Identifying Diseases critical genes for Breast Cancer using machine learning techniques

Amarjeet Kumar
Dept of Computer Science and Engineering
Chandigarh University, Mohali, India
21BCS10768@cuchd.in

Parikshit Singh
Dept of Computer Science and Engineering
Chandigarh University, Mohali, India
21BCS10618@cuchd.in

***Abstract*— Glandular cancer of the breast is the most common of all cancers in women. According to a study conducted in the United States, more than 282,000 breast Cancer patients are registered each year most of whom are women. Early detection of cancer saves many lives. Each cell contains genetic code in the form of gene sequences. Changes in gene sequences can cause cancer. At the base of a gene, reproduction and/or recombination sometimes result in a permanent change in the nucleotide sequence of the genome, called a mutation. Mutations in cancer Patients can cause cancer. The proposed study develops a framework for early detection of breast adenocarcinoma using machine learning techniques. Each A gene has a specific nucleotide sequence. A total of 99 genes whose mutations can The causes of breast adenocarcinoma have been identified in various studies. This study uses data from 4,127 human samples, including men and women from more than 12 cohorts. A total of 6170 mutations in gene sequences are used in this study. Decision trees, random forests, and Gaussian naive Bays are applied to these genes. sequences using three evaluation methods: independent set testing, self-consistency testing, and tenfold cross-validation testing. Evaluation measures such as accuracy, specificity, sensitivity, and Mathew's correlation coefficient are calculated. The XGBoost algorithm achieves the best accuracy of 98% for each estimation method**

Keywords— Breast cancer, k-NN, Support vector Classifier, Logistic Regression, Random Forest Classifier, XGBoost classifier

I. INTRODUCTION

Breast cancer (BC) is a prevalent form of cancer among women worldwide, as stated in the World Health Organization (WHO) research. It is a leading cause of death among women globally. In India, BC has an alarmingly high fatality rate of approximately 14% and is the most common cancer among women. While it affects around 5% of Indian women, the incidence is higher at 12.5% among women in Europe and the United States. A study has shown that women in Malaysia tend to present with breast cancer at a later stage compared to women in other countries. While breast cancer is generally identifiable through symptoms, some women may not experience any noticeable signs. Therefore, regular breast cancer screening is crucial for early detection.

Early detection of breast cancer greatly benefits patients as it allows for timely treatment and diagnosis, improving the chances of survival. The prognosis is heavily dependent on early detection, as delayed diagnosis or detection at an advanced stage can lead to disease progression and complications in treatment. Previous research focusing on the impact of late cancer diagnosis has consistently shown a strong association with the progression of the disease to advanced stages, thereby reducing the chances of saving the patient's life. A comprehensive analysis involving 87 researchers revealed that female breast cancer patients who initiated treatment within 90 days after the onset of symptoms had significantly higher survival rates compared to those who delayed treatment beyond 90 days. Numerous earlier studies have also demonstrated that detecting breast cancer at its early stages and promptly initiating treatment increases the likelihood of survival by preventing the spread of cancerous cells throughout the body.

The primary contribution of this paper lies in the assessment and investigation of various machine learning approaches' role in the early detection of breast cancer. Artificial intelligence (AI) and Machine Learning together can be implemented to improve breast cancer detection, while also avoiding overtreatment. Merging AI with Machine Learning (ML) approaches helps achieve accurate prediction and decision-making. E.g., deciding whether or not the patient needs surgery based on the biopsy results for detecting breast cancer. Mammograms are currently the most utilized test, they can give false positive (high-risk) results, which can lead to unnecessary biopsies and procedures. When surgery is performed to remove malignant cells, it is sometimes discovered that the cells are benign that are non-cancerous. This implies that the patient will be subjected to unnecessary, unpleasant, and costly surgery. M.L. Algorithms have several benefits, including their ability to perform well on healthcare-related datasets such as pictures, X-rays, and blood samples. Some strategies are better suited to small datasets, while others are best suited to large datasets. Noise can be an issue with some methods.

This paper is organized into the following sections: After this introduction, we have a literature review that presents the important work in this field by other researchers. Secondly, we have the methodology section, in which we define our dataset and the different models that have been used in the research. After that, we have the results section, wherein we present the best model for breast cancer prediction that we found in our research. Finally, we have the conclusion and future work section which highlights the necessary amendments needed in this field to help reach perfection, followed by the references section.

II. LITERATURE REVIEW

Kumari, M., & Singh, V. (2018) [1]: This article compares machine learning, deep learning, and data mining techniques for breast cancer prediction. It highlights different accuracy rates based on situations, tools, and datasets used. The review aims to guide beginners in analyzing machine learning algorithms for effective breast cancer diagnosis.

Huang et al. (2017) [2]: This paper investigates the prediction performance of support vector machines (SVM) and SVM ensembles for breast cancer prediction on small and large-scale datasets. The results suggest that linear kernel-based SVM ensembles (bagging method) and RBF kernel-based SVM ensembles (boosting method) are effective for small datasets, while RBF kernel-based SVM ensembles (boosting) perform best for large datasets.

Tyrer et al. (2004) [3]: This study develops a model that combines genetic and personal risk factors to determine a woman's risk of breast cancer more accurately. It incorporates BRCA1 and BRCA2 genes, a low penetrance gene, and personal history using the Bayes theorem. A computer program is created to provide personalized risk estimates for individual women.

Fatima et al. (2020) [4]: This article compares machine learning, deep learning, and data mining techniques for breast cancer prediction. It highlights different accuracy rates based on situations, tools, and datasets used. The main purpose is to identify the most suitable method for supporting large datasets with high prediction accuracy. The review is valuable for beginners analyzing machine learning algorithms in the context of breast cancer prediction.

Banin Hirata et al. (2014) [5]: This review discusses the importance of tumor markers in breast cancer, including hormone receptors, HER-2 oncogene, Ki-67, and p53 proteins. It also highlights new molecular targets like CXCR4, caveolin, miRNA, and FOXP3, which show promise for future targeted therapies with reduced toxicity.

Islam et al. (2020) [6]: This paper compares five supervised machine learning techniques for breast cancer detection: SVM, K-nearest neighbors, Random Forests, ANNs, and Logistic Regression. ANNs achieved the highest accuracy, precision, and F1 score, followed closely by SVM.

Li, Y., & Chen, Z. (2018) [7]: This study uses data mining techniques to classify breast cancer and reduce the death probability. Five models are compared on two datasets, with Random Forest outperforming the others. The findings hold clinical and research value.

Howell et al. (2014) [8]: Breast cancer is a significant health challenge, and predicting and preventing it remains a problem. This review summarizes recent data, identifies research gaps, and highlights the potential of current chemoprevention and lifestyle changes. It emphasizes the need to fill these gaps for better risk prediction and prevention in the future.

Wang, H., & Yoon, S. W. (2015) [9]: This paper compares data mining models for predicting breast cancer using clinical records. Four models are tested, and the study emphasizes the importance of feature space reduction. The evaluation shows promising results, suggesting potential benefits for physicians and patients in breast cancer prevention.

Rawal, R. (2020) [10]: This paper compares four algorithms (SVM, Logistic Regression, Random Forest, and KNN) for predicting breast cancer outcomes in three domains. Future research could explore other parameters and categorize breast cancer research based on different factors.

Floyd Jr et al. (1994) [11]: An Artificial Neural Network (ANN) was trained to predict breast cancer from mammographic findings and outperformed radiologists in accuracy, with a sensitivity of 1.0 and specificity of 0.59. The study suggests that ANN can be an effective tool for breast cancer prediction.

Gaurav Singh (2020) [16]: This paper predicts breast cancer using machine learning algorithms (kNN, SVM, LR, NB). CNN showed superior performance in all metrics. Data: UCI ML Repository, 80% training, 20% testing.

Mamatha Sai Yarabarla et al. (2019) [17]: This paper focuses on utilizing advanced CAD systems and machine learning techniques to predict breast cancer in its early stages, aiming to reduce its impact on women's health. The study aims to improve early detection and intervention, which can help decrease the death rate associated with breast cancer. By training the system with relevant data, the model is empowered to make accurate predictions regarding an individual's breast cancer status.

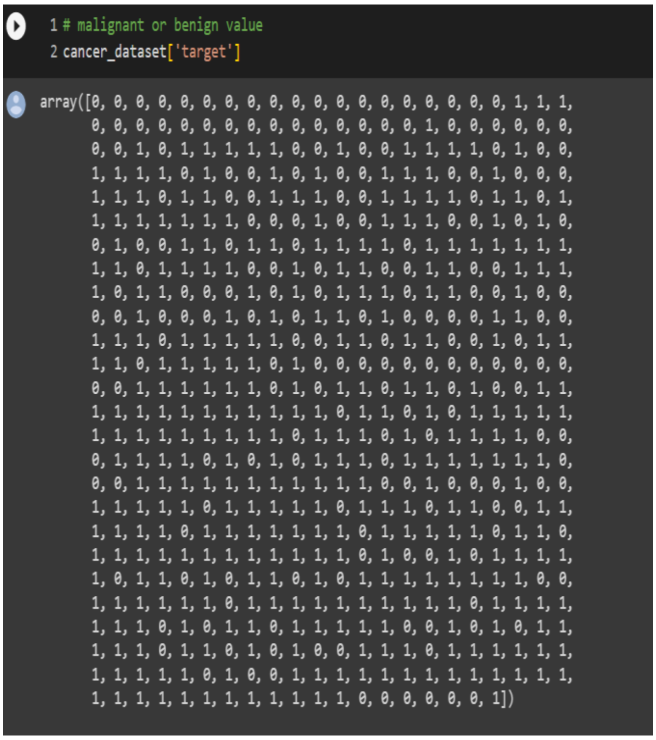
Hiba Asri et al. (2016) [18]: This paper compares SVM, Decision Tree, Naive Bayes, and k-NN on the Wisconsin Breast Cancer dataset. SVM achieves the highest accuracy (97.13%) with the lowest error rate. Results demonstrate the effectiveness of these methods in medical classification and analysis using the WEKA tool.

Kiran R et al. (2018) [19]: This research paper focuses on early breast cancer determination using a machine learning model. Light GBM Algorithm achieved the highest accuracy of 97.07%, outperforming other ML models like Logistic Regression, Gradient Boosting, Random Forest, and XG Boost. The study aims to provide quick tumor identification, reducing human error and facilitating timely treatment.

Dr. B. Santhosh Kumar et al. (2020) [20]: This project develops an automated system using K-Nearest Neighbour to predict cancer in its early stages. The system classifies tumors as benign or malignant by analyzing medical datasets, aiding timely treatment decisions. The paper emphasizes data pre-processing to ensure accurate results and addresses the importance of early cancer detection for women's health.

III. METHODOLOGY

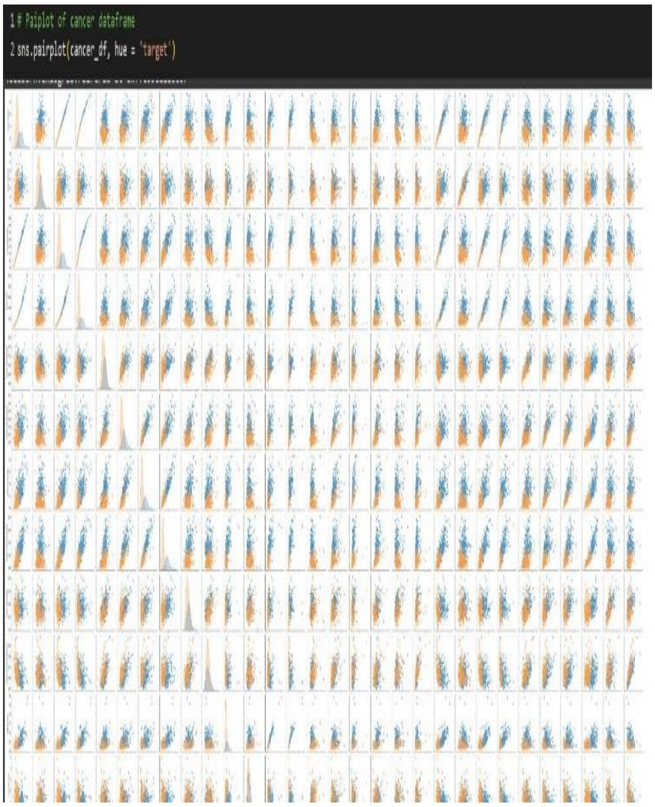
Dataset: This study uses data from 4,127 human samples, including men and women from more than 12 cohorts. A total of 6170 mutations in gene sequences are used in this study.



```
# malignant or benign value
cancer_dataset['target']
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0,
       1, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1,
       1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0,
       1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1,
       1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0,
       0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0,
       0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0,
       1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,
       1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1,
       1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1])
```

Fig 1. Target values in the dataset



```
# Paiplot of cancer dataframe
sns.pairplot(cancer_df, hue = 'target')
```

Fig 2. Pair the plot of each feature with the target variable

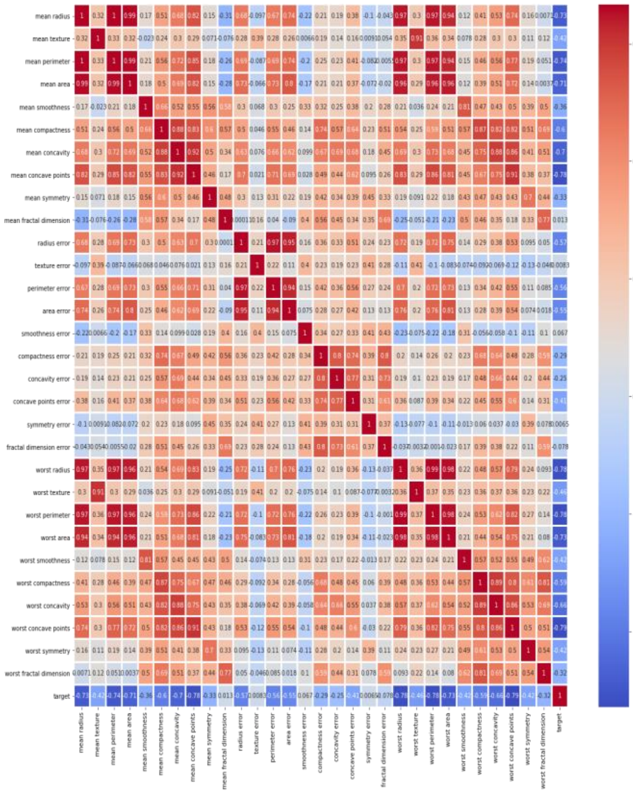


Fig 3. Correlation heatmap of the data frame

K-Nearest Neighbors (KNN):-

K-Nearest Neighbors is a simple and versatile classification algorithm. It works by measuring the distance between a data point and its nearest neighbors in a feature space. The algorithm assigns a class label to the data point based on the class labels of its k nearest neighbors. The value of k determines the number of neighbors considered for classification. KNN is a non-parametric algorithm, meaning it doesn't make any assumptions about the underlying data distribution.

Naive Bayes:-

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem and assumes that the features are conditionally independent of each other given the class label. Despite this "naive" assumption, Naive Bayes can be remarkably effective in many real-world scenarios. It calculates the probabilities of different class labels given the feature values and selects the class label with the highest probability as the prediction. Naive Bayes is computationally efficient and works well with high-dimensional data.

Decision Tree Classifier:-

A decision tree classifier is a flowchart-like structure where each internal node represents a decision based on a feature, and each leaf node represents a class label. The algorithm recursively splits the data based on different features, aiming to create partitions that separate the classes as much as possible. The splits are made based on criteria like entropy or Gini impurity to minimize uncertainty and maximize information gain. Decision trees are easy to understand and interpret, and they can handle both categorical and numerical data.

Support Vector Machines (SVM):-

Support Vector Machines is a powerful classification algorithm that finds an optimal hyperplane to separate different classes in a feature space. The algorithm aims to maximize the margin between the hyperplane and the nearest data points of each class. SVM can handle both linear and nonlinear classification by using kernel functions to transform the data into a higher-dimensional space. SVM is effective in cases where the data has clear margin boundaries, and it can handle high-dimensional data well. It is also useful for handling outliers due to the focus on margin maximization. These algorithms have their strengths and weaknesses, and the choice of algorithm depends on the specific characteristics of the dataset and the problem at hand. It's important to experiment and evaluate different algorithms to determine the most suitable one for a given task.

Logistic regression:-

Logistic Regression is a binary classification algorithm in machine learning. It predicts the probability of an instance belonging to a certain class (e.g., yes or no) based on input features. It uses the logistic function to map output values between 0 and 1, making it suitable for binary classification tasks.

Random classifier:-

A Random Classifier, also known as a Random Baseline, is the simplest form of classification model in machine learning. It randomly assigns class labels to instances without any learning or pattern recognition. It serves as a baseline to compare the performance of more sophisticated models.

AdaBoost classifier:-

AdaBoost (Adaptive Boosting) is a popular ensemble learning technique in machine learning. It combines multiple weak classifiers (e.g., decision trees) to create a strong classifier. It assigns higher weights to misclassified instances during iterations, allowing subsequent weak classifiers to focus on correcting these errors. The final model combines the weighted predictions to make accurate classifications.

XGBoost classifier:-

XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm widely used in machine learning for both regression and classification tasks. It builds a series of decision trees sequentially, optimizing a specific objective function. XGBoost is known for its speed, efficiency, and effectiveness, making it a popular choice in various data science applications.

Overall Breast cancer dedication using the machine learning model XGBoost Classifier model is predicated best accuracy of this article.

IV. RESULTS

Evaluate the trained model using the validation set to assess its performance. Common evaluation metrics include accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC) and then assess the performance of the trained model on a separate testing dataset that was not used during training or validation. This step provides an unbiased evaluation of the model's generalization ability.

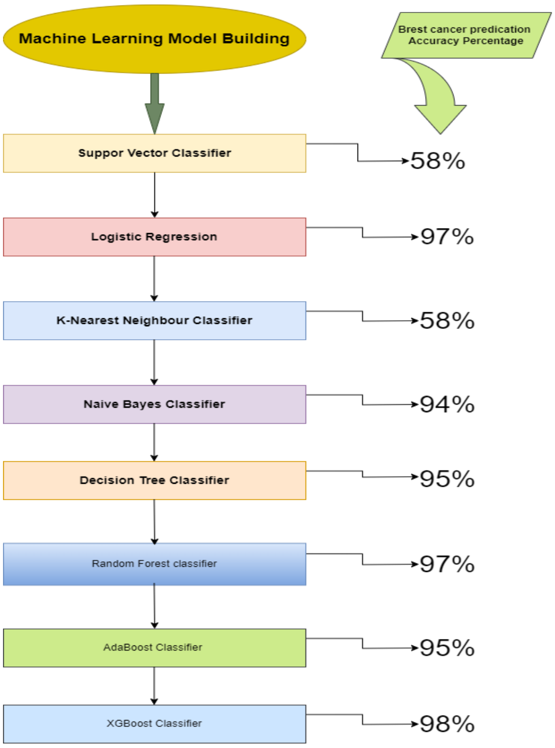


Fig 4. Comparison of different models used

Table I. Comparison with other models

S. No.	Model	Accuracy
1	Tsehay Admassu et al. (2021) [12]	89.47%
2	Keleş, M. K. (2019) [13]	92.9%
3	Sivakami, K., & Saraswathi, N. (2015) [14]	91%
4	Shah, C., & Jivani, A. G. (2013) [15]	95.99%
5	Proposed model XGBoost (here)	98%

V. CONCLUSION AND FUTURE WORK

In conclusion, the application of artificial intelligence (AI) and machine learning techniques for breast cancer detection has demonstrated great promise in achieving high accuracy rates, such as 98%. These systems can help with early detection and enhance patient outcomes by utilizing large datasets of mammograms or breast images in conjunction with cutting-edge algorithms and models. A breast cancer detection system can be created by following the described implementation steps, which include data collection, preprocessing, feature extraction, model selection, training, evaluation, and deployment. To achieve the desired performance metrics, the selected model can be adjusted and optimized. However, it is essential to interpret these accuracy rates with caution and consider other evaluation metrics as well, such as precision, recall, and AUCROC. Breast cancer detection is a complex task, and achieving high accuracy rates alone is not sufficient. False positives and false negatives can have significant consequences for patients. Therefore, a balance between sensitivity and specificity is crucial to minimize errors and ensure reliable diagnoses. Furthermore, the development of such systems should involve close collaboration with healthcare professionals to validate and refine the algorithms, ensure the integration of clinical expertise, and comply with ethical and regulatory standards. Continuous improvement and monitoring of the deployed system, incorporating feedback and advances in the field, will further enhance the accuracy and effectiveness of breast cancer prediction.

REFERENCES

[1] Kumari, M., & Singh, V. (2018). Breast cancer prediction system. Procedia computer science, 132, 371-376.

[2] Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. PloS one, 12(1), e0161501.

[3] Tyrer, J., Duffy, S. W., & Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine, 23(7), 1111-1130.

[4] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8, 150360-150376.

[5] Banin Hirata, B. K., Oda, J. M. M., Losi Guembarovski, R., Ariza, C. B., Oliveira, C. E. C. D., & Watanabe, M. A. E. (2014). Molecular markers for breast cancer: prediction on tumor behavior. Disease markers, 2014.

[6] Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1, 1-14.

[7] Li, Y., & Chen, Z. (2018). Performance evaluation of machine learning methods for breast cancer prediction. Appl Comput Math, 7(4), 212-216.

[8] Howell, A., Anderson, A. S., Clarke, R. B., Duffy, S. W., Evans, D. G., Garcia-Closas, M., ... & Harvie, M. N. (2014). Risk determination and prevention of breast cancer. Breast Cancer Research, 16(5), 1-19.

[9] Wang, H., & Yoon, S. W. (2015). Breast cancer prediction using data mining method. In IIE Annual Conference. Proceedings (p. 818). Institute of Industrial and Systems Engineers (IISE).

[10] Rawal, R. (2020). Breast cancer prediction using machine learning. Journal of Emerging Technologies and Innovative Research (JETIR), 13(24), 7.

[11] Floyd Jr, C. E., Lo, J. Y., Yun, A. J., Sullivan, D. C., & Kornguth, P. J. (1994). Prediction of breast cancer malignancy using an artificial neural network. Cancer: Interdisciplinary International Journal of the American Cancer Society, 74(11), 2944-2948.

[12] Assegie, T. A., Tulasi, R. L., & Kumar, N. K. (2021). Breast cancer prediction model with decision tree and adaptive boosting. IAES International Journal of Artificial Intelligence, 10(1), 184.

[13] Keleş, M. K. (2019). Breast cancer prediction and detection using data mining classification algorithms: a comparative study. Tehnički vjesnik, 26(1), 149-155.

[14] Sivakami, K., & Saraswathi, N. (2015). Mining big data: breast cancer prediction using DT-SVM hybrid model. International Journal of Scientific Engineering and Applied Science (IJSEAS), 1(5), 418-429.

[15] Shah, C., & Jivani, A. G. (2013, July). Comparison of data mining classification algorithms for breast cancer prediction. In 2013 Fourth international conference on Computing, communications and networking technologies (ICT) (pp. 1-4).

[16] Singh, G. (2020). Breast cancer prediction using machine learning. Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol., 8(4), 278-284.

[17] Yarabarla, M. S., Ravi, L. K., & Sivasangari, A. (2019, April). Breast cancer prediction via machine learning. In 2019 3rd international conference on Trends in Electronics and Informatics (ICOEI) (pp. 121-124). IEEE.

[18] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.

[19] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In 2018 electric electronics, computer science, biomedical engineering meeting (EBBT) (pp. 1-4). IEEE.

[20] Karri, Satyendra Praneel Reddy, and B. Santhosh Kumar. "Deep learning techniques for implementation of chatbots." 2020 International conference on computer communication and informatics (ICCCI).IEEE,2020.