Intelligent Environments 2021
E. Bashir and M. Luštrek (Eds.)

doi:10.3233/AISE210098

Prediction of Breast Cancer Using AI-Based Methods

Sanam AAMIR [a], Aqsa RAHIM [a], Sajid BASHIR [b] and Muddasar NAEEM [c]
[a] *National University of Science And Technology*
[b] *National University of Technology*
[c] *ICAR-CNR*

Abstract. Breast cancer has made its mark as the primary cause of female deaths and disability worldwide, making it a significant health problem. However, early diagnosis of breast cancer can lead to its effective treatment. The relevant diagnostic features available in the patient's medical data may be used in an effective way to diagnose, categorize and classify breast cancer. Considering the importance of early detection of breast cancer in its effective treatment, it is important to accurately diagnose and classify breast cancer using diagnostic features present in available data. Automated techniques based on machine learning are an effective way to classify data for diagnosis. Various machine learning based automated techniques have been proposed by researches for early prediction/ diagnosis of breast cancer. However, due to the inherent criticalities and the risks coupled with wrong diagnosis, there is a dire need that the accuracy of the predicted diagnosis must be improved. In this paper, we have introduced a novel supervised machine learning based approach that embodies Random Forest, Gradient Boosting, Support Vector Machine, Artificial Neural Network and Multilayer Perception methods. Experimental results show that the proposed framework has achieved an accuracy of 99.12%. Results obtained after the process of feature selection indicate that both preprocessing methods and feature selection increase the success of the classification system.

Keywords. Breast Cancer, WDBC, RFE, k-NN, Support Vector Machines, Random Forest, Decision Tree, ANN, MLP, Machine Learning

1. Introduction

In the modern world, individuals are more vulnerable to cancer than they have ever been. Cancer is a fatal disease that is present worldwide. Approximately 9.6 million people died because of cancer in 2018. One out of six deaths are caused by cancer globally [36]. Nearly 70% of all cancer-related deaths occur in middle-income and low-income countries. Other reasons contributing to cancer-related deaths are low fruit and vegetable consumption, body mass index, lack of physical activity and alcohol use. A research estimated the death of approx. 40,920 women in 2018 due to breast cancer alone [36]. According to statistics of the World Health Organization (WHO), 2.09 million women are diagnosed with breast cancer every year [1]. As is the case with any type of cancer, early diagnosis is the only cure to breast cancer as well. Researchers and scientists have conducted a variety of experiments for early detection of breast cancer, so that it can be

effectively cured and risks to patients' life may be significantly reduced.

Cancer is a term that is used for a large group of diseases that affect various parts of the human body. Cancer is mainly characterized by spread of abnormal cells rapidly. This spread of abnormal cells is so rapid that cells go beyond their normal limits and invade adjacent parts of the body and spread to other organs which causes deterioration finally leading to death. This process is known as metastasis [1]. The main cause of cancer-related deaths is metastasis. Other terms used for cancer are malignant tumors and neoplasms [1]. Breast cancer in women has an extremely high mortality rate. In Breast Cancer, rapidly dividing cells form breast masses in breast cancer. Such masses are named as tumors that are malignant (cancerous) of benign (non-cancerous) Malignant Tumors penetrate healthy tissues in the body and cause damage. Cancer spreads when malignant cells infect healthy cells, and these malignant cells spread very quickly [16]. Therefore, it is very important that not only cancer is diagnosed, but also that it is diagnosed in its early stage so that it can be cured. Evolution in the field of technology and medicine has led to availability of a large amount of data that is stored and provided to researchers that make innovative use of several techniques for detection of the disease.

The main challenge is to detect if the tumor is benign or malignant. Various models have been developed so for detection of breast cancer. The models have used risk factors, blood analysis data and features extracted from x-ray images to detect cancer. Features are extracted from Breast Cancer mammograms (x-ray images), that include many attributes i.e. Clump Thickness, Marginal Adhesion, Uniformity of Cell Size/ Shape, Single Epithelial Cell Size, Bland Chromatin, Normal Nucleoli, Bare Nuclei and Mitosis. They are normally used for the diagnosis and detection of breast cancer. Other researches have used risk factors i.e. age, number of previous biopsies, race, number of first-degree relatives affected with breast cancer, not only to predict breast cancer at the first instance but to predict its recurrence as well. Researchers have also used blood analysis data, including BMI, age, HOMA, Glucose, Leptin, Resistin, MCP.1 and Adiponectin to diagnose Breast Cancer using Machine Learning techniques. Other information that has been utilized for aiding diagnosis of Breast Cancer include race/ethnicity, pregnancy history, breastfeeding history, being overweight, exposure of chest or face to radiation before the age of 30, exposure to chemicals, low vitamin D levels, menstrual history and lack of exercise.

Recently, the development in computing technology and the introduction of new machine learning algorithms e.g. reinforcement learning [23], neural network [21] the goal of Artificial Intelligence (AI) has become a step closer. AI has important application in diverse fields including:healthcare [12], [22], robotics and autonomous control, vision enhancing method for low vision impairments [19], natural language processing, dynamic normative environments [31], risk management [26], intelligent environments [10], games and self-organized system [11], ambient assisted living techniques to improve the quality of life of elderly [13], Social humanoid robot [9] can help to monitor indoor environmental quality [29] and distributed fuzzy system able to infer in real-time critical situations [30].

In this paper, we propose a novel approach for the diagnostic prediction of Breast Cancer by careful feature selection and data handling. The diagnostic characteristics from the WDBC dataset [5] are used. Our aim is to predict the tumor as malignant or benign with a reduced set of features and improved accuracy.

The remaining part of the paper is organized as follows: section 2 discusses literature, section 3 discusses the proposed work of this study, section 4 explains the methodology

and experimentation, section 5 gives detailed information about the experimental results and analyzes the results of the experiment and provides comparative analysis with previous studies, section 6 discusses results of application of the framework on other datasets and section 7 concludes the work.

2. Literature Review

Cancer spreads when malign cells infect healthy cells, and these malign cells spread very quickly [1]. Therefore, it is very important that not only cancer is diagnosed, but also that it is diagnosed in its early stage so that it can be cured. Evolution in the field of technology and medicine has led to availability of a large amount of data that is stored and provided to researchers that contributed to innovative use of several techniques for detection of the symptoms of the disease. A challenge that radiologists usually face is that after the tumor is discovered, how to distinguish if it is a malignant or benign tumor [16].

In recent years, modern research and machine learning techniques have been taken into consideration for the treatment of Breast Cancer. Breast Cancer is not only detected earlier but also predictions can be made about whether a person will be able to survive it and about how likely is that the cancer cells will start recurring [25]. Careful analysis of different models and features can help in the development of a model that achieves high accuracy. SVM and ANN are two of the most widely used classification algorithms for solving the breast cancer prediction problem [24].

The most common datasets used for prediction of Breast Cancer are the Breast Tissue Dataset (BTD) [3], the Coimbra Breast Cancer (CBC) dataset [2], the SEER Breast Cancer (SEERBCD) dataset [4], the Wisconsin (Original) Breast Cancer (WOBC) dataset [6], the Wisconsin (Diagnostic) Breast Cancer (WDBC) dataset and the Wisconsin (Prognostic) Breast Cancer (WPBC) dataset [7]. BTD [3] contains the impedance measurement of freshly excised breast tissue that were obtained at various different frequencies. It comprises of a total of 6 classes. The CBC [2] dataset contains anthropometric data and parameters that are gathered in routine blood analysis. There are a total of 10 predictors. The predictors are quantitative and the presence or absence of breast cancer is indicated by a binary dependent variable. The SEERBCD was obtained in November 2017 under the SEER program of NCI.

This program provides information on population-based cancer statistics. It contains the data of 4024 patients that includes their age, race, marital status, tumor size, estrogen status, progesterone status, information about regional nodes and some other factors. The WOBC dataset was obtained from clinical cases reported by Dr. Wolberg. The databases reflect chronological grouping of data with 8 groups having number of cases recorded between January 1989 and November 1991. The WPBC contains information about 30 features that are computed from digitized images. Each record is a representation of the follow-up data for a single breast cancer case. It consists of a total of 34 attributes. The WDBC dataset is similar to WPBC.

Authors in [17] predicted a patient survival chance, by feeding various prognostic variables involving time factor, to the neural network. The predict results of the neural network were compared to that of a regression model, using data of 1373 patients to determine the performance of the models. For predictions of malignant probabilities for

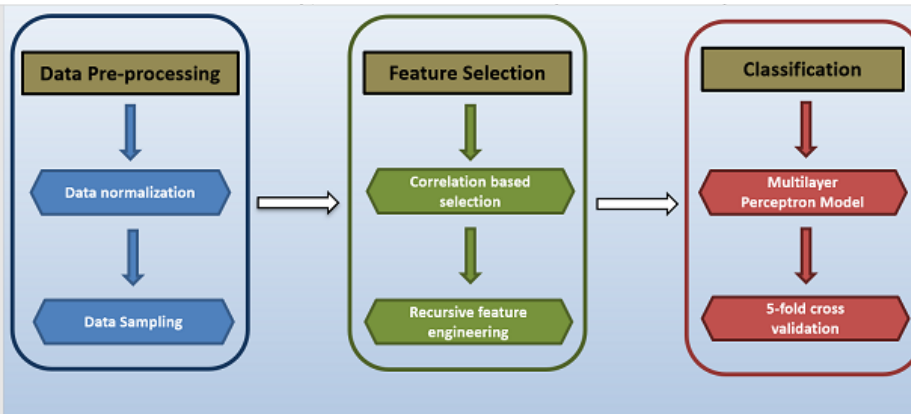

Figure 1. Flowchart of BCAD Framework

non-recurring cases and the recurring time period of diseases, the work of [28] designed a linear diagnostics model. This model was tested on the dataset for 569 patient and involved a cross-validation approach, giving the accuracy of up to 97.5%.

Quinlan in [20] also developed a model for medical diagnostics and predictions by incorporating Minimum Description length (MDL) penalty to the C4.5 decision tree algorithm, gaining accuracy of 94.74%. Delen in [27] then compared the Decision Tree model such as C4.5 with Neural Networks and Linear Regression models, using large datasets of around 200000 patient records. They concluded that a Decision tree algorithm such as C4.5 can often outperform the other two, on large datasets, by achieving accuracy of around 93.6% or more.

Feature selection and classification effect accuracy up to a great extent. Feature selection improves the overall classification accuracy and processing time since if proper features are selected, then it will not only benefit accuracy but also improve overall training time. Most of the above studies focused mainly on preparing the data using feature selections, feature extraction, various forms of data representations, and tuning model parameter, instead of dealing with model structures and changing them to obtain better performances. Even in model tuning, processes such as differential evolution leads to greater risk of overfitting the model.

3. System Model

Figure 1 shows the flowchart of BCAD Framework including data preprocessing steps, feature selection using filter-based feature selection methods and classification using Multilayer perceptron model.

First, data pre-processing is performed, data is checked for any inconsistencies or missing values, followed by random sampling. Sampling generates a unique sampling distribution that is based on the actual data. Sampling is done so that the result can be a more accurate estimate of the features selected. The data is then normalized.

Feature selection is then performed by correlation-based selection and recursive feature elimination methods. First, a correlation analysis of the features is done. The features that are highly co-related are set aside and one of them is selected in such a manner that

Table 1. COMPARISON OF MACHINE LEARNING ALGORITHMS ON WISCONSIN BREAST CANCER DATASET

Paper	Features	Classifier	Validation	Accuracy
[8]	30	Gradient Boosting	10-fold	98.88%
[34]	20	Random Forest	10-fold	98.77%
[14]	–	K-means Clustering	–	92.00%
[32]	30	SMO	10-fold	97.71%
Our approach	11	MLP	5-fold	99.12%

if three of the features are highly correlated, then one of them is selected.
The reason for this is that highly correlated features have the same impact on the result. Since there contribution to the result will be same, using all the features will not be a sensible approach, as the impact on the result would be the same. This results in a reduced feature set of 16 features. Recursive feature elimination (RFE) is then applied to the data in order to obtain the optimal features. This results in the resulting set of 11 features that are then sent as input to the classification models.

4. Results

Table 1 represents various machine learning models proposed by different authors and their achievements. In the approach proposed in this paper, for the purpose of feature selection, I have first eliminated the features that were highly correlated in such a manner that only one of the features that are highly correlated are selected i.e. if three of the features are highly correlated, one of them is selected. This results in a reduced feature set of 16 features. Further, I have applied recursive feature selection method to further reduce the number of features. This results in the selection of 11 best features for the purpose of classification that are sent to the classification models. The best accuracy is achieved by random forest.
The proposed approach was aimed at Feature space selection in order to test the influence of the feature space. For this purpose, a hybrid of correlation-based feature selection and Recursive feature elimination (RFE) were used to reduce the feature space to 11 features. This is particularly important to cater the problem of overfitting in Machine Learning.

Table 2 represents the comparison of our proposed framework with various machine learning based approaches proposed by various different authors and their achievements on the WPBC and WOBC datasets.
Table 2 shows the results of our approach on WOBC and WPBC datasets. We have applied our framework on both the datasets and recorded the results. First, standardization is done to ensure good normalization of features followed by feature selection. The analysis to identify the strongest predictors is on filter-based feature selection methods: correlation analysis followed by recursive feature elimination. It is evident that our proposed EDFBC framework outperforms state-of-the-art approaches. Optimal feature selection enhances the overall classification process, increasing the accuracy and reducing overall training time.

Table 2. COMPARISON OF BCAD FRAMEWORK ON WOBC AND WPBC DATASETS

Paper	Features	Classifier	Validation	Accuracy
[33]	10	J48 and MLP	10-fold	97.28%
[15]	–	Random Forest	10-fold	96.70%
Our approach	8	MLP	5-fold	98.20%
[35]	–	Fuzzy C-means clustering	4-fold	97.13%
[18]	19	Linear Regression	10-fold	84.34%
Our approach	16	MLP	5-fold	98.33%

n

5. Conclusion and Future Work

We have present BCAD framework for classification of breast cancer using machine learning models with focus on careful feature selection. Using a hybrid of correlation-based feature selection and recursive feature elimination, useful features are extracted from the WDBC dataset and used for classification using the Multilayer Perceptron Models. Experiments have shown that our framework outperforms state-of-the-art methods. We used the Wisconsin Diagnostic Breast Cancer (WDBC) Dataset that is used as a benchmark to improve the early diagnosis of breast cancer and eliminate challenges faced by radiologists in determining if tumor is malignant or benign. We compare the performance of proposed models including SVM, Random Forest, Gradient Boosting, Artificial Neural Network and Multilayer Perceptron Model. The best performing algorithm was the Multilayer Perceptron Model with an accuracy of 99.12%.
In future, it is recommended that large standardized public datasets must be constructed and then combined with the application of different feature selection and classification techniques to provide promising tools for the detection of breast cancer in its early stages.

References

[1] (2018, 10.01.2018). cancer. available: $http://www.who.int/en/news-room/fact-sheets/detail/cancer$.
[2] Coimbra breast cancer dataset. available at: $https://archive.ics.uci.edu/ml/datasets/breast+cancer+coimbra$.
[3] $https://archive.ics.uci.edu/ml/datasets/breast+tissue$.
[4] Seer breast cancer dataset. available at: $https://ieee-dataport.org/open-access/seer-breast-cancer-data$.
[5] Wisconsin diagnostic breast cancer dataset. available at: $http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)$.
[6] Wisconsin original breast cancer dataset. available at: $https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\%28original\%29$.
[7] Wisconsin prognostic breast cancer dataset. available at: $https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\%28prognostic\%29$.
[8] Samyam Aryal and Bikalpa Paudel. Supervised classification using gradient boosting machine: Wisconsin breast cancer dataset. 2020.
[9] Marina Bonomolo, Patrizia Ribino, and Gianpaolo Vitale. Explainable post-occupancy evaluation using a humanoid robot. *Applied Sciences*, 10(21):7906, 2020.
[10] A. Coronato and G. De Pietro. Tools for the rapid prototyping of provably correct ambient intelligence applications. *IEEE Transactions on Software Engineering*, 38(4):975–991, 2012.
[11] A. Coronato and G. D. Pietro. Formal design of ambient intelligence applications. *Computer*, 43(12):60–68, 2010.

[12] Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020.

[13] Claudia Di Napoli, Patrizia Ribino, and Luca Serino. Customisable assistive plans as dynamic composition of services with normed-qos. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–26, 2021.

[14] Ashutosh Kumar Dubey, Umesh Gupta, and Sonal Jain. Analysis of k-means clustering approach on the breast cancer wisconsin dataset. *International journal of computer assisted radiology and surgery*, 11(11):2033–2047, 2016.

[15] P Hamsagayathri and P Sampath. Performance analysis of breast cancer classification using decision tree classifiers. *Int J Curr Pharm Res*, 9(2):19–25, 2017.

[16] Murat Karabatak. A new classifier for breast cancer detection based on naïve bayesian. *Measurement*, 72:32–36, 2015.

[17] Yılmaz Kaya and Murat Uyar. 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, 2013.

[18] Rafaqat Alam Khan, Nasir Ahmad, and Nasru Minallah. Classification and regression analysis of the prognostic breast cancer using generation optimizing algorithms. *International Journal of Computer Applications*, 68(25):42–47, 2013.

[19] Carmelo Lodato and Patrizia Ribino. A novel vision-enhancing technology for low-vision impairments. *Journal of medical systems*, 42(12):1–13, 2018.

[20] Olvi L Mangasarian, W Nick Street, and William H Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.

[21] Muddasar Naeem, Giovanni Paragiola, Antonio Coronato, and Giuseppe De Pietro. A cnn based monitoring system to minimize medication errors during treatment process at home. In *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, pages 1–5, 2020.

[22] Muddasar Naeem, Giovanni Paragliola, and Antonio Coronato. A reinforcement learning and deep learning based intelligent system for the support of impaired patients in home treatment. *Expert Systems with Applications*, page 114285, 2020.

[23] Muddasar Naeem, S Tahir H Rizvi, and Antonio Coronato. A gentle introduction to reinforcement learning and its application in different fields. *IEEE Access*, 2020.

[24] Abdullah-Al Nahid and Yinan Kong. Involvement of machine learning for breast cancer image classification: a survey. *Computational and mathematical methods in medicine*, 2017, 2017.

[25] World Health Organization et al. *WHO position paper on mammography screening*. World Health Organization, 2014.

[26] Giovanni Paragliola and Muddasar Naeem. Risk management for nuclear medical department using reinforcement learning algorithms. *Journal of Reliable Intelligent Environments*, 5(2):105–113, 2019.

[27] J Ross Quinlan. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, 4:77–90, 1996.

[28] Peter M Ravdin and Gary M Clark. A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast cancer research and treatment*, 22(3):285–293, 1992.

[29] Patrizia Ribino, Marina Bonomolo, Carmelo Lodato, and Gianpaolo Vitale. A humanoid social robot based approach for indoor environment quality monitoring and well-being improvement. *International Journal of Social Robotics*, pages 1–20, 2020.

[30] Patrizia Ribino and Carmelo Lodato. A distributed fuzzy system for dangerous events real-time alerting. *Journal of Ambient Intelligence and Humanized Computing*, 10(11):4263–4282, 2019.

[31] Patrizia Ribino and Carmelo Lodato. A norm compliance approach for open and goal-directed intelligent systems. *Complexity*, 2019, 2019.

[32] Gouda I Salama, M Abdelhalim, and Magdy Abd-elghany Zeid. Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, 32(569):2, 2012.

[33] Gouda I Salama, M Abdelhalim, and Magdy Abd-elghany Zeid. Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, 32(569):2, 2012.

[34] Ahmet SAYGILI. Classification and diagnostic prediction of breast cancers via different classifiers. *International Scientific and Vocational Studies Journal*, 2(2):48–56, 2018.

[35] PB Tintu and R Paulin. Detect breast cancer using fuzzy c means techniques in wisconsin prognostic breast cancer (wpbc) data sets. *International Journal of Computer Applications Technology and*

Research, 2(5):614–617, 2013.

[36] Haowen You and George Rumbe. Comparative study of classification techniques on breast cancer fna biopsy data. 2010.