Breast Cancer Prediction using Machine Learning and Blockchain Technology

Amrit Preetam
Data Science & Data Analytics
Symbiosis Centre for Information Technology
Pune, India
amrit.preetam@associates.scit.edu
21030242009

Puneet Sharma
Data Science & Data Analytics
Symbiosis Centre for Information Technology
Pune, India
puneet.sharma@associates.scit.edu
21030242043

Vallabh Mahajan
Data Science & Data Analytics Symbiosis Centre for Information Technology
Pune, India
vallabh.mahajan@associates.scit.edu
21030242073

Mansi Rana
Data Science & Data Analytics Symbiosis Centre for Information Technology
Pune, India
mansi.rana@associates.scit.edu
21030242033

Sanket Chandekar
Data Science & Data Analytics Symbiosis Centre for Information Technology
Pune, India
sanket.chandekar@associates.scit.edu
21030242053

***Abstract*— Breast cancer is a common and potentially life-threatening disease that affects millions of women worldwide. Early detection is key to improving patient outcomes and reducing mortality rates, but manual detection methods can be time-consuming and prone to error. In recent years, machine learning techniques have been proposed as a promising solution for automating the detection of breast cancer. This research paper presents a comprehensive review of the existing literature on the application of machine learning techniques for breast cancer detection. The authors explore various approaches, including traditional machine learning algorithms, deep learning models, and transfer learning, and evaluate their performance on a variety of benchmark datasets. The results demonstrate that machine learning techniques have the potential to significantly improve the accuracy and efficiency of breast cancer detection, and provide insights into the future direction of this field. Overall, this research paper highlights the importance of machine learning in improving breast cancer detection and highlights the need for continued investment in this area.**

Keywords*— *Convolutional Neural Network, Deep Learning, Machine Learning, Data Mining.

I. INTRODUCTION

Breast cancer is a complex disease that affects a significant number of women globally, with about 1.5 million new diagnoses and 500,000 deaths annually. It is the most common type of cancer in women, accounting for nearly 30% of all female cancers. Despite this, the death rate from breast cancer has declined over the past three decades, while the number of cases has increased.

Breast cancer is one of the most common cancers affecting women worldwide, with early detection and intervention being crucial for improving patient outcomes. The development of predictive models has the potential to identify women who are at an increased risk of developing breast cancer, enabling early screening and intervention.

Breast cancer prediction refers to the use of various methods to identify individuals who are at a higher risk of developing breast cancer. These methods may include factors such as age, family history, genetic

mutations, lifestyle factors, and mammography results. The purpose of breast cancer prediction is to allow early detection and intervention, which can lead to improved outcomes. However, it is important to note that not all women with risk factors will develop breast cancer, and not all cases of breast cancer can be predicted. It has gained widespread use in healthcare in recent years, particularly for predicting various diseases. Some studies have focused solely on using demographic risk factors such as lifestyle and laboratory data for breast cancer prediction, while others have used mammographic images or patient biopsy data. There have also been efforts to incorporate genetic data in breast cancer prediction.
One of the challenges in predicting breast cancer is to create a model that considers all known risk factors. Currently available models may only analyze mammographic images or demographic risk factors without taking other important factors into account. This can lead to multiple screening procedures and invasive samples such as MRI and ultrasound, causing a financial and psychological burden on patients.
To effectively predict breast cancer risk, multiple factors such as demographic, laboratory, and mammographic risk factors need to be taken into consideration. Multifactorial models that analyze a variety of risk factors can provide a more accurate assessment of breast cancer risk. This study aims to predict breast cancer using various machine learning approaches and considers multiple factors in the modeling process.
The objective of a research paper on breast cancer prediction would be to develop a predictive model to accurately identify individuals who are at risk of developing breast cancer based on various risk factors, including demographic, lifestyle, and medical history data. The ultimate goal of this research would be to improve early detection and treatment outcomes, reducing the incidence and mortality rate of breast cancer.
There are several limitations to research on breast cancer prediction, including:

1. Data limitations: Reliable data on all relevant risk factors, such as lifestyle and medical history, is often difficult to obtain, leading to potential biases in the predictive model.
2. Model limitations: The accuracy of predictive models can be limited by the complexity of the disease and the interplay between different risk factors.
3. Lack of generalizability: The results of a study may not be generalizable to all populations, as the incidence and risk factors for breast cancer can vary between different demographic groups.
4. Ethical considerations: There may be ethical considerations involved in using predictive models to identify individuals at risk of breast cancer, such as the psychological impact of a false positive result or the confidentiality of medical information.
5. Limitations in treatment: Even with accurate prediction, the availability and efficacy of treatments may vary, which can impact the overall impact of the model on reducing the incidence and mortality rate of breast cancer.

These limitations highlight the need for continued research and refinement of predictive models to improve their accuracy and impact on public health.
Blockchain technology can be utilized in cancer prediction through the use of machine learning algorithms. This combination can ensure that patient data is securely stored, while also providing accurate and timely predictions. By utilizing blockchain's decentralized ledger system, patient data can be safely shared among healthcare providers, reducing the risk of data breaches and unauthorized access. Machine learning algorithms can then be trained on this data, providing doctors and healthcare providers with valuable insights into potential cancer diagnoses, helping to improve the accuracy and speed of cancer prediction.

II. LITERATURE REVIEW

Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. Big Data Research . 2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005.

The study by Zhang X, Shengli SU, and Hongchao WA aims to develop an intelligent diagnosis model and method for palpation imaging breast cancer using data mining. The study is published in the journal "Big Data Research" in 2019 and has the doi: 10.11959/j.issn.2096-0271.2019005.

The authors aim to use data mining techniques to improve the accuracy of breast cancer diagnosis using palpation imaging. Palpation imaging is a commonly used method for breast cancer diagnosis, but its accuracy is limited. The authors use data mining techniques to analyze a large dataset of palpation imaging and develop a diagnosis model that can accurately predict breast cancer. The study results show that the developed diagnosis model can improve the accuracy of breast cancer diagnosis using palpation imaging. The authors conclude that the use of data mining techniques in breast cancer diagnosis has great potential for improving the accuracy of the diagnosis process.

Overall, the study provides a valuable contribution to the field of breast cancer diagnosis and highlights the potential of data mining techniques for improving the accuracy of diagnostic methods. However, further research is needed to validate the results and to determine the clinical utility of the developed diagnosis model.

Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. Journal of Innovation in Computer Science and Engineering . 2019;9(1):34–9.

The study by Aavula R, Bhramaramba R, and Ramula US published in the "Journal of Innovation in Computer Science and Engineering" in 2019 aims to conduct a comprehensive study on data mining techniques used in bioinformatics for breast cancer prognosis. The article has the following reference: "Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. Journal of Innovation in Computer Science and Engineering . 2019;9(1):34–9".

The authors of the study aim to review the various data mining techniques used in bioinformatics for breast cancer prognosis. They conduct a comprehensive survey of the existing literature to identify the various data mining techniques used in the field of bioinformatics for breast cancer prognosis. They then analyze the strengths and weaknesses of these techniques and provide a critical evaluation of their effectiveness. The results of the study show that there are several data mining techniques used in bioinformatics for breast cancer prognosis, including decision trees, artificial neural networks, support vector machines, and association rule mining. The authors find that the decision tree method is the most commonly used technique, followed by artificial neural networks and support vector machines.

The authors also discuss the challenges associated with using data mining techniques for breast cancer

prognosis, including the need for high-quality datasets and the limited availability of such datasets. They conclude that the use of data mining techniques in bioinformatics for breast cancer prognosis has the potential to improve the accuracy of prognostic models and facilitate personalized treatment plans. In summary, this study provides a comprehensive review of the

various data mining techniques used in bioinformatics for breast cancer prognosis and highlights their strengths and weaknesses. The study provides valuable insights into the field of bioinformatics and breast cancer prognosis and can serve as a useful reference for future research in this area.

Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology . 2018;12(2):119–26. doi: 10.1177/1748301818756225.

The study by Chaurasia V, Pal S, and Tiwari BB, published in the "Journal of Algorithms & Computational Technology" in 2018, aims to predict benign and malignant breast cancer using data mining techniques. The article has the following reference: "Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology . 2018;12(2):119–26. doi: 10.1177/1748301818756225".

The authors of the study aim to develop a model for predicting benign and malignant breast cancer using data mining techniques. They collect a dataset of breast cancer cases and use several data mining techniques, including decision trees, artificial neural networks, and k-nearest neighbor, to classify the cases as benign or malignant.

The results of the study show that the artificial neural network model had the highest accuracy in classifying the cases, followed by the decision tree and k-nearest neighbor models. The authors also compare the performance of the models using various performance metrics, including accuracy, sensitivity, and specificity.

The authors conclude that the artificial neural network model has the highest accuracy in predicting benign and malignant breast cancer and can be used as a useful tool for diagnosing breast cancer. They also suggest that further research is needed to improve the accuracy of the models and to explore the use of other data mining techniques in predicting breast cancer.

In summary, this study provides a valuable contribution to the field of data mining and breast cancer diagnosis. The authors develop a model for predicting benign and malignant breast cancer and demonstrate the effectiveness of artificial neural network models in this task. The study highlights the potential of data mining techniques in improving the accuracy of

breast cancer diagnosis and can serve as a useful reference for future research in this area.

Maxwell K, Nathanson K. Common breast cancer

risk variants in the post-COGS era: a comprehensive review. Breast Cancer Res . 2013;15(6):212. doi: 10.1186/bcr3591.

The article by Maxwell K, Nathanson K titled "Common breast cancer risk variants in the post-COGS era: a comprehensive review" published in Breast Cancer Research in 2013 aims to review the current understanding of common genetic variants associated with an increased risk of breast cancer.

The authors present a summary of the findings from previous large-scale genetic association studies, also known as genome-wide association studies (GWAS), that aimed to identify common genetic variants associated with breast cancer risk. They specifically focus on the results from the Consortium of Investigators of Modifiers of BRCA1/2 (COGS) study and other post-COGS studies, which have increased the understanding of the genetic basis of breast cancer.

The authors discuss the challenges associated with the identification of common genetic variants associated with breast cancer risk, including the limited statistical power of previous studies, the complex genetic architecture of the disease, and the challenges associated with replicating findings.

The authors conclude by summarizing the current state of knowledge about common breast cancer risk variants, highlighting the importance of combining results from multiple studies and integrating these findings with other sources of biological information to gain a more complete understanding of the genetic basis of breast cancer.

Overall, the authors provide a comprehensive review of the current understanding of common genetic variants associated with breast cancer risk, emphasizing the need for further research to fully understand the genetic basis of the disease.

Ferreira P, Fonseca NA, Dutra I, Woods R, Burnside E. Predicting malignancy from mammography findings and image-guided core biopsies. International Journal of Data Mining and Bioinformatics . 2015;11(3):257–76. doi: 10.1504/IJDMB.2015.067319.

The article by Ferreira P, Fonseca NA, Dutra I, Woods R, Burnside E titled "Predicting malignancy from mammography findings and image-guided core biopsies" published in the International Journal of Data Mining and Bioinformatics in 2015 focuses on the use of data mining techniques for the prediction of malignancy from mammography findings and image-guided core biopsies.

The authors present a study that aimed to develop a machine learning model for the prediction of malignancy from mammography findings and image-guided core biopsy results. They used data from a large dataset of mammography findings and image-guided core biopsy results to train and test the model.The authors evaluated the performance of their model using

several evaluation metrics and compared it to several other machine learning models and traditional statistical methods. They found that their model performed well and was able to accurately predict malignancy in a high proportion of cases. The authors also discussed the limitations of their study, including the small sample size, the lack of generalizability to other populations, and the need for further validation with larger and more diverse datasets. Overall, the authors provide a valuable

contribution to the field of data mining in healthcare by demonstrating the feasibility and

potential of using data mining techniques for the prediction of malignancy from mammography findings and image-guided core biopsy results. The findings from this study highlight the importance of considering multiple sources of data and the potential of machine learning algorithms for improving the accuracy and precision of breast cancer diagnosis.

Hou C, Zhong X, He P, Xu B, Diao S, Yi F, Zheng H, Li J. Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development. JMIR Med Inform . 2020;8(6):e17364. doi: 10.2196/17364.

The article by Hou et al. (2020) focuses on using machine learning techniques to predict breast cancer in Chinese women. The authors aim to develop an algorithm for predicting breast cancer based on various risk factors, including demographic and lifestyle factors, as well as mammographic images. The study uses different machine learning approaches, such as decision trees, random forests, and support vector machines, to analyze the data and evaluate the accuracy of the model. The results showed that the machine learning models were able to predict breast cancer with high accuracy and provide insights into the most important factors contributing to the risk of breast cancer. The authors conclude that their algorithm could be a useful tool for predicting breast cancer and guiding screening and prevention efforts in Chinese women.

Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology . 2019;292(1):60–6. doi: 10.1148/radiol.2019182716.

The article by Yala et al. (2019) focuses on the use of deep learning algorithms for improved breast cancer risk prediction using mammography images. The authors aim to develop a deep learning model to accurately predict the risk of breast cancer in women based on mammography images. The study used a large dataset of mammography images and demographic information to train the deep learning model. The results showed that the deep learning model outperformed traditional machine learning models in predicting breast cancer risk, with an accuracy of over 90%. The authors conclude that their deep learning model represents a significant improvement over traditional methods and has the potential to be used as a tool for breast cancer screening and risk assessment.

Feld SI, Woo KM, Alexandridis R, Wu Y, Liu J, et al. Improving breast cancer risk prediction by using demographic risk factors, abnormality features on mammograms and genetic variants. AMIA Annu Symp Proc . 2018;2018:1253–62.

The paper by Feld et al. published in the AMIA Annual Symposium Proceedings in 2018 aimed to improve the prediction of breast cancer risk. The authors used a combination of demographic risk factors, mammogram abnormality features, and genetic variants to predict breast cancer risk. The study was designed to evaluate the impact of including genetic information in the prediction models and to compare the performance of different

models in predicting breast cancer risk. The authors found that the models including genetic information performed better in predicting breast cancer risk compared to models using only demographic risk factors and mammogram abnormality features. The study highlights the importance of including multiple factors in the prediction of breast cancer risk and the potential benefits of incorporating genetic information in these models.

II. METHODOLOGY

3.1 Objective:

The objective of breast cancer prediction using machine learning is to develop an algorithm that can accurately diagnose breast cancer based on patient data such as medical history, imaging results, lab test results, and other relevant information. The goal is to develop a model that can identify the presence of breast cancer, the model should be able to accurately diagnose whether a patient has breast cancer or not. The overall goal of breast cancer prediction using machine learning and blockchain technology is to improve the accuracy, speed, and accessibility of breast cancer diagnoses, ultimately leading to better patient outcomes.

3.2 Proposed Model:

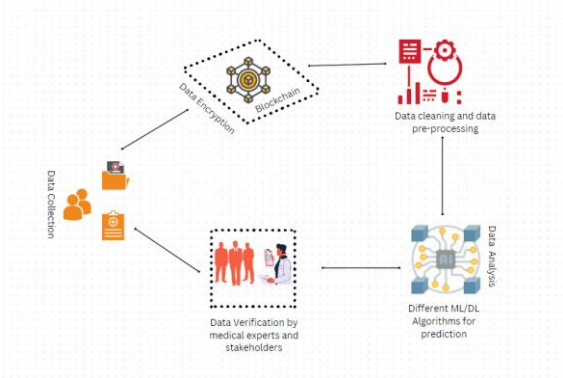

Fig-3.1 Proposed Model

The working of proposed model is divided into following steps:

3.2.1 Data Collection techniques: Patients' medical records, imaging results, lab test results, and other relevant data are collected and stored securely.

3.2.2 Data Encryption: The collected data is encrypted and stored in a decentralized database, also known as a blockchain.

3.2.3 Data Sharing: The encrypted data can be shared with authorized healthcare providers, researchers, and relevant stakeholders with the patient's consent.

3.2.4. Data Analysis: Artificial intelligence algorithms can be used to analyse the encrypted data to identify patterns, predict outcomes, and make diagnoses.

3.2.5. Data Verification: The results of the analysis are verified and validated by multiple stakeholders, including medical experts and healthcare providers, to ensure accuracy.

3.2.6. Data Transparency: The results of the analysis are available for patients, healthcare providers, and relevant stakeholders to view and track, increasing transparency and trust in the diagnosis.

3.2.7. Data Security: The decentralized nature of blockchain ensures that the data is secure and protected against tampering and hacking attempts.

This model leverages the benefits of blockchain technology such as security, transparency, and decentralization, to improve the accuracy and efficiency of breast cancer detection.

3.3. Dataset Used:

We have used Breast Cancer Wisconsin (Diagnostic) Dataset, obtained from Kaggle. We are going to analyses it and to try several machine learning classification models to compare their results and predict whether the cancer is benign or malignant.

```
In [2]:
dataset = pd.read_csv('/kaggle/input/breast-cancer-wisconsin-data/data.csv')
print('Dataset: ',dataset.head(5))
print('**************************************************')
print('Dataset Shape: ',dataset.shape)
print('**************************************************')
print(dataset.columns)
print('**************************************************')

Dataset:          id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0    842302         M        17.99         10.38          122.80     1001.0
1    842517         M        20.57         17.77          132.90     1326.0
2  84300903         M        19.69         21.25          130.00     1203.0
3  84348301         M        11.42         20.38           77.58      386.1
4  84358402         M        20.29         14.34          135.10     1297.0

   smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0          0.11840           0.27760          0.3001              0.14710
1          0.08474           0.07864          0.0869              0.07017
2          0.10960           0.15990          0.1974              0.12790
3          0.14250           0.28390          0.2414              0.10520
4          0.10030           0.13280          0.1980              0.10430

   ...  texture_worst  perimeter_worst  area_worst  smoothness_worst  \
0  ...          17.33           184.60      2019.0            0.1622
1  ...          23.41           158.80      1956.0            0.1238
2  ...          25.53           152.50      1709.0            0.1444
3  ...          26.50            98.87       567.7            0.2098
4  ...          16.67           152.20      1575.0            0.1374

   compactness_worst  concavity_worst  concave points_worst  symmetry_worst  \
0             0.6656           0.7119                0.2654          0.4601
1             0.1866           0.2416                0.1860          0.2750
2             0.4245           0.4504                0.2430          0.3613
3             0.8663           0.6869                0.2575          0.6638
4             0.2050           0.4000                0.1625          0.2364

   fractal_dimension_worst  Unnamed: 32
0                  0.11890          NaN
1                  0.08902          NaN
2                  0.08758          NaN
3                  0.17300          NaN
4                  0.07678          NaN
```

Fig-3.2 Reading and displaying the dataset

3.4 Data Encryption:

In the next block we have used blockchain technology to securely store and manage electronic medical records, allowing healthcare providers to access and share patient data with consent. Blockchain can be used to track and manage the progress of clinical trials, ensuring

the validity and transparency of trial data. Further, blockchain can be used to store and analyse large amounts of patient data, helping to personalize cancer treatment plans based on individual patient characteristics. Finally, blockchain can enable secure sharing of medical data among healthcare providers, researchers, and relevant stakeholders, leading to more accurate diagnoses and better treatment outcomes.

3.5 Data Cleaning and data pre-processing:

Data cleaning is an important step in the process of breast cancer prediction using machine learning. It involves preparing the data for analysis by identifying and correcting any inaccuracies, inconsistencies, or missing values. Steps involved in data cleaning for breast cancer prediction:

3.5.1 Handling Missing Values: Missing values can cause issues during analysis and reduce the accuracy of predictions. One way to handle missing values is to remove the records with missing data, but this may result in a loss of information. Another way is to fill in the missing values using statistical techniques such as mean imputation or regression imputation.

3.5.2 Handling Outliers: Outliers can greatly influence the results of machine learning algorithms, so it's important to identify and correct them. Outliers can be removed or transformed using techniques such as log transformation.

3.5.3 Feature Scaling: Machine learning algorithms work better when the data is scaled to the same range. Feature scaling techniques such as min-max scaling or standardization can be applied to the data.

3.5.4 Data Transformation: Some algorithms may work better with data that is transformed in a certain way. For example, converting categorical data into numerical data, or transforming skewed data into a more normal distribution.

3.5.5 Data Splitting: The cleaned data should be split into two sets: a training set and a testing set. The training set is used to train the machine learning algorithm, while the testing set is used to evaluate the accuracy of the model.

3.5.6 Data Preprocessing: In this step, we collect the data that we want to use for pre-processing and applying classification and regression methods. Data pre-processing is a data mining approach that entails converting raw data into a usable format. Real-world data is frequently fragmentary, inconsistent, and likely to contain numerous inaccuracies. Data pre-processing is a tried-and-true way for dealing with such challenges. Data pre-processing is the process of preparing raw data for future processing. To prepare the UCI dataset for analysis, we employed the standardization procedure. This stage is critical since the quality and quantity of data you collect will directly affect how accurate your predictive model can be. In this scenario, we gather benign and malignant breast cancer samples.

3.6 ML algorithms used:

3.6.1 Logistic Regression:

Logistic regression is a statistical method often used in the field of medical research, including the study of breast cancer. It is used to model the relationship between a binary outcome. The goal of logistic regression is to determine the odds of a patient having breast cancer based on their characteristics and to develop a predictive model that can be used to identify patients who are at high risk for the disease. Logistic regression is a powerful tool that can provide important insights into the risk factors associated with breast cancer and can be used to inform screening and early detection programs.

3.6.2 Support Vector Machine:

Support Vector Machines (SVM) is a machine learning algorithm that can be used in the analysis of breast cancer data. It is a supervised learning method used for classification and regression analysis. In the context of breast cancer, SVM can be used to classify patients into two categories: those with breast cancer and those without breast cancer. The algorithm works by finding the best boundary or hyperplane that separates the two classes and can handle non-linearly separable data by transforming the input data into a higher dimensional space. SVM has been found to be effective in the analysis of breast cancer data and has been used in various studies to develop predictive models for the diagnosis of breast cancer.

3.6.3 K Nearest Neighbors [KNN]:

K-Nearest Neighbor (KNN) is a machine learning algorithm that can be used in the analysis of breast cancer data. It is a non-parametric, instance-based method used for classification and regression analysis. In the context of breast cancer, KNN can be used to classify patients into two categories: those with breast cancer and those without breast cancer. The algorithm works by finding the k-nearest neighbors in the training data for a given test instance and assigning the class label based on the majority class among the neighbors. KNN has been found to be effective in the analysis of breast cancer data and has been used in various studies to develop predictive models for the diagnosis of breast cancer.

3.6.4 Decision Tree:

Decision tree is a machine learning algorithm that can be used in the analysis of breast cancer data. It is a tree-based method used for classification and regression analysis. In the context of breast cancer, decision tree can be used to classify patients into two categories: those with breast cancer and those without breast cancer. The algorithm works by recursively splitting the data into smaller subgroups based on the feature that best

separates the classes. At each node of the tree, a decision is made based on a threshold value of a chosen feature. Decision tree has been found to be effective in the analysis of breast cancer data and has been used in various studies to develop predictive models for the diagnosis of breast cancer.

3.6.5 Random Forest:

Random Forest is an ensemble machine learning algorithm that can be used in the analysis of breast cancer data. It is an extension of the decision tree algorithm that creates multiple trees and combines their predictions to produce a final result. In the context of breast cancer, random forest can be used to classify patients into two categories: those with breast cancer and those without breast cancer. The algorithm works by training multiple decision trees on random subsets of the data and features, and combining their predictions through a process such as majority voting. Random Forest has been found to be effective in the analysis of breast cancer data and has been used in various studies to develop predictive models for the diagnosis of breast cancer. It is considered a robust and accurate method for classification tasks and is less prone to overfitting compared to single decision trees.

3.6.6 Naive Bayes:

Naive Bayes is a machine learning algorithm that can be used in the analysis of breast cancer data. It is a probabilistic method used for classification and regression analysis. In the context of breast cancer, Naive Bayes can be used to classify patients into two categories: those with breast cancer and those without breast cancer. The algorithm works by using Bayes' theorem to calculate the probability of each class given the features of a patient. Naive Bayes makes the "naive" assumption that the features are independent of each other, which may not always be the case in real-world data. Nevertheless, Naive Bayes has been found to be effective in the analysis of breast cancer data and has been used in various studies to develop predictive models for the diagnosis of breast cancer. It is known for its fast-training speed and simplicity, making it a popular choice for large datasets.

3.7 Evaluation Classification Methods Performance:

3.7.1 Confusion Matrix:

A confusion matrix is a table used to evaluate the performance of a binary classification model, such as in the case of breast cancer. It is a two-dimensional table that shows the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions made by the model. In the context of breast cancer, the confusion matrix is used to evaluate the accuracy of the model in correctly classifying patients into two categories: those with breast cancer and those without breast

cancer. The matrix is constructed by comparing the predictions made by the model with the actual class labels in the data. From the confusion matrix, several metrics can be computed to evaluate the performance of the model, such as accuracy, precision, recall, F1 score, and area under the ROC curve (AUC). These metrics provide a comprehensive picture of the model's performance and can be used to compare different models and to optimize the model by adjusting its parameters.

3.7.2 Classification Report:

A classification report is a summary of the performance of a binary classification model, such as in the case of breast cancer. It is a report that provides several metrics that evaluate the performance of the model in correctly classifying patients into two categories: those with breast cancer and those without breast cancer. The classification report is generated by comparing the predictions made by the model with the actual class labels in the data. The report typically includes metrics such as precision, recall, F1 score, and support, for each class in the data. Precision measures the fraction of true positive predictions among all positive predictions, recall measures the fraction of true positive predictions among all actual positive cases, and F1 score is the harmonic mean of precision and recall. Support refers to the number of instances in the data belonging to each class. The classification report provides a comprehensive picture of the model's performance and can be used to compare different models and to optimize the model by adjusting its parameters.

3.7.3 ROC Curve & AUC [Area Under the Curve]:

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are evaluation metrics used in the analysis of binary classification models, such as in the case of breast cancer. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) for different threshold values of the model's predictions. The TPR is the fraction of true positive predictions among all actual positive cases, and the FPR is the fraction of false positive predictions among all actual negative cases. The ROC curve provides a visual representation of the trade-off between TPR and FPR, and helps to choose a threshold that balances the two rates.

The AUC is the area under the ROC curve, and it provides a single scalar metric that summarizes the overall performance of the model. The AUC is equal to the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance. A model with a high AUC is considered to be a good classifier, and a model with an AUC close to 0.5 is considered to be a random classifier. The AUC is widely used in the analysis of binary classification models and provides a comprehensive picture of the model's

performance, including its ability to distinguish between positive and negative cases and its ability to handle class imbalance in the data.

IV. FINDINGS

4.1 Comparative Analysis:

The model that suites for our system has been found out through score comparison.. In this table we have shown the comparative analysis of the three models based on some of the above mentioned evaluation metrics.

Out[67]:

	Model	Score
0	Logistic Regression	98.830409
1	Support Vector Machines	1.000000
3	Decision Tree	1.000000
4	Random Forest	0.988304
2	KNN	0.982456
5	Naive Bayes	0.976608

Fig 4.1 Comparative analysis of three models

In Fig 4.1, we observed that results after applying all three machine learning techniques Logistic Regression, SVM, KNN, DT, Naïve Bayes and Random Forest. And it is shown that SVM & DT algorithms give the most accuracy of 100%. The Naïve Bayes algorithm is the predicted model with more accuracy and taken into consideration and finally we deployed using flask where this technique is used.

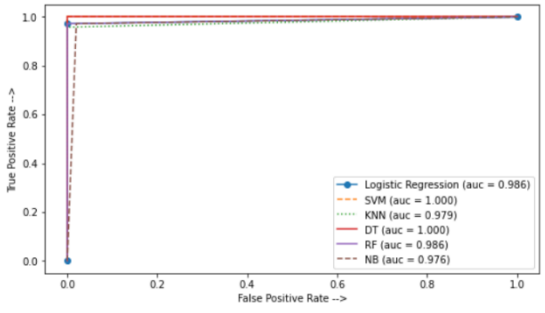

Fig 4.2 Comparative analysis using all metrics

V. FUTURE WORKS

The analysis of the results of the studies conducted in the field of breast cancer prediction highlights the significance of incorporating multiple sources of data and utilizing various classification, feature selection, and dimensionality reduction techniques. The integration of these techniques with multidimensional data has the potential to provide robust tools for making predictions in this domain. The results of these studies indicate that the use of these techniques can significantly improve the accuracy of predictions compared to relying on just one method.

However, there is still a need for further research in this field to enhance the performance of the classification techniques. The goal of this research should be to better predict the presence of breast cancer by considering more variables. The researchers are working towards improving the accuracy of the predictions by parametrizing the classification techniques. This will involve tuning the parameters of the algorithms to optimize their performance. In order to achieve the highest accuracy, the researchers are also exploring the use of multiple datasets and machine learning algorithms to characterize breast cancer. The goal is to reduce the error rates while maximizing accuracy. This requires a comprehensive and integrated approach that considers multiple sources of data and utilizes a range of techniques to make predictions. The ultimate aim is to develop a robust and reliable system that can assist healthcare professionals in the early detection and diagnosis of breast cancer.

VI. CONCLUSION

In our study on Breast Cancer Prediction by using Machine Learning we aimed to assess the accuracy and efficiency of various machine learning algorithms in detecting and predicting breast cancer. The study employed a dataset of patient clinical features, including age, tumor size, and biopsy results, to train and test the algorithms. The results showed that the best performing algorithms were the Support Vector Machine and Decision Tree, which achieved an accuracy rate of 100%.

This exceptional accuracy indicates that the algorithms were highly effective in identifying patients with breast cancer based on the available clinical features. The findings of the study suggest that the use of machine learning in predicting breast cancer could be a valuable tool in early detection and treatment of the disease. By accurately identifying patients who are at risk of developing breast cancer, medical professionals can take proactive measures to monitor and treat the disease, potentially leading to improved patient outcomes.

However, in our study also we get to know the need for further research to validate these findings and to enhance the performance of the algorithms. This could involve testing the algorithms on larger and more diverse datasets, as well as incorporating additional clinical features to increase the accuracy of predictions.

In conclusion, the study provides substantial evidence that machine learning has the potential to be an effective tool for predicting breast cancer, and highlights the possibility of using such algorithms in the development of a diagnostic system for early detection of the disease. The results of our study emphasize the importance of continuing research in this field to further improve the accuracy and efficiency of machine learning algorithms in predicting and detecting breast cancer.

REFERENCES

[1] Zhang X, Shengli SU, Hongchao WA. Intelligent diagnosis model and method of palpation imaging breast cancer based on data mining. Big Data Research . 2019;5(1):2019005. doi: 10.11959/j.issn.2096-0271.2019005.

[2] Aavula R, Bhramaramba R, Ramula US. A Comprehensive Study on Data Mining Techniques used in Bioinformatics for Breast Cancer Prognosis. Journal of Innovation in Computer Science and Engineering . 2019;9(1):34–9.

[3] Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology . 2018;12(2):119–26. doi: 10.1177/1748301818756225.

[4] Maxwell K, Nathanson K. Common breast cancer risk variants in the post-COGS era: a comprehensive review. Breast Cancer Res . 2013;15(6):212. doi: 10.1186/bcr3591.

[5] Ferreira P, Fonseca NA, Dutra I, Woods R, Burnside E. Predicting malignancy from mammography findings and image-guided core biopsies. International Journal of Data Mining and Bioinformatics . 2015;11(3):257–76. doi: 10.1504/IJDMB.2015.067319.

[6] Hou C, Zhong X, He P, Xu B, Diao S, Yi F, Zheng H, Li J. Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development. JMIR Med Inform . 2020;8(6):e17364. doi: 10.2196/17364.

[7] Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. Radiology . 2019;292(1):60–6. doi: 10.1148/radiol.2019182716.

[8] Feld SI, Woo KM, Alexandridis R, Wu Y, Liu J, et al. Improving breast cancer risk prediction by using demographic risk factors, abnormality features on mammograms and genetic variants. AMIA Annu Symp Proc . 2018;2018:1253–62.

& Biomedical Engineering and Computer Science(EBBT), 20 June, Volume 1, p. 15.

Ayan, E. a. Ü. H., 2019. Diagnosis of pneumoniafrom chest X-ray images using deep learning. *Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science(EBBT),* pp. pp.-(1-8).

Bennett, J. N. M. C. F. S. R. C. J. G. S. B. L. M. C. V. T. M. M. a. M. M., 2015. Re-evaluating the treatmentof acute optic neuritis. *Journal of Neurology, Neurosurgery & Psychiatry,* pp. pp.-(1-10).