International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 32035-32045

http://www.ripublication.com

A Survey on Various Predictability and Survivability Factors in Breast Cancer Using Data Mining and Soft Computing Techniques

K.Sangeetha

SNS College of Technology Coimbatore
sangithaprakash@gmail.com

P.S.Periasamy

K.S.R College of Engineering Tiruchengode
psperiasamy.ps@gmail.com

S.Prakash

Sri Shakthi Institute of Engineering and Technology Coimbatore
prakashdharsan@gmail.com

Abstract

In the field of Medical science, there was a vast technological improvement. Even though, some of the dreadful disease causing factors was not predicted in earlier stages**.**Predicting the outcome of a disease is one of the most interesting and challenging tasks where data mining techniques have to be applied. The Main aim of this paper is to identify the most effective way to reduce cancer deaths by detecting it earlier. The Classification technique is deployed with the various data mining factor to yield the better prediction rate. Cancer is a dreadful disease which kills several thousand people life. If it is predicted earlier based on the food habits, age, sex, and other risk factors the death rate can be still reduced. The supervised technique is used to classify the risk causing factors and the association rule mining is used to build the rules which helps for easier prediction. In the future, classification is done and ontology framework will be developed, which may lead to better results in prediction.

Keywords: Classification, Decision tree, Association Rule Mining, Prediction, Risk factors, Artificial Neural Networks, Naïve Bayes, Genetic Algorithm, survivability, predictability

INTRODUCTION

Cancer is one of the dangerous dieses which cause death in both the gender. The main objective of this proposal is to identify the causes of cancer patterns that affect the female groups. This work aims to focus on identifying the various breast cancer predictability model and survivability model used in the field of medical data mining. Further we are interested to make an analysis by grouping the people surviving in the urban and rural area by considering their age groups of less than 10 years, 10-20, 21-30, 31-40 and above 40 Years. The importance is to identify the age groups which are likely to be affected by the disease and the age groups that can adhere to the survival.The parameters like food habits of children's and adults, Radiations, hazardous chemicals, lack of physical activities, overweight, alcohol intake, soft drinks, junk foods, and food items which includes more carbohydrates. The statistical collection has to be made based on the education status, the nature of job, the stress level and modern life style.. The cancer pattern is to classified based on numerous attributes, which includes the various habits of human. This particular classification is going to be carried out effectively based on different data mining techniques and tools [18]. The prediction plays a major role to identify the vulnerability of a cancer in future. The cancer has become leading causes of death. This statistical analysis in cancer research provides valuable suggestions and information to organize the cancer control programmes. It also helps to create the awareness in the society.The Idea is to collect the real time data set from Southern part of Coimbatore region and need to apply the following methodology to identify the parameters of breast and lung cancer. This review contains the various Predictability and survivability methods based on Classification and soft computing techniques. It contains three sections, in which after this preface section follows two other sections. In the section of related works explanation about the data mining techniques and its issues is discussed. Further the final section describes the comparative analysis.

RELATED WORK:

If-Then Rules & Naïve Bayes:

In [13], Krishnaiah.V et al proposed a lung cancer disease prediction system. They used the efficient data mining classification technique in order to extract the disease knowledge from the database. The prediction is made using the Naive Bayes with IF-THEN Rules. Whereas the attribute relationships produced by the Neural Networks are quite difficult and vague.The Enhanced system can be built using the clustering and association rules, categorical data, mining unstructured data.

Decision Tree:
In [11], S. Jothi, S.Anita proposed a predictive model based on decision tree, a classification approach. The classification aims to build models which are used in predicting the class of objects. The main advantage is that the generated rules are easy to understand. The main drawback is that, comparatively they do not produce better performance. Linear Regression can be used for Prediction tasks as it yields more approximate results.

Logistic Regression:
In [14], Kung-Min Wang et al proposed survival prediction models using the data mining methods like logistic regression and decision tree. These methods of data mining are used to investigate the survivability of breast cancer patients. The result obtained using the logistic regression outperforms the decision tree. The logistic regression does not suffer from Multicollinearity whereas the decision trees are affected by Multicollinearity problem.

ANN, Decision tree, logistic regression:
In [7], Dursun Delen et al reported on the survivability model for breast cancer. They focused on processing huge volume of data which is available from the database. They used three different types of classification models: artificial neural network (ANN), decision tree, and logistic regression along with a 10-fold cross-validation technique to compare the accuracy of these classification models. The extended scope is to develop and use the model as a web-based decision support system (DSS) to provide the accurate prediction models based on hybrid approach. They used the SEER data's twenty variables. The decision tree accuracy and ANN accuracy were found more superior to logistic regression accuracy.

ANN & GA:
In [4], Chang Pin-Wie et al conducted a study on which they investigated the Artificial intelligence and data mining techniques to the prediction models of breast cancer. The artificial neural network, decision tree, logistic regression, and genetic algorithm were used. They focused on the accuracy and positive predictive values.Based on the investigation the accuracy of the genetic algorithm was significantly higher. Similarly the outcome of the logistic regression model was higher than that of the neural network model. They explored that based on the parameters like overall accuracy, the expression and complexity of the classification rule, results of breast cancer patients obtained using genetic algorithm outperforms than the data mining models. The accuracy and positive predictive value of each algorithm were used as the evaluation indicators.

Functional Classifiers:
In [20], Shelly Gupta et al focused the research using the data mining techniques to enhance the breast cancer diagnosis and prognosis. The model was based on various neural network structures like Multi-Layer Perceptrons (MLP), Probabilistic Neural Network (PNN), Self Organizing Map (SOM) and Radial Basis Function. They

formulated the functional classifiers for breast cancer Observed that the accuracy for the diagnosis analysis of various applied data mining classification techniques is highly acceptable and can help the medical professionals in decision making for early diagnosis. The prognostic problem is mainly analyzed under ANNs and its accuracy came higher in comparison to other classification techniques for the same dataset. Among these classifiers, RBF and PNN classify the tumors accurately. Decision tree, Support Vector Machines are other well known classification methods, they are implemented by constructing tree forest with the given dataset. If the ROI curve has the largest area then SVM shows good performance results when compared to other methods.

Neural Networks:
In [19], Soltani Sarvestani.A, et al provided a comparison among the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map (SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN) which are used to classify WBC and NHBCD data. The performance of these neural network structures was investigated for breast cancer diagnosis problem. RBF and PNN were proved as the best classifiers in the training set. But the PNN gave the best classification accuracy when the test set is considered.

Decision Tree:
In [2], Anunciacao Orlando et al explored the applicability of decision trees for detection of high-risk breast cancer groups over the dataset produced by Department of Genetics of Universidade Nova de Lisboa. To statistically validate the association found, permutation tests were used. They found a high-risk breast cancer group composed of 13 cases and only 1control, with a Fisher Exact Test (for validation). These results showed that it is possible to find statistically significant associations with breast cancer by deriving a decision tree.

Support Vector Machine:
In [1], Abdelaal Ahmed Mohamed Medhat et al investigated the capability of the classification SVM with Tree Boost and Tree Forest in analyzing the DDSM dataset for the extraction of the mammographic mass features along with age that discriminates true and false cases. Here, SVM techniques show promising results for increasing diagnostic accuracy of classifying the cases comparable to values for tree boost and tree forest.

Particle Swarm Optimization Algorithm:
In [8], Rajiv Gandhi.K et al in their work constructed classification rules using the Particle Swarm Optimization Algorithm (PSOA) for breast cancer datasets. The problem of feature subset selection as a pre-processing step was used which learns fuzzy rules bases using GA implementing the Pittsburgh approach.The proposed Particle Swarm Optimization Algorithm (PSOA) follows genetic algorithm method of classification. During this pre-processing stage fuzzy rules were framed that accumulated smaller number of rules with high accuracy. This method provides good

results when compared to other classification methods. They opted for a higher accuracy system with less fuzzy rules. The datasets obtained after the after feature selection are used in PSOA. The developed rules met the effective rate of accuracy.

Multilayer Perceptron & Radial Basis Function:
In [17], Padmavati.J performed a comparative study on WBC dataset for breast cancer prediction using RBF and MLP along with logistic regression. It was observed that neural networks took slightly higher time than logistic regression but the sensitivity and specificity of both neural network models had a better predictive power over logistic regression. When comparing RBF and MLP neural network models, it was found that RBF had good predictive capabilities and also time taken by RBF was less than MLP.

Rough Set Theory:
In [15], Lee Chul-Heui et al (2001) in their study proposed a new classification method based on the hierarchical granulation structure using the rough set theory. The hierarchical granulation structure was adopted to find the classification rules effectively. The classification rules had minimal attributes and the knowledge reduction was accomplished.

Rough Set Reduction Technique:
In [9], Aboul Ella Hassanien, and Jafar M.H.Ali presented simplification algorithm based on a rough set method for generating less classification rules from the observed samples. The attributes were selected, normalized and then the rough set dependency rules were generated directly from the real value attribute vector. Then the rough set reduction technique was applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. They showed that the total number of generated rules was reduced from applying the proposed simplification algorithm. They also made a comparison between the obtained results of rough sets with the well known ID3 decision tree and concluded rough sets showed higher accuracy and generated more compact rules.

SVM & ANN:
In [22], Sudhir D. Sawarkar et al applied SVM and ANN on the WBC data.The results of SVM and ANN prediction models were found comparatively more accurate than the human being. The 97% high accuracy of these prediction models can be used to take decision to avoid biopsy.

ANN and Multiwavelet Technique:
In [10], Sepehr M. H. Jamarani et al presented an approach for early breast cancer diagnosis by applying combination of ANN and multiwavelet based sub band image decomposition. The proposed approach was tested using the mammographic databases and images collected from local hospitals. Yields best performance. The proposed approach could assist the radiologists in mammogram analysis and diagnostic decision making.

Decision tree induction learning Technique:
In [3], Bellaachia Abdelghani and Erhan Gauven explored a study based on C4.5 which is decision tree induction learning technique, Naive Bayes and Back-Propagated Neural Network method. They presented an analysis of the prediction of survivability rate of breast cancer patients using above mentioned method.The analysis was carried out based on the SEER data set with parameters like STR(Survival Time Recode), VSR(Vital Status Recode), and COD(Cause Of Death).The C4.5 algorithm outperforms the Naive Bayes and Back-Propagated Neural Network.

ANN:
In [5], Chih-Lin Chi et al used the ANN model for Breast Cancer Prognosis on WPBC data and Love data. In their research they used recurrence at five years as a cut point to define the level of risk. The applied models successfully predicted recurrence probability and separated patients with good(>5 yrs) and bad(<5 yrs) prognosis.

Hybrid Network:
In [6], Choi Pill Jong et al explored the performance of an Artificial Neural Network (ANN), a Bayesian Network (BN) and a Hybrid Network (HN) which is used to predict breast cancer prognosis. The hybrid Network is developed by combining both ANN and Bayesian Network. Compared to the accuracy of the three methods ANN, and Hybrid Network, outperforms the Bayesian Network. The proposed Hybrid model can be used to take better decisions.

Fuzzy Decision Trees:
In [12], Muhammad Umer Khan et al investigated a hybrid scheme based on fuzzy decision trees on SEER Data. They conducted experiments using various combinations of decision tree rules, different types of fuzzy Membership functions and inference techniques. They compared the performance of each for cancer prognosis and concluded that hybrid fuzzy decision tree classification is more unique and balanced than the independently applied crisp classification.

Neural Network:
In [21], Shewta Karya et al explored that they outperformed the Neural Network (NN) classifiers in the prediction process of breast cancer. The Markov Blanket and Tree augmented Naive Bayes are different aspects of BBN. The Markov Blanket Estimation creates a Bayesian network by framing the relationship between attributes that are independent. It is based on the concept that every attribute is present in the classifier as the class attribute. The tree augmented Naive Bayes is an enhanced approach of Naive Bayes which allows inter dependency between attributes along with relationship between target attribute. The classification accuracy increased considerably using these two methods.

COMPARATIVE ANALYSIS

In this section the pros and cons involved in using several techniques is outlined.

Table1. Pros and Cons of Several Mining techniques

Author	Proposed Mining Technique	Identified Pros and Cons
V. Krishnaiah et al (2013)	Lung cancer prediction system	Extracts the hidden knowledge form the database. Difficult to understand. Better enhanced by using association and clustering.
S. Jothi, S.Anita (2012)	Predictive model based on decision trees	Aims to build models which are used in predicting the class of objects. Better works for all types of cancer. They do not produce better performance. Linear Regression can be used for Prediction tasks as it yields more approximate results.
Kung-Min Wang et al (2012)	Survival prediction models using logistic regression and decision tree.	Logistic regression will be prone to use high correlation variables and solve the problem of Multicollinearity. Decision trees are not well suited and leads to Multicollinearity problem
Dursun Delen et al (2005)	Survivability model for breast cancer using Artificial Neural Networks (ANN), decision tree, and logistic regression	Provides better performance. The Present model does not support for highly accurate system. Not suitable for all types of cancer data sets. Hybrid models can be developed.
Wei-Pin et al	Investigation on prediction models of breast cancer using Artificial neural network(ANN), decision tree, logistic regression, and Genetic Algorithm(GA)	Accuracy of the genetic algorithm was significantly higher. The logistic regression does not outperforms the ANN. Enhancement can be made using the two-stage hybrid model

Shelly Gupta et al (2011)	Proposed a model based on various neural network structures like Multi-Layer Perceptrons (MLP), Probabilistic Neural Network (PNN), Self Organizing Map (SOM) and Radial Basis Function.	RBF and PNN classifiers classify the tumors accurately. Based on either diagnosis or prognosis.
Soltani Sarvestani et al	Analysed the Multilayer Perceptron (MLP), Self Organizing Map(SOM), Radial Basis Function (RBF) and Probabilistic Neural Network(PNN)	Radial Basis Function and Probabilistic Neural Network are identified as best classifiers in training set. Radial basis Function does not outperforms in test set.
Orlando Anunciacao et al (2010)	Used decision trees for detection of high-risk breast cancer	Helps to find statistically significant associations with breast cancer by deriving a decision tree.
Medhat Mohamed Ahmed Abdelaal et al (2010)	Used classification SVM with Tree Boost and Tree Forest	SVM improves the classification accuracy in diagnosis
K. Rajiv Gandhi et al (2010)	Constructed the classification rules using particle swarm optimization(PSO)	It was used to produce a smaller fuzzy rule based system with higher accuracy
J. Padmavati (2011)	Performed a study on WBC dataset using logistic regression.	Neural networks took slightly higher time than logistic regression. Sensitivity is low in logistic regression.
Chul-Heui Lee et al (2001)	hierarchical granulation structure using the rough set theory	Made the information analysis easier Used only minimal attributes
Aboul Ella Hassanien et al (2004)	a rough set method for generating classification rules	Results in Higher accuracy and generated more compact rules.
Sudhir D. Sawarkar et al (2006)	Applied both SVM and ANN techniques	Provides the prediction models with more accuracy rate.
Sepehr M. H. Jamarani et al (2005)	Early breast cancer diagnosis by applying combination of ANN and Multiwavelet based sub band image decomposition	The best performance was achieved

Bellaachia Abdelghani and Erhan Gauven (2006)	Prediciton survivability based on C4.5(decision algorithm) Naive Bayes and Back-Propagated Neural Network	Model generated by C4.5 algorithm for the given data has a much better performance than the other two techniques
Chih-Lin Chi et al (2007)	ANN model for Breast Cancer Prognosis	Better predicts the recurrence probability and classify as good and bad
Jong Pill Choi et al (2009)	performance of an Artifical Neural Network, a Bayesian Network and a Hybrid Network used to predict breast cancer prognosis	Accuracy of ANN(88.8%) and Hybrid Network(87.2%) were very similar and they both outperformed the Bayesian Network
Muhammad Umer Khan et al (2008)	Hybrid scheme based on fuzzy decision tree approach	Hybrid fuzzy decision tree classification produces Unique results.
Shewta Karya et al (2008)	Developed a model based on the Neural Network classifier for breast cancer prediction	Accuracy improved while using the Naive Bayes and the Markov Blanket

CONCLUSION

This paper provides a study of various technical and review papers on breast cancer predictability and survivability models for the past one decade and explores that data mining techniques when combined with neural networks or with genetic algorithm, better results are produced which assists in decision making.

REFERENCES

[1] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed, 2010 “Using data mining for assessing diagnosis of breast cancer, ” *in Proc. International multiconfrence on computer science and information Technology*,, pp. 11-17.

[2] Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and Rueff Jose,, 2010 “A Data Mining approach for detection of high-risk Breast Cancer groups, ” *Advances in Soft Computing*, vol. 74, pp. 43-51.

[3] Bellaachia Abdelghani and Erhan Guven, 2006 "Predicting Breast Cancer Survivability using Data Mining Techniques, " *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining.*

[4] Chang Pin Wei and Liou Ming Der, “Comparision of three Data Mining techniques with Genetic Algorithm in analysis of Breast Cancer data”[Online]. Available:http://www.ym.edu.tw/~dmliou/Paper/compar_threedata.pdf

[5] Chi C.L., Street W.H. and Wolberg W.H. 2007. "Application of Artifical Neural Network- based Survival Analysis on Two Breast Cancer Datasets", *Annual Symposium Proceedings / AMIA Symposium*.

[6] Choi J.P., Han T.H. and Park R.W., 2009, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis", *JKorean Soc Med Inform*, pp. 49-57.

[7] Delen Dursun, Walker Glenn and Kadam Amit June 2005, "Predicting breast cancer survivability: a comparison of three data mining methods, " *Artificial Intelligence in Medicine*, vol. 34, pp. 113-127.

[8] Gandhi Rajiv K., Karnan Marcus and Kannan S. 2010, "Classification rule construction using particle swarm optimization algorithm for breast cancer datasets, " *Signal Acquisition and Processing. ICSAP, International Conference*, pp.233 – 237.

[9] Hassanien Ella Aboul and Ali H.M. Jafar, 2004, "Rough set approach for generation of classsification rules of Breast cancer data, " *Journal Informatica*, vol. 15, pp. 23–38.

[10] Jamarani S. M. H., Behnam H. and Rezairad G. A., 2005, "Multiwavelet Based Neural Network for Breast Cancer Diagnosis", *GVIP 05 Conference*, pp. 19-21.

[11] Jothi.S, Anita.S, October - 2012 "Data Mining Classification Techniques Applied For Cancer Disease – A Case Study Using Xlminer"International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, ISSN: 2278-0181 1

[12] Khan M.U., Choi J.P., Shin H. and Kim M, 2008, "Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare", *Conf Proc IEEE Eng Med Biol Soc.*, pp. 48-51.

[13] Krishnaiah.V, Narsimha.G, Subhash Chandra,, 2013 "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1), 39 – 45

[14] Kung-Min Wang, Bunjira Makond, Wei-Li Wu, K.-J. Wang, Y. S. Lin June 2012, *International Journal of Innovative Management, Information & Production.*

[15] Lee Heui Chul, Seo Hak Seon and Choi Chul Sang, 2001, "Rule discovery using hierarchial classification structure with rough sets, " *IFSA World Congress and 20th NAFIPS International Conference*, vol.1, pp. 447-452.

[16] Lundin M., Lundin J., BurkeB.H., Toikkanen S., Pylkkanen L. and Joensuu H., 1999, "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", *Oncology International Journal for Cancer Resaerch and Treatment*, vol. 57.

[17] Padmavati J., Jan. 2011 "A Comparative study on Breast Cancer Prediction Using RBF and MLP, "International Journal of Scientific & Engineering Research, vol. 2.

[18] Ramachandran.P, Girija.N, Bhuvaneswari.T, June 2013 "Cancer Spread Pattern – an Analysis using Classification and Prediction Techniques"

International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6.

[19] Sarvestan Soltani A., Safavi A. A., Parandeh M. N. and Salehi M.,, 2010 "Predicting Breast Cancer Survivability using data mining techniques, " Software Technology and Engineering (ICSTE), 2nd International Conference, vol.2, pp.227-231.

[20] Shelly Gupta, Dharminder Kumar, Anand Sharma, Apr-May 2011 "Data mining classification techniques applied for breast cancer diagnosis and prognosis", Indian Journal of computer Science and Engineering ISSN: 0976-5166, Vol. 2 No, 2.

[21] Shweta Kharya, April 2012 "Using Data Mining Techniques For Diagnosis and Prognosis of Cancer Disease" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2,, DOI: 10.5121/ijcseit.2012.2206.

[22] Sudhir D., Ghatol Ashok A., Pande Amol P. 2006, "Neural Network aided Breast Cancer Detection and Diagnosis", 7 th WSEAS International Conference On Neural Networks,.