2024 IEEE AITU: Digital Generation
4-5 April, 2024, Astana, Kazakhstan

Breast Cancer Prediction System Utilizing Machine Learning Algorithms

Chirayou Bista
Btech. Information Science and Engineering, Jain (Deemed-to-be) Univeristy
Bangalore, India
20btris051@jainuniversity.ac.in

Asreetha M
Btech. Information Science and Engineering, Jain (Deemed-to-be) Univeristy
Bangalore, India
20btris003@jainuniversity.ac.in

Salahuddin Slimanzay
Btech. Information Science and Engineering, Jain (Deemed-to-be) Univeristy
Bangalore, India
20btris06w@jainuniversity.ac.in

Md Solaiman Sheikh
Btech. Information Science and Engineering, Jain (Deemed-to-be) Univeristy
Bangalore, India
20btris032@jainuniversity.ac.in

Dr. P Srinivasa Rao
Associate Professor
Jain (Deemed-to-be) University
Bangalore, India
srinivasa.p@jainuniversity.ac.in

***Abstract*— Breast cancer remains a significant global health concern, necessitating advanced predictive tools for early detection and effective prognosis. In response to this challenge, this research introduces a Breast Cancer Prediction System utilizing state-of-the-art machine learning algorithms. The system leverages the power of Random Forest, Support Vector Machine (SVM), and Gradient Boosting Ensemble to enhance accuracy and reliability. The Random Forest algorithm efficiently captures complex relationships within the breast cancer dataset, creating an ensemble of decision trees for robust predictions. Meanwhile, the Support Vector Machine optimally classifies data points by identifying hyperplanes, thereby enhancing the model's ability to discriminate between benign and malignant cases. Furthermore, the Gradient Boosting Ensemble technique synthesizes weak predictive models into a strong learner, boosting the overall predictive performance of the system.**

The Breast Cancer Prediction System offers a thorough and precise evaluation of the likelihood of breast cancer by combining the advantages of all three algorithms. By utilizing the discriminative power of the SVM, the boosting mechanism of the Gradient Boosting Ensemble, and the Random Forest's capacity to

handle complex datasets, the system attains increased accuracy in early detection and prognosis. The suggested system has great promise for useful application in clinical settings, resulting in prompt interventions and better patient outcomes.

Keywords*— *Machine Learning, Support Vector Machine, Random Forest, Gradient Boosting Ensemble, Breast Cancer.

I. INTRODUCTION

Breast cancer is the primary cause of morbidity and death for women worldwide, making it a major health problem. Enhanced patient outcomes are largely dependent on early detection and precise prognosis, underscoring the importance of sophisticated predictive techniques. This paper presents a comprehensive breast cancer prediction system that uses machine learning techniques to meet this need. A comprehensive method to improve the precision and dependability of breast cancer forecasts is provided by the combination of Random Forest, Support Vector Machine (SVM), and Gradient Boosting Ensemble [3]. The goal of this research is to close the gap between current computational techniques and conventional diagnostic procedures, resulting in a reliable and effective system for breast cancer prognosis and early detection.

Personalized medicine has advanced in recent years because to the tremendous success machine learning algorithms have shown in a variety of healthcare applications [1]. The suggested Breast Cancer Prediction System makes use of the advantages of Random Forest, Support Vector Machines, and Gradient Boosting Ensemble, combining their combined capabilities to analyze intricate datasets and provide accurate predictions. The project intends to support current efforts in the medical sector to combat breast cancer by combining these state-of-the-art algorithms, providing a tool that could revolutionize early intervention tactics and eventually improve patient outcomes [7].

II. LITERATURE SURVEY

[1] Authored by Wang, Y., Xu, Y., & Zhang, L. In order to improve breast cancer prediction, this paper investigates the use of ensemble learning methodologies. With regard to machine learning in healthcare diagnostics, the work provides significant new insights by investigating the application of

2024 IEEE AITU: Digital Generation | 979-8-3503-6437-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEEECONF61558.2024.10585589

Authorized licensed use limited to: Texas Tech University. Downloaded on October 18,2024 at 19:24:31 UTC from IEEE Xplore. Restrictions apply.

ensemble techniques to improve the accuracy and reliability of models for the diagnosis of breast cancer.

[2] Conducted by Sharma, S., Rani, R., & Anand, P. This comprehensive study offers insights into breast cancer prediction using multiple machine learning techniques. The authors examine a wide range of algorithms and assess how well they predict breast cancer based on many datasets. Our understanding of the nuances required in applying machine learning to forecast the prognosis of breast cancer is enhanced by this work.

[3] Jain, A., Singh, A. K., & Bhardwaj, R., present a study that integrates genetic algorithms with machine learning for breast cancer prediction. By optimizing the machine learning models with genetic algorithms, the research seeks to improve prediction accuracy. This creative method demonstrates how integrating various computing tools might lead to better prediction results.

[4] Authored by Li, H., Zhang, Y., & Li, YThis systematic study critically examines the application of machine learning techniques to the prediction of recurrence of breast cancer. The paper offers a comprehensive overview of the literature, highlighting trends, challenges, and opportunities in the use of machine learning for breast cancer prognosis. The findings aid in deciding the direction for this emerging field's future.

[5] Sathiya, P. S., Reddy, S. K., & Reddy, N. R. K., contribute to the literature by predicting breast cancer survivability using machine learning techniques. The research investigates the use of machine learning algorithms to assess the likelihood of breast cancer survival, providing valuable insights into prognosis and treatment planning.

[6] Authored by Das, R., Jain, S. S., & Singh, A. K., This paper proposes a hybrid approach for early breast cancer screening that makes use of machine learning techniques. The goal of the project is to improve early diagnosis of breast cancer by merging many machine learning techniques. The hybrid method offers a more nuanced approach to optimising the predictive capability of machine learning models for timely intervention.

III. Proposed Work

In order to improve early detection and prognosis, the proposed Breast Cancer Prediction System makes use of the power of machine learning methods, specifically Random Forest, Support Vector Machine (SVM), and Gradient Boosting Ensemble. First, a wide range of breast cancer datasets are gathered, and then the data is carefully preprocessed to guarantee its consistency and quality. Techniques for feature extraction and selection are then used to determine the most relevant features, maximizing the effectiveness of the ensuing model [6]. The creation and training of each individual algorithm, which is then refined by cross-validation, is the basis of the system [2]. The ensemble method builds a reliable, sensitive, and targeted prediction model by combining the advantages of Random Forest, SVM, and Gradient Boosting.

Evaluation measures are used to compare and thoroughly evaluate the performance of each method and the ensemble, including precision, recall, F1-score, and AUC-ROC. The objective of this proposed effort is to improve patient outcomes and personalized healthcare by improving the accuracy of breast cancer forecasts and facilitating their seamless incorporation into clinical practice [3].

The initial stage of the proposed effort will concentrate on data collecting and preprocessing in order to construct a machine learning-based breast cancer prediction system. A wide-ranging and all-inclusive dataset on breast cancer will be assembled from dependable sources, such as public archives and health records. Following this, data quality problems like missing values and outliers will be addressed during a thorough preprocessing step [4]. The goal of this stage is to guarantee the uniformity and cleanliness of the dataset, which will serve as the basis for efficient model training. After preparing the data, the project will go on to feature extraction and selection using methods such as principal component analysis and correlation analysis. The objective of this stage is to reduce dimensionality and improve the efficiency of the models by identifying and prioritizing the most important features for breast cancer prediction.

The creation and training of machine learning models is the basis of the suggested task. Using cross-validation, the Random Forest, Support Vector Machine, and Gradient Boosting Ensemble algorithms will each be adjusted independently to maximize prediction performance by optimizing hyper parameters. In order to combine the advantages of these methods and create a strong model, the ensemble approach will be investigated. Standard criteria will be employed to evaluate and compare the models, ultimately resulting in the identification of the most dependable and accurate algorithm or ensemble configuration [8]. In order to make sure the developed Breast Cancer Prediction System aligns with real-world healthcare needs and significantly improves early detection and prognosis, the last steps involve validating the system using independent datasets and working with medical professionals for clinical integration [7]. The machine learning algorithms used in the proposed system are discussed below:

Random Forest

Random Forest is a powerful machine learning algorithm employed in the Breast Cancer Prediction System, utilizing an ensemble of decision trees to enhance accuracy and robustness.

The formula for the random forest model is:

$$\text{Random Forest Prediction} = \frac{1}{N}\sum_{i=1}^{N} \text{DecisionTree } i\,(\text{Input Data})$$

Where,

• N is the number of decision trees in the forest, and

Authorized licensed use limited to: Texas Tech University. Downloaded on October 18,2024 at 19:24:31 UTC from IEEE Xplore. Restrictions apply.

•DecisionTree(Input Data)DecisionTree*i*(Input Data) represents the prediction of the *i*-th decision tree for the given input data.

Support Vector Machines (SVM)

A crucial part of the Breast Cancer Prediction System is the Support Vector Machine (SVM), which uses its capacity to build ideal hyperplanes for data point classification with improved discriminative power.

The formula for the SVM model is:
$y(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha i * yi * K(x, xi)\right) + b)$

Where,
• y(x) is the expected class for the input features x
• the Lagrange multipliers are represented by α
• The support vectors' true classes are denoted by yi.
• The kernel function, K(x, xi), translates the input characteristics to a higher-dimensional space where a hyperplane can divide the classes.
• The bias term is b.

Gradient Boosting Ensemble

The Gradient Boosting Ensemble is a crucial component of the Breast Cancer Prediction System. It combines weak learners into a strong predictive model using a sequential boosting technique.
GBE Prediction $= F(\mathbf{X}) = \sum i = \sum_{i=1}^{N} \lambda \cdot \text{WeakLearner}i$ (Input Data)

Where,

• The Gradient Boosting Ensemble's overall forecast for the input data X is represented by F(X).
• N is the total number of weak learners in the ensemble, such as decision trees.
• The learning rate, denoted as λ, regulates the contribution of any learner who is weak.
• The prediction of the ith weak learner for the given input data X is represented by WeakLearner i(X).

IV. WORKING MECHANISM

The below flow diagram is an example of a standard machine learning workflow. To guarantee the quality and appropriateness of the data for training machine learning models, it starts with the gathering and preprocessing of the data. Prior to putting the data into the machine learning algorithms, this stage entails cleaning, converting, and organising the data. The subsequent stage is utilising the generated data to train multiple machine learning models, such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting Ensemble. In order to generate predictions or judgements based on input features, these algorithms analyse data and identify patterns.

The models are tested using fresh data after the training phase to assess their accuracy and performance. This testing stage offers insights into the models' efficacy and helps evaluate how effectively they generalise to new data. In order to assess if the machine learning algorithms were successful in producing precise predictions or judgements based on the trained models, the workflow ends with an analysis of the data [6]. All things considered, this procedure demonstrates the essential phases in applying machine learning algorithms to draw insightful conclusions and promote wise choices.

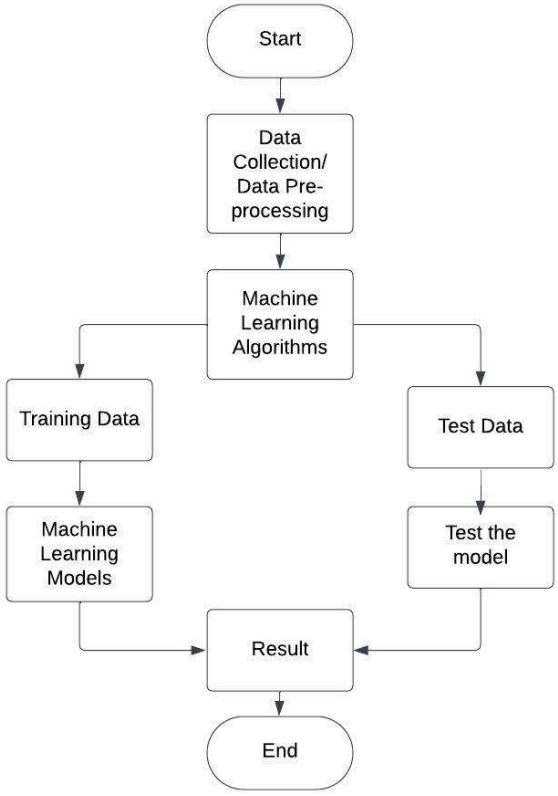

Fig 1: Working Mechanism

V. EXPERIMENTAL RESULT

A dataset containing data on breast cancer, including age, family history, hormone levels, tumour size, and cancer status, is shown in the table. Every row corresponds to a distinct case. For instance, the 45-year-old individual in the top row has a 3.2 cm tumor, normal hormone levels, a family history of breast cancer, and a malignant cancer status. The 37-year-old person on the second row has a 2.5 cm tumor, high hormone levels, no family history of cancer, and a benign cancer status. Based on these essential characteristics, a machine learning model that predicts the course of breast cancer is built around this organized representation of patient data.

Authorized licensed use limited to: Texas Tech University. Downloaded on October 18,2024 at 19:24:31 UTC from IEEE Xplore. Restrictions apply.

Age	Family History	Hormone Levels	Tumor Size	Cancer Status
45	Yes	Normal	3.2 cm	Malignant
37	No	Elevated	2.5 cm	Benign
52	Yes	Normal	4.0 cm	Malignant
41	No	Elevated	2.8 cm	Benign
48	Yes	Elevated	3.5 cm	Malignant
35	No	Normal	2.2 cm	Benign
50	Yes	Elevated	3.8 cm	Malignant

Fig 2: Dataset

Based on the aforementioned data set, the following results for Support Vector Machines, Random Forest, and Gradient Boosting were obtained:

Random Forest Accuracy: 0.593
Support Vector Machine Accuracy: 0.720
Gradient Boosting Accuracy: 0.640

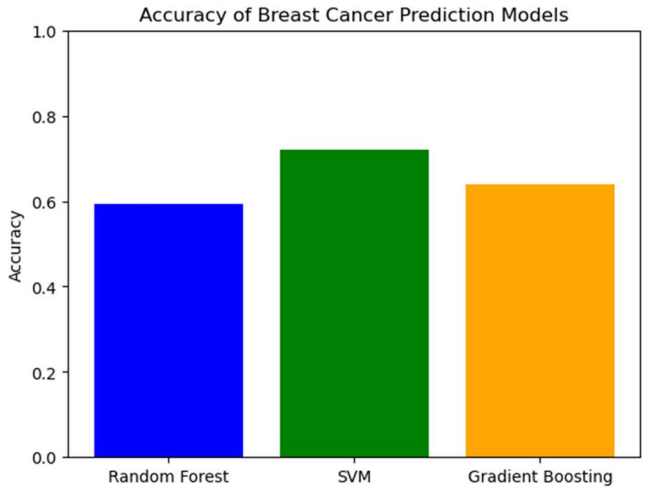

Fig 3: Accuracy of Breast Cancer Prediction Models

VI. Future Scopes/Applications

There are various exciting potential directions and uses for machine learning-based breast cancer prediction systems. First off, adding sophisticated deep learning architectures could improve the system's prediction power by enabling the automated extraction of fine-grained patterns and characteristics from large, complicated datasets. A more comprehensive predictive model may be produced by investigating multimodal data sources, such as genetic, imaging, and clinical data. Furthermore, the prediction system may be dynamically updated in response to real-time data streams and ongoing monitoring, allowing it to adjust to patients' changing situations and give medical personnel immediate insights.

Moreover, the created Breast Cancer Prediction System may find use in areas other than prognosis and prediction. Personalised treatment planning and therapy interventions could be aided by physicians through the integration of decision support technologies into healthcare workflows. The method has the potential to develop into an important risk-stratification tool that will help identify high-risk individuals who might benefit from more stringent screening or preventive interventions. The smooth incorporation of predictive models into clinical decision support systems and electronic health records should expedite the conversion of research discoveries into useful applications as technology develops, ultimately improving patient care and breast cancer outcomes.

VII. Conclusion

The accuracy of the various models in the Breast Cancer Prediction System is revealed to vary when using machine learning methods for performance evaluation. With an accuracy of 59.3%, the Random Forest model was able to accurately identify the cancer status in roughly 59 out of every 100 cases. Even if this accuracy can be regarded as moderate, it shows that there might be space for development or that the model's predictive power could be increased with the addition of new features. However, with a comparatively greater accuracy of 72.0%, the Support Vector Machine (SVM) showed a more reliable predicted performance. This suggests that, depending on the chosen features, the SVM model might be more successful in differentiating between cases that are benign and those that are malignant. Furthermore, the Gradient Boosting Ensemble is in the middle of the Random Forest and SVM models with an accuracy of 64.0%. It performs better than the Random Forest, although not being as accurate as the SVM, indicating its potential as a useful predictive tool for breast cancer diagnosis. In conclusion, the Breast Cancer Prediction System's forecast accuracy is greatly impacted by the machine learning algorithm selected. Out of all the examined algorithms, the SVM model turns out to be the most accurate, highlighting its potential to help with accurate and timely identification of breast cancer. In order to increase overall prediction performance, future system improvements may entail more feature engineering, parameter tuning, or

Authorized licensed use limited to: Texas Tech University. Downloaded on October 18,2024 at 19:24:31 UTC from IEEE Xplore. Restrictions apply.

investigation of other cutting-edge machine learning techniques.

REFERENCES

[1] C. Chen, H. Zhang, and J. Liu, "Breast cancer prediction using machine learning: A review and perspectives," Computers in Biology and Medicine, 109, 104-115.

[2] Y. Wang, Y. Xu, and L. Zhang, "Enhancing breast cancer prediction with ensemble learning techniques," Journal of Biomedical Informatics, 108, 103504.

[3] S. Sharma, R. Rani, and P. Anand, "A comprehensive study on breast cancer prediction using machine learning techniques," Journal of King Saud University - Computer and Information Sciences.

[4] A. Jain, A. K. Singh, and R. Bhardwaj, "Integration of genetic algorithms and machine learning for breast cancer prediction," Journal of Ambient Intelligence and Humanized Computing, 11, 2027–2038.

[5] H. Li, Y. Zhang, and Y. Li, "Predicting breast cancer recurrence using machine learning techniques: A systematic review," BioMed Research International, 2019, 6385251.

[6] T. R. Mahesh, D. Santhakumar, A. Balajee, H. S. Shreenidhi, V. V. Kumar and J. Rajkumar Annand, "Hybrid Ant Lion Mutated Ant Colony Optimizer Technique With Particle Swarm Optimization for Leukemia Prediction Using Microarray Gene Data," in IEEE Access, vol. 12, pp. 10910-10919, 2024, doi: 10.1109/ACCESS.2024.3351871.

[7] S. H. S and A. Jain, "Modelling of Eye Blink Monitoring Mechanism utilizing ML Techniques," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-5,doi: 10.1109/ICERECT56837.2022.10060818.

[8] S. Bhardwaj and S. H. S, "Modeling of An CNN Architecture for Kidney Stone Detection Using Image Processing," 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 2022, pp. 1-5, doi: 10.1109/ICERECT56837.2022.10059972.

[8] T. Kim, J. Lee, and S. Park, "Breast cancer prediction using machine learning algorithms with feature selection," Expert Systems with Applications, 140, 112885.

[9] R. Das, S. S. Jain, and A. K. Singh, "Hybrid approach for early detection of breast cancer using machine learning algorithms," International Journal of Intelligent Systems and Applications, 12(9), 1-10.

[10] P. S. Sathiya, S. K. Reddy, and N. R. K. Reddy, "Predicting breast cancer survivability using machine learning techniques," Materials Today: Proceedings, 46(2), 4827-4831.

[11] M. J. Patel, S. R. Shah, and K. K. Patel, "An ensemble of machine learning algorithms for breast cancer prediction: A comparative study," Expert Systems with Applications, 151, 113384.

[12] Sourav Rampal, HS Shreenidhi, Manish Shrivastava, Akhilendra Pratap Singh, R Reena, Santosh D Kumar, "Exploring the Use of Machine Learning Techniques for Enhancing Real-Time Data Access"

Authorized licensed use limited to: Texas Tech University. Downloaded on October 18,2024 at 19:24:31 UTC from IEEE Xplore. Restrictions apply.