XBPF: An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability

Ravi Aavula, R. Bhramaramba

***Abstract:** Breast cancer is the second most lethal type of cancer causing death of woman. As a thumb rule prevention is better than cure. Prevention is possible with life style changes and healthy habits. It is also important to have early detection of it to prevent death. Many researchers contributed towards early detection, prognosis and better treatment of breast cancer in the last two decades causing decline of mortality rate. However, the breast cancer problem is still alarming and needs further research in the area of betterment of detection and prediction besides methods for treating it. Breast cancer prognosis is the holistic approach that covers three important aspects of research including prediction of susceptibility, recurrence and survivability. In this paper we propose an Extensible Breast Cancer Prognosis Framework (XBPF) for breast cancer prognosis which includes susceptibility or risk assessment, recurrence or redevelopment of the cancer after resolution, and survivability. We proposed a representative feature subset selection (RFSS) algorithm that is used along with SVM to improve efficiency in prognosis. SEER dataset is used to have experiments. A prototype is built to demonstrate proof of the concept. Our empirical study revealed that the framework is useful in prognosis of breast cancer instead of focusing on a particular aspect like susceptibility, survivability and recurrence individually. SVM-RFSS has shown significant performance improvement over state of the art prognosis methods.*

Keywords: Breast cancer, prognosis, survivability prediction, recurrence prediction, susceptibility prediction.

I. INTRODUCTION

When cells in human body grow out of control, then we call it cancer. Cancer can start in any part of human body. When cancer is in place, the body cannot function as it is supposed to. Cancer is not a single disease. In fact it is a collection of diseases as cancer can start in the colon, breast, lungs and even blood. Different cancers grow differently. The most common cancer type in woman is breast cancer. Breast cancer, as said earlier about cancer, starts when cell in woman's breast start growing abnormally. Thus a tumour is formed and that can be detected with a simple x-ray or it can be felt as a lump. The tumour is said to be malignant when it grows into surrounding tissues and distant areas of human body. Though men can get breast cancer, it is almost entirely occurring in woman. Different parts of the breast can be subjected to cancer. The ducts that carry milk to the nipple are subjected to breast cancer most of the cases. Other kinds of breast cancers are less common. Some cancers start in other tissues of breast. They are known as sarcomas and lymphomas and these are not really Cancer. It is essential to understand the fact that most of the breast lumps are not cancer. They are known as benign. They are abnormal growths but do not spread further and they are not life threatening. Checking lump to know whether it is malignant (actual breast cancer) or benign is essential [1]. According to [2], breast cancer is the second lethal cancer which needs to be addressed in the real world. It has 14.1% of all cancer cases in the world. It is really alarming that women are killed with breast cancer across the globe. The technological advances, fortunately, and innovative treatment survival rate of women who are diagnosed breast cancer is more. Nevertheless, breast cancer is still the problem to be addressed efficiently. In this research the focus is to investigate on data mining techniques for bioinformatics in order to bring about efficient prognosis method. From the literature it is understood that different approaches are used to deal with prediction of susceptibility, survivability and recurrence of breast cancer. A framework to deal with complete prognosis is desired. Different methods are found in the literature for breast cancer prognosis. Our contributions in this paper are as follows.

A. We proposed a framework with holistic approach to concentrate on breast cancer prognosis with three aspects of breast cancer research known as susceptibility, survivability and recurrence.

B. A representative feature subset algorithm is proposed to improve performance of classification performance with SVM.

C. Different data mining algorithms are explored for prediction of breast cancer susceptibility, survivability and recurrence and compared with the proposed framework where SVM-RFSS is used for prediction efficiency.

D. A prototype application is built to demonstrate proof of the concept. The empirical study revealed the usefulness of the framework.

The remainder of the paper is structured as follows. Section 2 reviews literature on breast cancer research that covers all aspects of prognosis. Section 3 presents the proposed methodology used to achieve breast cancer prognosis. Section 4 presents experimental results and evaluation. Section 5 throws light into the threats to validity. Section 6 concludes the paper and gives directions for future work.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)

Mr. Ravi Aavula*, Research Scholar, Department of Information Technology, GIT, GITAM, Visakhapatnam , Associate Professor, Department of C.S.E, Guru Nanak Institutions Technical Campus Hyderabad, (T.S), India

Dr. R. Bhramaramba, Professor, Department of Information Technology, GIT, GITAM, Visakhapatnam, (A.P), India

II. RELATED WORK

This section reviews literature on breast cancer prognosis. It is divided into three categories namely breast cancer susceptibility, recurrence and survivability. These three constitute the review of breast cancer prognosis.

A. Breast Cancer Susceptibility

Susceptibility refers to the likelihood of having breast cancer. In the breast cancer research, Rani et al. [1] studied the performance of synthetic minority oversampling technique (SMOTE) to deal with imbalanced breast cancer data. Velandia-Brinez et al. [2] explored classification of instances in breast cancer dataset using Complex Event Processing (CEP). Normal, benign and cancer are the class labels. Demigha [4] employed data mining techniques for breast cancer screening. Image segmentation techniques are used in [5] for breast cancer detection. Muthuselvan et al. [6] studied classification algorithms like Random Tree, J48, Naive Bayes, One R and Zero R. J48 showed better performance. Shen et al. [7] proposed a feature selection method known as INTERACT for selecting features and SVM for classification for better prediction. Seker et al. [8] focused on breast cancer prognosis using fuzzy logic based fuzzy kNN. Curz and Wishart [11] used machine learning methods for breast cancer prediction. The methods studied include decision trees, Naive Bayes, k-Nearest Neighbour, Neural Network (NN), Support Vector Machine (SVM) and Genetic Algorithm (GA). Durai et al. [14] explored linear regressive classification (LRC) method for breast cancer diagnosis and compared with BF tree, ID3, J48 and SVM. LRC could achieve highest accuracy. Anothaisintawee et al. [20] made a review of many risk prediction models related to breast cancer. Sena and Salem [22] investigated SVM, Tree Forest and Tree Boost in order to find efficiency in predicting breast cancer. Agarap [29] used data mining algorithm to predict breast cancer. The algorithms used to evaluate their prediction accuracy include Linear Regression (LG), MLP, NN and SVM. Dyrstad et al. [30] proposed a method for breast cancer prediction using risk analysis method. Bazazeh and Shubair [31] used machine learning algorithms like Bayesian Networks (BN), SVM and Random Forest (RF) to diagnose breast cancer. Precision, recall and ROC are used for evaluation. In terms of accuracy, SVM was found to be better than others. Eyl et al. [32] employed DT, PNN and NB to predict breast cancer. Wang [35] investigated Bayesian analysis in order to find the causal relationships between cancer diagnosis and clinical variables. Doreswamy and Salma [36] studied Fast Modular – ANN for breast cancer prediction using Wisconsin Breast Cancer Diagnostic Data (WBCD). Pack et al. [37] investigated multi-parameter SVM with fuzzy logic in order to have Computer Aided Diagnosis (CAD) of breast cancer. SVM showed better performance. Rani et al. [38] on the other hand derived breast cancer data using Natural Language Processing (NLP) and classified it to predict breast cancer cases. Arora and Tagra [39] explored many neural network algorithms for breast cancer diagnosis. The algorithms include MLP, SVM, PNN, CNN, RNN, and Neuro-fuzzy to have an expert system for breast cancer diagnosis.

B. Breast Cancer Recurrence

Predicting breast cancer recurrence has its utility in making decisions. Fan et al. [3] explored data mining techniques for prediction of recurrence using SEER dataset. They used algorithms like QUEST, CART, CHAID and C5.0. These are decision tree based mechanisms for prediction. Richter and Khoshgoftaar [10] focused on predictive modelling of recurrence using data mining techniques like decision trees, logistic regression (LR), Artificial Neural Network (ANN) and SVM. Roshani et al. [12] proposed a fuzzy expert system for predicting recurrence. They found it to be more accurate than C4.5, SVM and RBF network. Ojha and Goel [17] on the other hand employed decision tree and SVM classifiers in order to predict recurrence. They compared performance difference between classification and clustering algorithms. SVM based classifiers are used by Maglogiannis et al. [18] for breast cancer diagnosis. They compared SVM with ANN**.** They found that SVM could provide superior performance over ANN. Pritom et al. [23] explored breast cancer recurrence probability using UCI datasets. They used three data mining algorithms like SVM, C4.5 and Naive Bayes. They also used an efficient feature selection algorithm in order to improve the accuracy in prediction. Shin and Nam [24] proposed a coupling approach for breast cancer prognosis. It has a predictor and descriptor in order to have SVM supervised learning and for post processing respectively. Richter et al. [25] reviewed techniques used to find the recurrence of breast cancer. The algorithms they explored are SVM, Decision Tree and Artificial Neural Network (ANN). Abreu et al. [34] made a review of literature on breast cancer recurrence and found that it is difficult to find best dataset for breast cancer research. They also found it to be an open problem still.

C. Breast Cancer Survivability

Survivability statistics help in understanding the progress in breast cancer research. Seker et al. [8] made survival analysis of breast cancer datasets using fuzzy based k-nearest neighbour method. A survey of data mining methods for breast cancer research is found in [9]. Xu et al. [13] focused on Adaboost algorithms for survivability prediction. With feature selection and Adaboost algorithms, classification accuracy is improved. Adaboost could enhance prediction accuracy. Thongkam et al. [15] proposed a hybrid approach for making breast cancer survivability prediction models. They employed two important approaches like outlier detection and over-sampling approach. They proposed a framework to achieve this. They made experiments with different algorithms like C4.5, Adaboost, SVM and bagging. Out of them SVM was found to be more accurate. Behera and Rani [16] studied density based outlier detection approaches for breast cancer research. They are known as LOF, DENCLUE, DBSCAN and OPTICS. Kim and Shin [19] proposed a methodology for breast cancer survivability using datasets containing labelled, unlabelled and pseudo-labelled data.

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication (BEIESP)

They employed semi-supervised learning (SSL) as it can make use of unlabelled data. They also employed the concept of tagging virtual labels or pseudo labels. It is evaluated with data pertaining to epidemiology and surveillance. Khan et al. [21] employed fuzzy decision trees for survivability prediction using SEER datasets. Khan et al. [26] proposed Weighted Fuzzy Decision Trees (WFDT) in order to predict survivability of breast cancer. They employed weighted fuzzy decision trees in order to achieve this. They also studied the tradeoffs between linguistic fuzzy modelling and precise fuzzy modelling. The algorithm is compared with that of C4.5 decision tree in order to evaluate it using the state of the art techniques. Usage of data mining techniques in breast cancer research is carried out in [7]. The techniques include C4.5 and ID3. Delen et al. [28] proposed a methodology for finding breast cancer survivability using two mining algorithms like decision trees and ANN. They found that these algorithms can be used to provide a reliable decision making. Jhajharia et al. [33] employed three data mining algorithms such as NB, DT, SVM, IBK and One R for breast cancer survivability. From the literature it is revealed that the existing methods need to be improved further with efficient feature selection methods. Towards this end, we proposed a representative feature selection algorithm to enhance performance of classifiers.

III. PROPOSED PROGNOSIS FRAMEWORK

A framework is proposed to have breast cancer prognosis which includes susceptibility prediction, recurrence prediction and survivability prediction. The framework is flexible and extensible so as to support future techniques with ease. It is named as eXtensible Breast Cancer Prognosis Framework (XBPF) for breast ancer prognosis. The framework is extensible and supports new algorithms in future. First of all datasets are collected for having effective prediction models. Latest datasets provided by SEER for breast cancer prognosis are used for empirical study. The datasets are subjected to pre-processing as needed. Then there are three sets of experiments made. First set of experiments are associated with susceptibility prediction. This is nothing but risk assessment which is carried out based on the vital signs of patients. Then the second set of experiments is made on the recurrence probability of breast cancer. After resolution of breast cancer it may recur again. This is called recurrence probability. Then the third set of experiments focus on survivability of breast cancer patients. These three experiments are made with the proposed framework and the underlying techniques proposed for each set of experiments. The prediction results are interpreted to make well informed decisions.

A. Overview of Prediction Models

The prediction models are made using the combination of proposed RFSS algorithm and SVM. The prediction models for the three aspects of breast cancer prognosis are presented in Figure 2. SEER dataset is collected as described in Section 4 which has details of breast cancer associated with its instances. The SEER dataset has different columns related to cancer incidence data. However, all are not required by each aspect of prognosis. Therefore, some sort of pre-processing is needed in order to generate training and testing files separately for susceptibility, recurrence and survivability predictions. Therefore pre-processing results in training and testing datasets for the three kinds of prediction.

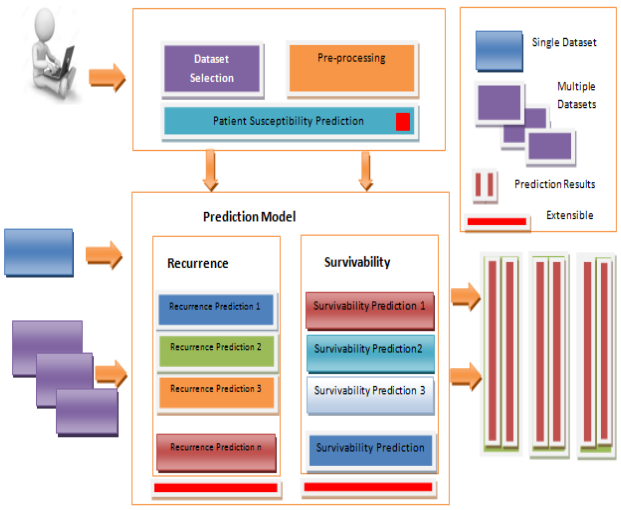
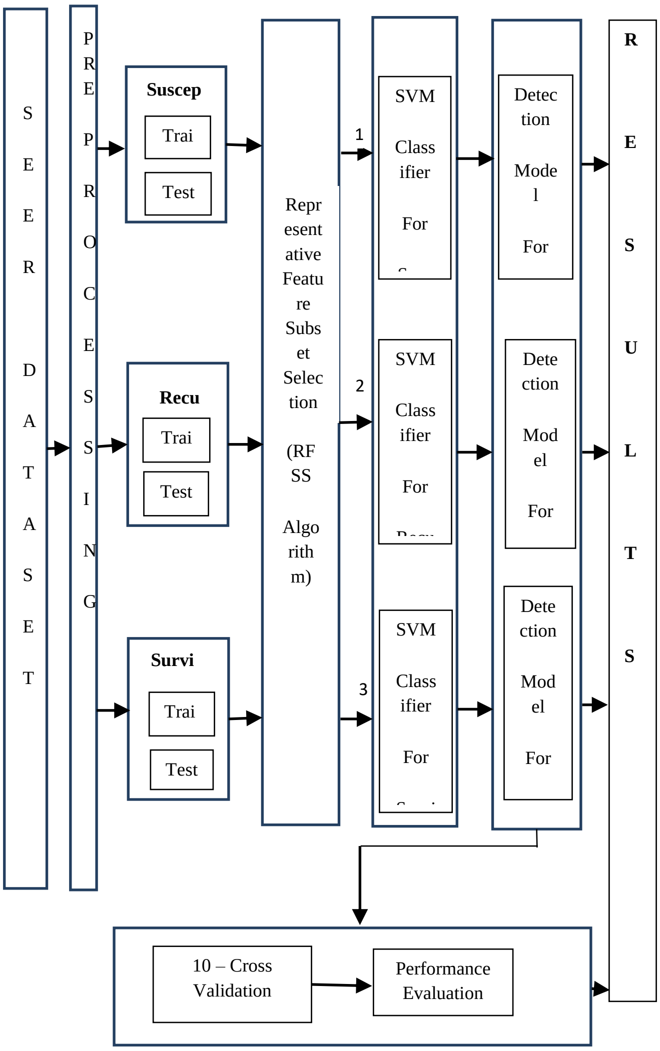


Figure 1: Overview of Proposed Framework XDBF



1. Features selected for susceptibility **2.** Features selected for recurrence **3.** Features selected for survivability

Figure 2: Overview of generating prediction models for susceptibility, recurrence and survivability predictions

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication (BEIESP)

The SEER dataset is considered for the research of breast cancer prognosis as it is the most reputed and reliable dataset in this area. The dataset is subjected to pre-processing in order o have training and testing datasets. This process is carried out for susceptibility, recurrence and survivability predictions. Then the training dataset is used for representative feature selection in order to enhance prediction accuracy of classifiers like SVM. Towards this end an algorithm is proposed as in Section 3.2. Three different SVM classifiers are trained. The first classifier is given features selected for susceptibility. The second classifier is given features selected for recurrence while the third classifier is given features selected for survivability. With representative features selected, the SVM classifiers learned provide corresponding detection models like detection model for susceptibility, detection model for recurrence and detection model for survivability. These models work on the corresponding testing sets in order to provide prediction results. The results include risk analysis results, recurrent prediction results and survivability prediction results. The classifiers are subjected to 10 – fold cross validation in order to evaluate performance. The results of performance of the proposed methodology and comparison with state of the art are presented in Section 5.

B. Representative Feature Subset Algorithms

```
Algorithm: Representative Feature Subset Algorithm
Inputs: Dataset D
Outputs: Chosen Feature Subset FS

Initialize gain threshold gt
Initialize entropy threshold et
Initialize feature subset vector FS
Initialize feature vector F
Initialize attributes vector A
Initialize concept c
Initialize tree T
Extract attributes of D into A based on c
Extract Relevant Features
For each attribute a in A
    Compute entropy e
    Compute gain g
    Associate a weight with a
    IF e>et and g>gt THEN
        Add a to F
    END IF
End For
Construction of Tree
For each f in F
    Add f to T
End For
Find Representative Features
For each node in T
    IF a feature pair is correlated THEN
        Add one feature to FS
    END IF
End For
Return FS
```

Algorithm 1: Representative feature subset selection algorithm

This algorithm extracts subset of features that are representative to reduce computational complexity besides reducing dimensionality to improve performance of machine learning algorithms use in breast cancer prognosis.

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x) \qquad (1)$$

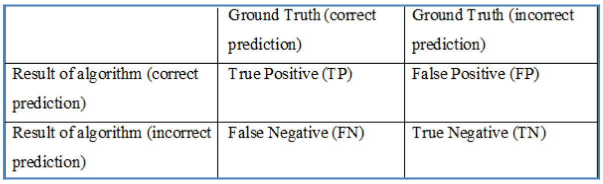
$$\text{Gain}\,(X/Y) = H(X) - H(X/Y) \qquad (2)$$
$$= H(Y) - H(Y/X)$$

Equations (1) and (2) are used to compute entropy and gain respectively. Entropy refers to disorder or uncertainty. Entropy and Gain are the two statistical measures used to make decisions pertaining to removal of irrelevant features and choosing relevant features that are relevant to the chosen concept. These two are also used to know correlation between two attributes in the given dataset. Entropy characterizes impurity of an attribute while the Gain is the expected reduction in entropy when examples are partitioned according to given attribute.

C. Evaluation Metrics

Evaluation of the proposed framework is evaluated using different metrics. The basis for the metrics is the confusion matrix presented in Table 1. It provides details of True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

Table 1: Confusion matrix



	Ground Truth (correct prediction)	Ground Truth (incorrect prediction)
Result of algorithm (correct prediction)	True Positive (TP)	False Positive (FP)
Result of algorithm (incorrect prediction)	False Negative (FN)	True Negative (TN)

From the confusion matrix, the following metrics are constructed. They are used to evaluate performance of proposed framework and compare with state of the art approaches.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

Accuracy is the measure used to know the ability of the algorithm to differentiate correct prediction from others. It makes use of all measures presented in confusion matrix. It is mathematically represented as in Eq. (3).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad (4)$$

The sensitivity is the measure used to know the ability of algorithm to determine risk/recurrence/survivability class correctly. It needs TP and FN values. This is mathematically represented as in Eq. (4).

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (5)$$

The specificity on the other hand is the measure used to know the ability of algorithm to determine a class which is not in risk/recurrence/survivability correctly. It needs TN and FP values. This is mathematically represented as in Eq. (3).

IV. DATASET COLLECTION

The Surveillance, Epidemiology, and End Results (SEER) are a program initiated by National Cancer Institute (NCI) of USA. As per this program cancer data is collected from time to time and that dataset is named as SEER dataset which contains cancer incidence and survival data. The dataset covers incidents from different countries and found to be highly reliable for cancer research. The registries of SEER dataset are updated from time to time with data of cancer patients including demographic, tumour morphology and stage, primary tumour site, first course of treatment and subsequent follow-ups for maintaining vital status. It is the population based dataset and considered comprehensive with population of different regions. The mortality data of SEER is provided by National Centre for Health Statistics (NCHS). Census Bureau periodically obtains cancer rates among population. Therefore SEER dataset is useful to general public, community groups, policy makers, legislators, healthcare units, clinicians and researchers. SEER dataset is collected from [40] by making a request and obtaining credentials to gain access to data. The dataset is of 1973-2015 November submission which includes difference cases pertaining to cancer research.

V. EXPERIMENTAL RESULTS

This section provides results of experiments and evaluation of the same. The results include susceptibility prediction, recurrence prediction and survivability prediction. In other words, the results constitute the results of breast cancer prognosis.

A. Results of Susceptibility Prediction

Results of risk prediction are presented here. The performance of the proposed SVM-RFSS method is compared with C4.5 and Naive Bayes algorithms.

Table 2: Experimental results of susceptibility

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
C4.5	98.09	54.78	59.86
Naïve Bayes	95.85	56.12	58.69
SVM-RFSS	98.90	57.54	58.48

As shown in Table 2, the accuracy, sensitivity and specificity are observed against proposed method SVM-RFSS and state of the art methods Naive Bayes and C4.5.

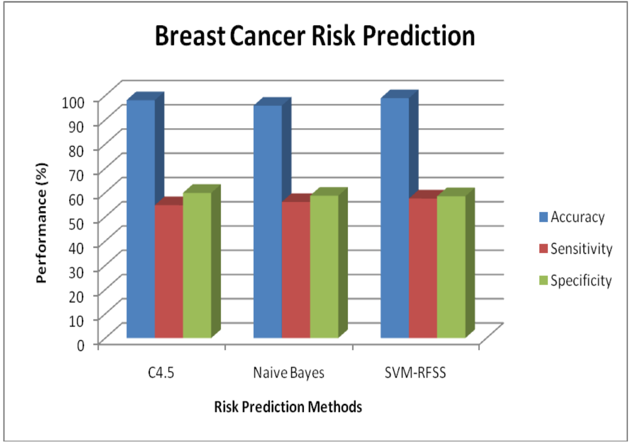


Figure 3: Performance comparison on breast cancer risk prediction

As presented in Figure 3, the risk prediction methods are presented in horizontal axis and the performance in terms of percentage of accuracy, sensitivity and specificity is presented on vertical axis. The results reveal that accuracy of the proposed SVM-RFSS is higher than other state of the art algorithms. The performance of C4.5 is better than that of Naive Bayes in terms of accuracy and specificity. Naive Bayes has shown better performance over C4.5 in terms of sensitivity.

B. Results of Recurrence Prediction

This subsection provides results of breast cancer recurrence. It shows comparison of performance among DT, SVM, ANN and SVM-RFSS.

Table 3: Results of breast cancer recurrence prediction

	Performance (%)		
Algorithm	**Accuracy**	**Sensitivity**	**Specificity**
DT	94.15	87.17	96.04
SVM	91.95	78.63	95.58
ANN	90.86	80.34	93.72
SVM-RFSS	97.14	89.5	96.9

As shown in Table 3, the performance of different algorithms for prediction of breast cancer recurrence is presented in terms of accuracy, specificity and sensitivity.

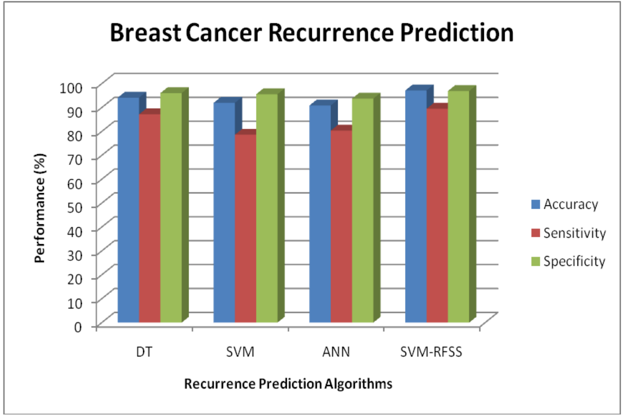


Figure 4: Performance comparison on breast cancer recurrence prediction

As presented in Figure 4, the recurrence prediction methods are presented in horizontal axis and the performance in terms of percentage of accuracy, sensitivity and specificity is presented on vertical axis. The results reveal that accuracy of the proposed SVM-RFSS is higher than other state of the art algorithms. The performance of DT is better than that of SVM and ANN. ANN has shown better performance over SVM in terms of sensitivity.

C. Results of Survivability Prediction

This section provides results of survivability prediction with 10-fold cross validation. The results of SVM-RFSS are compared against state of the art algorithms like MLP, DT Induction and LG.

Table 4: Results of breast cancer survivability in terms of accuracy

Fold No	Accuracy (%)			
	Neural Networks (MLP)	Decision Tree Induction (C5)	Logistic Regression	SVM-RFSS
1	0.9165	0.9363	0.8912	0.9564
2	0.9271	0.9403	0.8906	0.9678
3	0.9214	0.9356	0.8915	0.9518
4	0.9169	0.9356	0.8951	0.9651
5	0.9239	0.9348	0.8908	0.9456
6	0.8914	0.9363	0.89	0.9478
7	0.9081	0.9374	0.8949	0.9462
8	0.9024	0.9353	0.8894	0.9536
9	0.904	0.936	0.8923	0.9158
10	0.9098	0.9349	0.8938	0.9154

As shown in Table 4, the accuracy performance of breast cancer survivability is presented against 10 folds as in 10-fold cross validation.

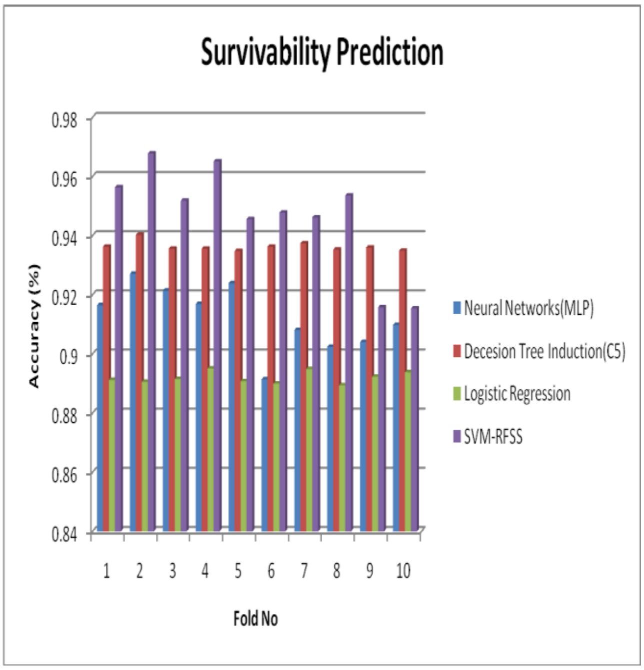


Figure 5: Accuracy of survivability prediction with 10-fold cross validation

As presented in Figure 5, the fold numbers in 10-fold cross validation are presented in horizontal axis while the accuracy (%) is presented in vertical axis for different algorithms like MLP, C5, LG and SVM-RFSS. The results revealed that performance of the proposed method SVM-RFSS has shown better performance due to the effectiveness of feature selection.

Table 5: Results of breast cancer survivability in terms of sensitivity

Fold No	Sensitivity (%)			
	Neural Networks (MLP)	Decision Tree Induction(C5)	Logistic Regression	SVM-RFSS
1	0.9535	0.9586	0.9015	0.8954
2	0.9578	0.9639	0.9018	0.8564
3	0.9537	0.9584	0.9011	0.8714
4	0.948	0.9582	0.9017	0.8956
5	0.9549	0.9607	0.9022	0.8835
6	0.9179	0.9604	0.8992	0.8657
7	0.929	0.9605	0.9037	0.8912
8	0.9442	0.9594	0.9001	0.8974
9	0.9454	0.9631	0.9036	0.8931
10	0.933	0.9591	0.9022	0.8967

As shown in Table 5, the sensitivity performance of breast cancer survivability is presented against 10 folds as in 10-fold cross validation.

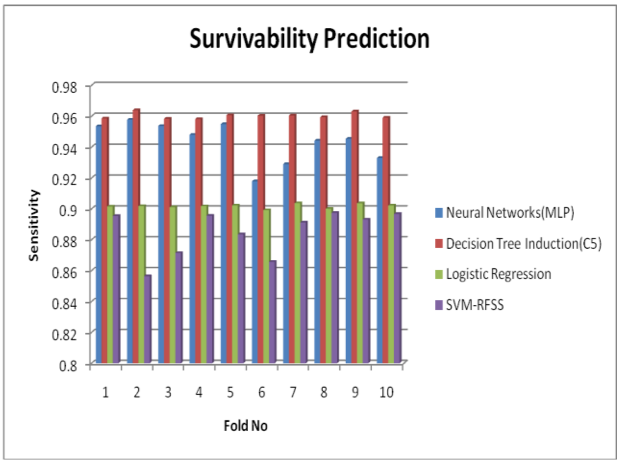


Figure 6: Sensitivity performance of survivability prediction

As presented in Figure 6, the fold numbers in 10-fold cross validation are presented in horizontal axis while the sensitivity (%) is presented in vertical axis for different algorithms like MLP, C5, LG and SVM-RFSS. The results revealed that performance of the proposed method SVM-RFSS has shown better performance due to the effectiveness of feature selection.

Table 6: Results of breast cancer survivability in terms of specificity

Fold No	Specificity (%)			
	Neural Networks (MLP)	Decision Tree Induction(C5)	Logistic Regression	SVM-RFSS
1	0.8719	0.9078	0.8766	0.8547
2	0.8905	0.9113	0.8756	0.8679
3	0.8818	0.9065	0.8779	0.8654
4	0.8798	0.9076	0.886	0.8735
5	0.8903	0.9021	0.8747	0.8694
6	0.8581	0.9062	0.8771	0.8638
7	0.8823	0.9095	0.8831	0.8728
8	0.8565	0.9068	0.8756	0.8642
9	0.8559	0.9027	0.8767	0.8756
10	0.8809	0.9052	0.8824	0.8751

As shown in Table 4, the specificity performance of breast cancer survivability is presented against 10 folds as in 10-fold cross validation.

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication (BEIESP)

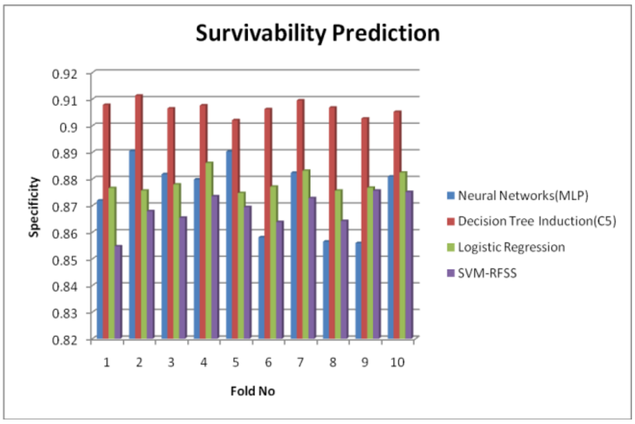


Figure 7: specificity performance of survivability prediction

As presented in Figure 7, the fold numbers in 10-fold cross validation are presented in horizontal axis while the specificity (%) is presented in vertical axis for different algorithms like MLP, C5, LG and SVM-RFSS. The results revealed that performance of the proposed method SVM-RFSS has shown better performance due to the effectiveness of feature selection. The results observed so far in Section5 showed that the SVM-RFSS provided better results. The rationale behind this is that feature subset selection avoids unnecessary processing of certain features thus reducing time and space complexity. This has resulted in the performance gain in the classification. Different classification algorithms are compared with the SVM-RFSS and found that the proposed method is capable of improving performance in breast cancer prognosis.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a framework for breast cancer prognosis. The framework is named as eXtensible Breast Cancer Prognosis Framework (XBPF). As the prognosis refers to prediction of breast cancer susceptibility, survivability and recurrence, this research assumes importance as it considers all aspects of the prognosis. As breast cancer is the second highest reason for cancer deaths in women across the globe, it is inevitable to take the research forward to contribute towards a solution. There has been considerable research and the results are significant as the recent developments helped increase the lifespan of breast cancer victims. In this paper, we proposed a framework to support breast cancer prognosis. The data used for experiments are collected from the real world cases present in the form of SEER datasets. Algorithms are explored towards realizing the framework for prediction of susceptibility, survivability and recurrence. We proposed a new prediction model that is the combination of RFSS and SVM for improving prediction performance. A prototype application is built to demonstrate proof of the concept. The empirical study revealed the utility of the framework. In future we intend to study ensemble approaches to understand possible synergy effect of using multiple approaches together stacked to have enhancements in breast cancer prognosis.

REFERENCES

1. K. Usha Rani . (2016).Perfomance Of Synthetic Minority Over Sampling Technique On Im Balanced Breast Cancer Data. IEEE, P1-5.
2. Cesar D. Velandia-Briñez, S.Don, Na-Yun Cho, Eunmi Choi And Dugki Min. (2009). Breast Cancer Image Classification Based On A Complex Event Processing Engine. IEEE, P60-64.
3. Qi Fan, Chang-Jie Zhu And Liu Yin. (2010). Predicting Breast Cancer Recurrence Using Data Mining Techniques. IEEE, P1-2.
4. Souad Demigha. (2015). Data Mining For Breast Cancer Screening. IEEE, P1-5.
5. B.K.Gayathri And P.Raajan. (2016). A Survey Of Breast Cancer Detection Based On Image Segmentation Techniques. IEEE, P1-5.
6. S. Muthuselvan. (2016). Prediction Of Breast Cancer Usingclassification Rule Mining Techniques In Blood Test Datasets. IEEE, P1-5.
7. Aaron N. Richter And Taghi M. Khoshgoftaar. (2016). Predicting Cancer Relapse With Clinical Data: A Survey Of Current Techniques. IEEE, P1-8.
8. Huseyin Seker, Michael O. Odetayo, Dobrila Petrovic, And Raouf N. G. Naguib. (2003). A Fuzzy Logic Based-Method For Prognostic Decision Making In Breast And Prostate Cancers. Ieee. 7 (2), P1-9.
9. Hemant Palivela, Yogish H K, Vijaykumar S And Kalpana Patil. (2013). Survey On Mining Techniques For Breast Cancer Related Data. IEEE, P1-7.
10. Runjie Shen, Yuanyuan Yang And Fengfeng Shao. (2014). Intelligent Breast Cancer Prediction Model Using Data Mining Techniques. IEEE, P1-4.
11. Joseph A. Cruz And David S. Wishart(2006). Applications Of Machine Learning In Cancer Prediction And Prognosis. Cruz And Wishart. P1-20.
12. Faezeh Roshani, M.H. Fazel Zarandi, I.B. Turksen And Maede Maftooni. (2015). Fuzzy Expert System For Prognosis Of Breast Cancer Recurrence. IEEE, P1-5.
13. Jaree Thongkam , Guandong Xu, Yanchun Zhang And Fuchun Huang. (2007). Breast Cancer Survivability Via Adaboost Algorithms. Australian Computer Society. P1-10.
14. [14]Samuel Giftson Durai, Dr. S. Hari Ganesh And A. Joy Christy. (2017). Novel Linear Regressive Classifier For The Diagnosis Of Breast Cancer. IEEE, P1-4.
15. [15]. Jaree Thongkam , Guandong Xu, Yanchun Zhang And Fuchun Huang. (2007). Toward Breast Cancer Survivability Prediction Models Through Improving Training Space. Expert Systems With Applications P1-10.
16. [16]Sourajit Behera And Rinkle Rani (2016). Comparative Analysis Of Density Based Outlier Detection Techniques On Breast Cancer Data Using Hadoop And Map Reduce. IEEE, P1-4.
17. Uma Ojha And Dr. Savita Goel. (2017). A Study On Prediction Of Breast Censer Recurence Using Data Mining Technique . IEEE, P1-4.
18. Ilias Maglogiannis , Elias Zafiropoulos And Ioannis Anagnostopoulos. (2007). An Intelligent System For Automated Breast Cancer Diagnosis And Prognosis Using Svm Based Classifiers. Springer Science. P1-13.
19. Juhyeon Kim And Hyunjung Shin. (2013). Breast Cancer Survivability Prediction Using Labeled, Unlabeled, And Pseudo-Labeled Patient Data.Ieee. P1-6.
20. Thunyarat Anothaisintawee , Yot Teerawattananon , Chollathip Wiratkapun , Vijj Kasamesup And Ammarin Thakkinstian. (2012). Risk Prediction Models Of Breast Cancer: A Systematic Review Of Model Performances. Breast Cancer Res Treat. P1-10.
21. Muhammad Umer Khan, Jong Pill Choi, Hyunjung Shin And Minkoo Kim . (2008). Predicting Breast Cancer Survivability Using Fuzzy Decision Trees For Personalized Healthcare. IEEE, P20-24.
22. Dr. Medhat Mohamed, Muhamed Wael Farouq,Hala Abou And Abdel Badeeh.(). Using Data Mining For Assessing Diagnosis Of Breast Cancer. Computer Science And Information Technology,P1-7.
23. Ahmed Iqbal Pritom, Shahed Anzarus Sabab And Md. Ahadur Rahman Munshi. (2016). Predicting Breast Cancer Recurrence Using Effective Classification And Feature Selection Technique. IEEE, P1-5.
24. Hyunjung Shin And Yonghyun Nam. (2014). A Coupling Approach Of A Predictor And A Descriptor For Breast Cancer Prognosis. Bmc, P1-12

25. Aaron N. Richter And Taghi M. Khoshgoftaar. (2016). Predicting Cancer Relapse With Clinical Data: A Survey Of Current Techniques. IEEE, P1-8.
26. Umer Khan, Hyunjung Shin, Jong Pill Choi And Minkoo Kim. (2008). Wfdt - Weighted Fuzzy Decision Trees For Prognosis Of Breast Cancer Survivability, P1-12.
27. B.Padmapriya And T.Velmurugan (2014). Survey On Breast Cancer Analysis Using Data Mining Techniques. IEEE, P1-4.
28. Dursun Delen, Glenn Walker, Amit Kadam(2005). Predicting Breast Cancer Survivability:
29. A Comparison Of Three Data Mining Methods.Computer Science ,P1-15.
30. Abien Fred M. Agarap.(2018). On Breast Cancer Detection: An Application Of Machine Learning Algorithms On The Wisconsin Diagnostic Dataset.ACM, P1-5.
31. Sara W. Dyrstad, Yan Yan , Amy M. Fowler And Graham A. Colditz. (2015). Breast Cancer Risk Associated With Benign Breast Disease: Systematic Review And Meta-Analysis. Springer, P1-7.
32. Dana Bazazeh And Raed Shubair.(2016). Comparative Study Of Machine Learning Algorithms For Breast Cancer Detection And Diagnosis. IEEE, P1-5.
33. Noa Eyal, Mark Last And Eitan Rubin.(2015). Comparison Of Three Classifiers For Breast Cancer Outcome Prediction, Acm,P1-5.
34. Smita Jhajharia,Seema Verma And Rajesh Kumar. (2016). Predictive Analytics For Breast Cancer Survivability: A Comparison Of Five Predictive Models. Acm, P1-5.
35. Pedro Henriques Abreu , Miriam Seoane Santos,Miguel Henriques Abreu,Bruno Andrade And Daniel Castro Silva. (2016). Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. ACM Computing Surveys. 49 (3), P1-40.
36. Limin Wang. (2015). Mining Causal Relationships Among Clinical Variables For Cancer Diagnosis Based On Bayesian Analysis. Wang Biodatamining, P1-15.
37. Doreswamy And Umme Salma M. (2015). Fast Modular Artificial Neural Network For The Classification Of Breast Cancer Data. ACM, P1-7.
38. Chulwoo Pack,Sung Shin,Seong Ho Son, Soon Ik Jeon. (2015). Computer Aided Breast Cancer Diagnosis System With Fuzzy Multiple-Parameter Support Vector Machine. ACM, P1-5.
39. Johanna Johnsi Rani G, Dennis G And Joy Mammen.(2015). Classification And Prediction Of Breast Cancer Derived Using Natural Language Processing. Acm,P1-5.
40. Manisha Arora And Dinesh Tagra. (2012). Neuro-Fuzzy Expert System for Breast Cancer Diagnosis. ACM, P1-7.
41. Seer Incidence Database (2018). Options for Accessing Daa and SEER*Stat Software. Retrieved from https://seer.cancer.gov/data/options.html.