment*, *44*(2), 167-178.
[20]. Kim, W., Kim, K. S., Lee, J. E., Noh, D. Y., Kim, S. W., Jung, Y. S., & Park, R. W. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of breast cancer*, *15*(2), 230-238.
[21]. Jonsdottir, T., Hvannberg, E. T., Sigurdsson, H., & Sigurdsson, S. (2008). The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining. *Expert Systems with Applications*, *34*(1), 108-118.
[22]. Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2008). Breast cancer survivability via AdaBoost algorithms. In *Proceedings of the second Australasian workshop on Health data and knowledge management-Volume 80* (pp. 55-64).
[23]. Bilal, E., Dutkowski, J., Guinney, J., Jang, I. S., Logsdon, B. A., Pandey, G., & Rueda, O. M. (2013). Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS computational biology*.
[24]. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, *34*(2), 113-127.
[25]. Park, K., Ali, A., Kim, D., An, Y., Kim, M., & Shin, H. (2013). Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, *26*(9), 2194-2205.
[26]. Kim, J., & Shin, H. (2013). Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association*, *20*(4), 613-618.
[27]. Xu, X., Zhang, Y., Zou, L., Wang, M., & Li, A. (2012, October). A gene signature for breast cancer prognosis using support vector machine. In *2012 5th International Conference on BioMedical Engineering and Informatics* (pp. 928-931). IEEE.
[28]. Yeh, W. C., Chang, W. W., & Chung, Y. Y. (2009). A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. *Expert Systems with Applications*, *36*(4), 8204-8211.
[29]. Kaya, Y., & Uyar, M. (2013). A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. *Applied Soft Computing*, *13*(8), 3429-3438.
[30]. Marcano-Cedeño, A., Quintanilla-Domínguez, J., & Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications*, *38*(8), 9573-9579.
[31]. Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *38*(7), 9014-9022.
[32]. Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in medicine*, *18*(3), 205-219.
[33]. Nahato, K. B., Harichandran, K. N., & Arputharaj, K. (2015). Knowledge mining from clinical datasets using rough sets and backpropagation neural

network. *Computational and mathematical methods in medicine*.
[34]. Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert systems with Applications*, *36*(2), 3465-3469.
[35]. Şahan, S., Polat, K., Kodaz, H., & Güneş, S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. *Computers in Biology and Medicine*, *37*(3), 415-423.
[36]. Sartakhti, J. S., Zangooei, M. H., & Mozafari, K. (2012). Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). *Computer methods and programs in biomedicine*, *108*(2), 570-579.
[37]. Liu, F., Gong, Z., Chen, Y., & Gu, Y. (2015). Segmentation of mass in mammograms by a novel integrated active contour method. *International Journal of Computational Science and Engineering*, *11*(2), 207-215.
[38]. Patil, B. M., Joshi, R. C., & Toshniwal, D. (2010). Effective framework for prediction of disease outcome using medical datasets: clustering and classification. *International Journal of Computational Intelligence Studies*, *1*(3), 273-290.
[39]. Xue, Y., Zhao, B., Ma, T., & Liu, A. X. (2018). An evolutionary classification method based on fireworks algorithm. *IJBIC*, *11*(3), 149-158.
[40]. Winkler, S. M., Affenzeller, M., Kronberger, G., Kommenda, M., Wagner, S., Dorfer, V., & Stekel, H. (2013). On the use of estimated tumour marker classifications in tumour diagnosis prediction–a case study for breast cancer. *International Journal of Simulation and Process Modelling*, *8*(1), 29-41.
[41]. Ramos-Pollán, R., López, M. Á. N. G., & Ramos, I. (2015). Machine learning methods for breast cancer CADx over digital and film mammograms. *International Journal of Image Mining*, *1*(2-3), 208-223.
[42]. Ed-daoudy, A., & Maalmi, K. (2020). Breast cancer classification with reduced feature set using association rules and support vector machine. *NetMAHIB*.
[43]. Liu, K., Kang, G., Zhang, N., & Hou, B. (2018). Breast cancer classification based on fully-connected layer first convolutional neural networks. *IEEE Access*, *6*, 23722-23732.
[44]. Zhao, Z., Liu, Y., Li, J., Liang, J., & Wang, J. (2017). A comparative study on disease risk model in exploratory spatial analysis. *International Journal of Computational Science and Engineering*, *15*(3-4), 285-294.
[45]. Zhang, M., Pan, H., Zhang, N., Xie, X., Zhang, Z., & Feng, X. (2018). Cost-sensitive ensemble classification algorithm for medical image. *International Journal of Computational Science and Engineering*, *16*(3), 282-288.
[46]. Wu, K., Kang, J., & Chi, K. (2018). The fault diagnosis method of RVM based on FOA and improved multi-class classification algorithm. *International Journal of High Performance Computing and Networking*, *12*(2), 118-127.
[47]. Wen, C. H., Liao, W. D., Hsieh, T. Y., Chen, D. Y., Lan, J. L., & Li, K. C. (2009). Computer-aided image analysis aids early diagnosis of connective-tissue diseases. *SPIE Newsroom, Biomedical Optics & Medical Imaging*.
[48]. Wen, C. H., Liao, W. D., & Li, K. C. (2007). Classification framework for nailfold capillary microscopy images. In *TENCON 2007-2007 IEEE Region 10 Conference*, 1-4.
[49]. Sarvaiya, L. , Yadav, H. & Aggarwal, C. (2019). A Literature review of Diagnosis of Heart Disease using Data Mining Techniques. *International Journal of Electrical, Electronics and Computer Engineering, 8*(1): 40-45.

How to cite this article: Kumari, M., Singh, V. and Ahlawat, P. (2020). Automated Decision Support System for Breast Cancer Prediction. *International Journal on Emerging Technologies*, *11*(4): 193–201.