BREAST CANCER PREDICTION USING MACHINE LEARNING

G. Saranya[1], Shaik Samreen[2], Shaik Roshini[3], Netuluri Akshara Pragna[4]
[1] Professor, [2, 3 & 4] Student
[1]gaddamsaranya4@gmail.com, [2] samreeenshaik@gmail.com, [3] shaikroshini22@gmail.com, [4] netuluripragna@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

***Abstract*— Breast cancer is a big problem, especially for women's health. It's important to raise awareness and support research to find a cure. Accurate prediction is crucial for early detection and effective treatment. Breast cancer starts when there are changes or mutations in the DNA of breast cells. These changes can be influenced by factors like age, family history, hormones, and environmental exposures. It's important to understand these risk factors and take steps towards early detection and prevention. Timely detection greatly improves outcomes. Machine learning is crucial in predicting breast cancer by analyzing large datasets to find patterns. In a recent study, researchers utilized five machine learning methods - Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) - using the Breast Cancer Wisconsin Diagnostic dataset. Their findings revealed SVM as the most effective predictor, boasting an amazing precision rate of 99.2%. This study was conducted within the Python programming language, utilizing the Anaconda environment and the Scikit-learn library. Such advancements in predictive modeling contribute significantly to the ongoing efforts in combating breast cancer.**

Keywords— Risk factors, Early detection, Women, Abnormal cells, Dataset, Python, Anaconda environment, Scikit-learn library, Accuracy, Machine learning algorithms.

I. INTRODUCTION

Breast cancer is among the most typical types of cancer and concerning health issues affecting women worldwide, with profound implications for public health and individual well-being (World Health Organization, 2022). According to the WHO around 2 million people are affected by every year, breast cancer becomes the most common cancer among women globally. It accounts for around 25% of all cancer cases. Breast cancer claimed the lives of almost 685,000 women worldwide in 2022, accounting for an estimated 2.3 million new instances of the disease that were identified globally. An estimated 50% more people will receive a cancer diagnosis in 2040 than in 2020, and the number of cases is expected to rise even more in the upcoming years. [1].

In fact, the most frequent kind of cancer among Indian women is breast cancer, making up about 27% of all cancer cases in females. The ICMR has provided data showing that the incidence of breast cancer has been on the rise..[2]. Breast cancer risk factors include various aspects like family history, genetics, hormones, lifestyle, and environment (American Cancer Society) [3]. Being a woman and getting older are big risk factors. Genetic mutations like BRCA1 and BRCA2 can increase risk, as can certain reproductive factors like starting menstruation early or having children late [4].

Not breastfeeding and hormonal factors like using hormone replacement therapy may also play a role. Lifestyle choices such as alcohol consumption, obesity, and lack of physical activity can increase risk too [5].

When it comes to the danger of acquiring radiation-associated breast cancer, women under the age of twenty have a higher risk than those who are exposed later in life.[6]. While having these risk factors doesn't mean someone will get breast cancer, knowing about them can help people and doctors take steps to reduce risk and catch cancer early.

Researchers are continuously making advancements in breast cancer prediction to better understand the risk levels and personalize treatments. By analyzing these multi-omics data, scientists hope to improve risk stratification and tailor interventions specifically to each patient's needs. It's a fascinating field of study with a bright future. [7].

In recent studies, researchers have highlighted the importance of incorporating multi-omics data into breast cancer prediction models to improve their predictive power and clinical utility [8].

Moreover, the integration of multi-omics data has the potential to uncover new insights into the underlying mechanisms of breast cancer and identify novel therapeutic targets [9]

By combining data from multiple omics platforms, researchers can develop more comprehensive and robust predictive models that account for the heterogeneity of breast cancer and individual patient characteristics [10].

By utilizing machine learning algorithms and evaluating their performance using metrics like confusion matrix, accuracy, precision, and sensitivity, you can determine which algorithm is most effective for diagnosing and predicting breast cancer. This approach will help you identify the most suitable algorithm that can provide accurate results.

II. LITERATURE REVIEW

Previous studies in breast cancer prediction and diagnosis using machine learning have shown promising results. Researchers have explored various machine learning algorithms to develop models for accurate diagnosis. They've also looked into different methods to select the most relevant features related to breast cancer detection, aiming to improve model performance. Additionally, researchers have tested ensemble learning techniques to combine classifiers in an effective manner. They've used evaluation metrics like confusion matrix, accuracy, precision, and recall to assess model performance [11].

Overall, these studies have laid the groundwork for enhancing breast cancer diagnosis using machine learning, providing valuable insights into improving early detection and treatment. The authors Youness Khoudfi et al. [12]in their study, compared different machine learning methods to see which one worked best for diagnosing breast cancer. They focused on Support Vector Machine (SVM) and also looked at K-Nearest Neighbors (K-NN), Random Forest (RF), and Naive Bayes (NB). After testing these methods, they found that SVM was the most accurate, with a success rate of 97.9%.

On the other side, Behravan et al. [13] utilized a database comprising 695 records encompassing demographic risk factors and genetic data to forecast breast cancer. Their research indicated that employing the XGBoost model with multiple factors yielded enhanced performance (AUC=0.788) compared to a model reliant solely on one set of factors (AUC=0.678).

Deng [14] employed the XGBoost algorithm to classify and predict breast cancer, achieving notable performance metrics. Their model achieved an accuracy of 0.96 and a recall of 0.97. Their approach demonstrated high precision in both correctly identifying instances of breast cancer and minimizing false negatives.

Mahesh et al. [15] developed a breast cancer prediction technique using XGBoost ensembles. They addressed data imbalance and noise using SMOTE and employed classifiers like naïve Bayes, decision tree, and random forest combined with XGBoost. The XGBoost-Random Forest ensemble achieved a high accuracy of 98.20% in early breast cancer detection.

Breast cancer prediction and diagnosis, previous research has demonstrated promising outcomes through the application of machine learning techniques. Additionally, researchers have experimented with different feature selection methods to identify the most relevant factors associated with breast cancer detection, aiming to enhance model performance. Furthermore, ensemble learning approaches have been explored to combine classifiers effectively, leading to more robust prediction models. To assess how well these models work, researchers have employed metrics like confusion matrix, accuracy, precision, and recall. Notably, studies conducted by Youness Khoudfi and Mohamed Bahaj highlighted the effectiveness of Support Vector Machine (SVM) in diagnosing breast cancer, achieving a remarkable success rate of 97.9% [16].

Similarly, Behravan and Hartikainen's research showcased the improved performance of the XGBoost model when considering multiple factors, leading to a higher Area Under the Curve (AUC) compared to models relying solely on single factors [17].

Furthermore, Deng et al. [18] demonstrated notable success with the XGBoost algorithm in classifying and predicting breast cancer, achieving high accuracy and recall rates.

Their approach emphasized precision in both accurately identifying instances of breast cancer and minimizing false negatives. Mahesh et al. employed XGBoost ensembles in early breast cancer detection, addressing data imbalance and noise through techniques like SMOTE. Their XGBoost-Random Forest ensemble achieved a remarkable accuracy of 98.20%, showcasing the possible of machine learning in improving breast cancer diagnosis and treatment outcomes.

To assess model efficacy, researchers have employed evaluation metrics such as confusion matrix, accuracy, precision, and recall. Notably, studies by Youness Khoudfi and Mohamed Bahaj highlighted the effectiveness of Support Vector Machine (SVM) in diagnosing breast cancer, achieving a success rate of 97.9%

III. DATASET DESCRIPTION

The WBCD was utilized for this research. 32 characteristics in the dataset are utilized in the breast cancer prediction. The characteristics are specifics extracted from images of cells within a breast tumor. They tell us things like the size, shape, and arrangement of the cell nuclei. Doctors use this information to understand what's going on in the lump and whether it might be cancerous.We obtained the dataset from Kaggle [19]. The dataset contains of 569 rows and 33 columns. In those columns, the first column represents ID number and the second column represents the target variable and the rest 30 columns represent the features.Attributes are shown in fig 1.

```
df.columns
✓ 0.0s
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst', 'Unnamed: 32'],
      dtype='object')
```

Fig. 1. Attributes diagram

IV. PROPOSED WORK

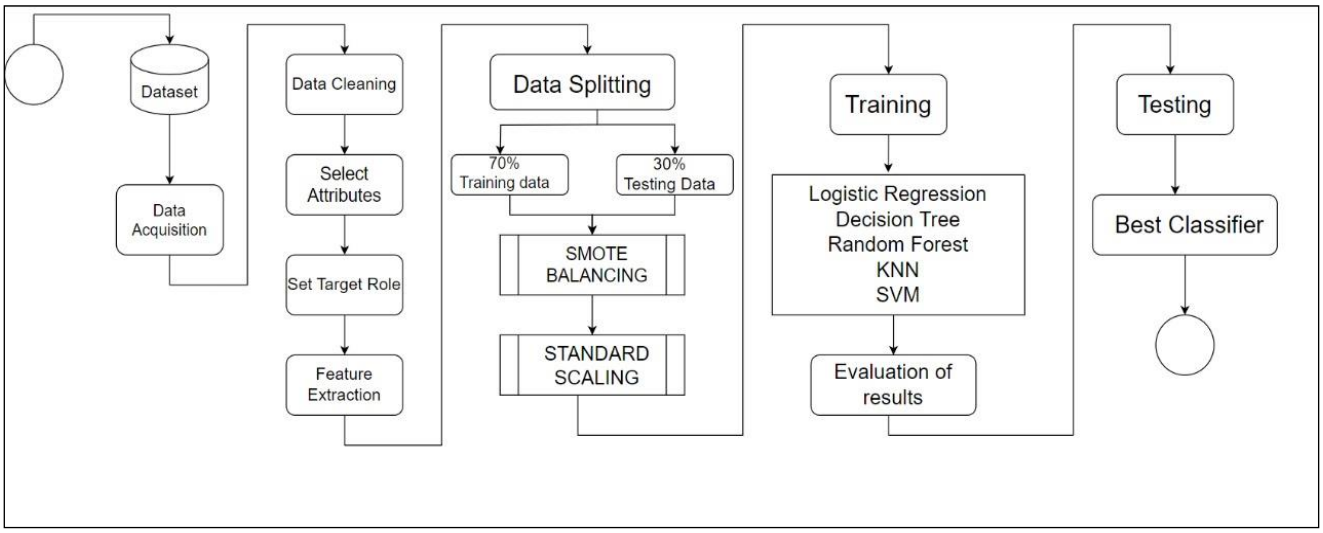

Fig. 2. Work flow diagram

In our experiment, we want to find the ideal method for detecting breast cancer. So, we tested different ones like SVM, Random Forests, Logistic Regression, Decision Tree, and KNN on a dataset called Breast Cancer Wisconsin Diagnostic. Then, we looked at the results to see which algorithm gave us the most accurate predictions. The proposed architecture is detailed in Figure 2.

Our proposed work for predicting breast cancer entails a systematic approach beginning with data acquisition and preprocessing. The process encompasses four essential steps: data cleaning, attribute selection, target role definition, and feature extraction. Ensuring data integrity and relevance, we meticulously address inconsistencies and choose pertinent attributes for prediction while designating the target variable. Additionally, we extract meaningful features to enhance the predictive power of our models. Before model construction, we apply SMOTE to mitigate class imbalance and standard scaling to standardize feature values. Subsequently, we build machine learning algorithms using the preprocessed data, aiming to accurately predict breast cancer outcomes.

Through rigorous evaluation using train-test splitting, we assess model performance and compare results to select the most effective algorithm. Our methodology integrates advanced techniques to improve prediction accuracy, contributing to early detection and enhanced patient care in breast cancer diagnosis.

A. Data Obtaining

In our study, we utilized the Breast Cancer Wisconsin Diagnostic dataset sourced from the University of Wisconsin Hospitals Madison Breast Cancer Database in our research. This dataset comprises features derived from digitized images of breast cancer samples obtained through fine-needle aspirate. It is often utilized for classification tasks, particularly for distinguishing between benign and malignant tumors based on various features derived from digitized images of breast cancer cells.

The dataset contains a total of 569 instances, with 357 classified as benign and 212 as malignant, thereby representing two distinct classes.

The distribution indicates that approximately 62.74% of cases are benign, while 37.26% are malignant. Furthermore, the dataset encompasses 11 integer-valued attributes, including information related to radius, texture, area, perimeter, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

B. Data Preprocessing

In our study, data preprocessing is crucial for ensuring the dataset's quality and relevance in predicting breast cancer:

1. Data cleaning: We carefully deal with missing information and get rid of any identical records to keep our dataset accurate and we find and fix any unusual data points to make sure they don't mess up our predictions. Here in this study, we dropped the attribute id since it does not contribute with the model building then we checked for the null columns and then dropped the unnamed column.

We noticed from Fig. 3 that the column labeled "Unnamed:32" has 569 missing values. So, in the process of data cleaning, we dropped that particular column. Fig. 4 shows the missing value count after data cleaning process. It is clear that after dropping the Unnamed:32 column there are no missing values left in the dataset.

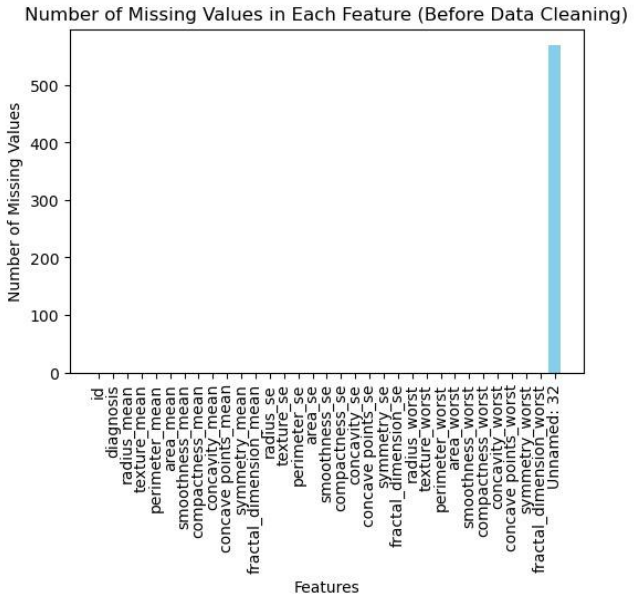

Fig 3. Missing Values before data cleaning

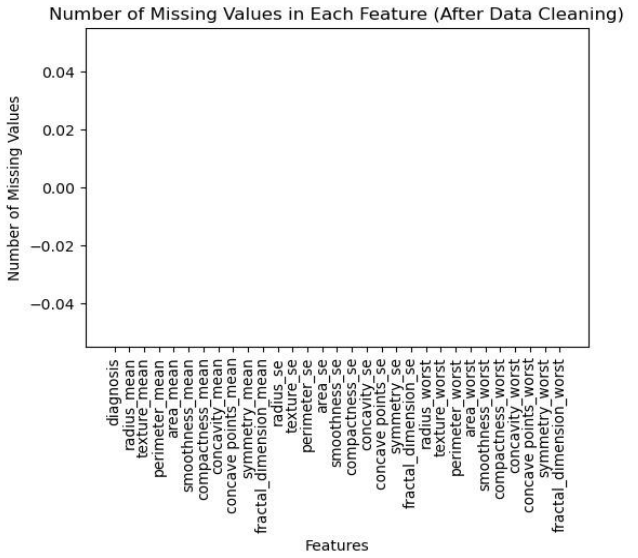

Fig.4. Missing Values after data cleaning

2. Target Variable: The target variable indicates whether breast cancer is present or absent. This binary variable helps our predictive models learn the patterns linked with diagnosing breast cancer.

3. Feature Selection: For feature selection, we used a simple method called a correlation matrix. If two features had a high correlation (above 0.9), we removed one of them. For example in our correlated matrix from below fig, we can see that the value at (radius_mean, perimeter_mean) is 0.99. Since the value exceeded 0.9, we dropped the column perimeter_mean. Similarly we dropped the features whose correlation value exceeded 0.9. This helped us avoid using redundant information in our analysis. After dropping the collinear features, we were remained with 21 features overall. Here is the lower triangle correlated matrix of those 21 features depicted in fig. 5

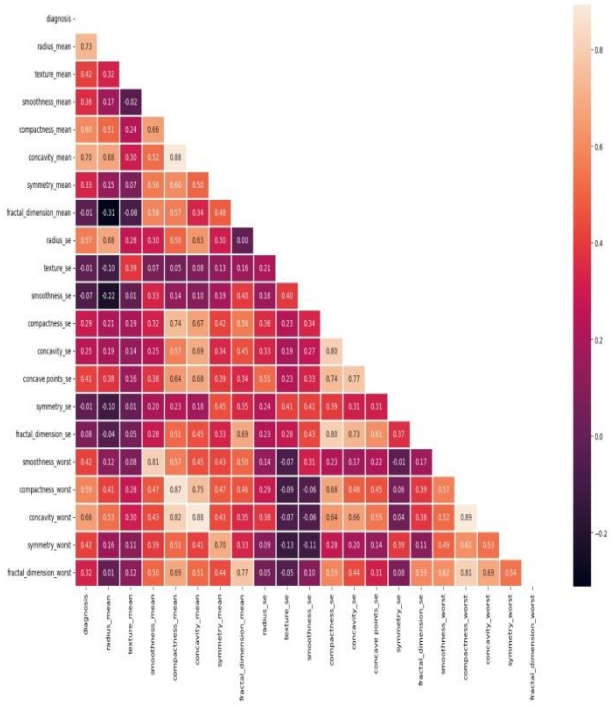

Fig. 5. Lower triangle correlated matrix after feature selection

C. Model Building

1. Data Splitting: We divided our dataset into two parts: a training set, which had 70% of the data, and a testing set, which had 30% of the data of the entire dataset respectively.

2. SMOTE Balancing: To balance the dataset, we used a method called Synthetic Minority Over-sampling Technique (SMOTE). Below fig. 6 and fig. 7 depict the class distribution before and after applying SMOTE balancing technique. This helps make sure that each class has a similar number of examples for better analysis.

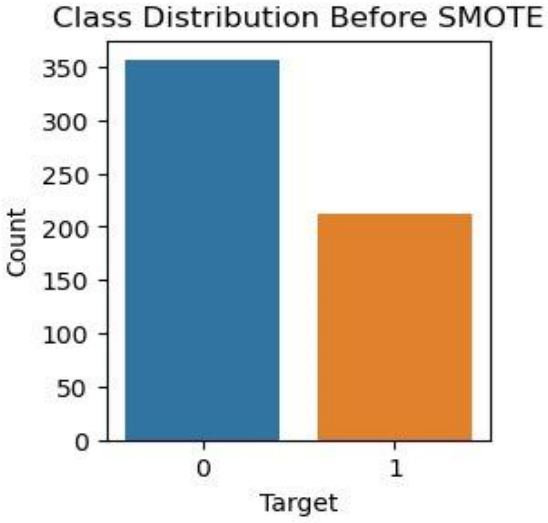

Fig. 6. Before SMOTE

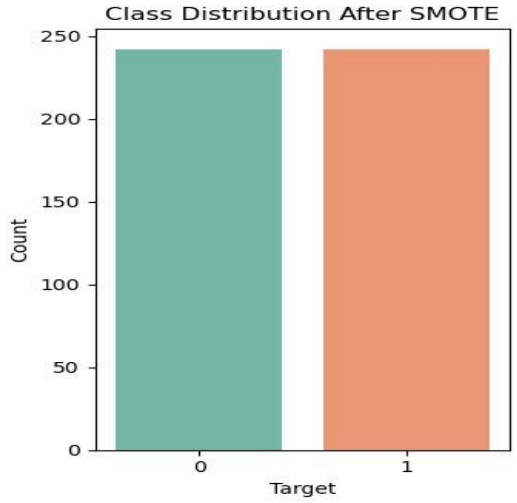

Fig. 7. After SMOTE

3. Standard Scaling: Additionally, we scaled the features using a method called standard scaling. Standard scaling, is a method used to put all features in a dataset on an equal scale.

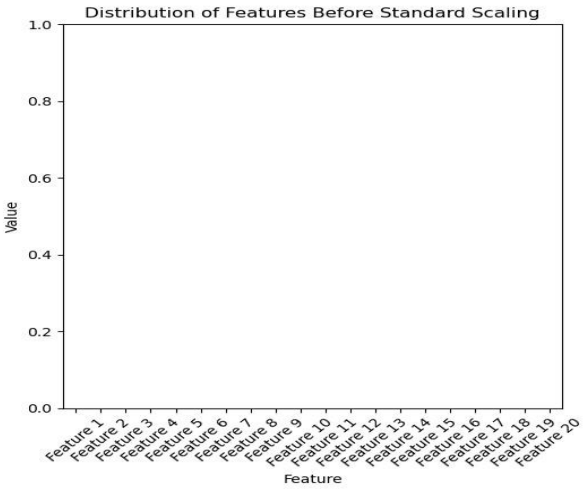

Fig.8. Before scaling

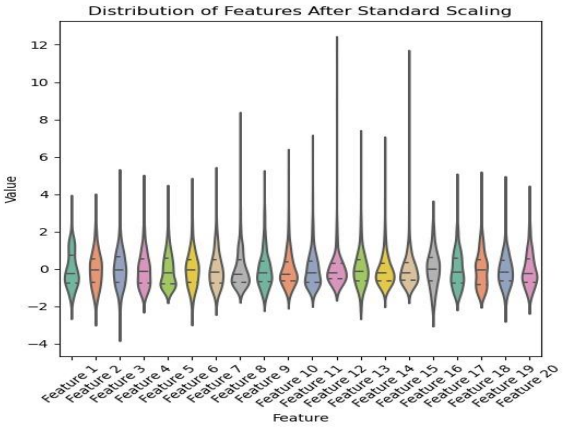

Fig.9. After scaling

Fig. 8 shows that the plot of features before standard scaling looks empty, it means that either all values in each feature are the same or there's very little variation in their values. This highlights the need for standard scaling to prepare the data for analysis.

Fig 9 shows that after applying standard scaling we can see there are clearer patterns. The plot now shows how the data is spread out after scaling, giving us a better understanding of its overall structure. This helps us see how scaling affects the data.

D. Training and Testing

Further we trained our data by applying five different machine learning algorithms – Logistic Regression, Decision Tree, Random Forest, KNN and SVM. We calculated both training and testing accuracies for each model. And we generated classification report with accuracy, recall and f1 score for training and testing data for every particular model.

E. Choosing the best classifier

In our breast cancer prediction research, the SVM model stood out with an impressive testing accuracy. This high accuracy underscores the SVM's effectiveness in correctly identifying breast cancer cases, suggesting its potential as a valuable tool in healthcare decision-making.

V. MACHINE LEARNING ALGORITHMS

In our study, we used machine learning to make predictions. The algorithms we used include:

Logistic regression is a statistical technique used for predicting outcomes with two possibilities, like yes/no or true/false. It is valuable for its interpretability, as you can understand how each input variable affects the likelihood of the outcome. It's widely applied across industries for tasks like risk assessment, marketing analytics, and medical diagnosis.

Decision tree is comparable to a flowchart in which each node stands for a question or decision based on input features. It's a popular tool in machine learning used for both classification and regression tasks. The tree splits the data into smaller groups based on features, aiming to create subsets that are as pure as possible, meaning they contain mostly one class or category.

Random forests are an ensemble learning method used for classification and regression tasks in machine learning. They are constructed from a multitude of decision trees during training and output the mode of the classes (classification) or mean prediction (regression) of the individual tree.

K-Nearest Neighbors (KNN) is a simple but powerful algorithm in machine learning that helps with sorting things into groups or predicting values**.** The idea behind KNN is simple: it classifies or predicts a data point based on the majority vote or average of its nearest neighbors in

the feature space. The parameter 'k' determines the number of neighbors considered for decision making.

Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression tasks. Its main objective is to find the best possible line or hyperplane that separates data points belonging to different classes with the widest margin. SVM achieves this by identifying support vectors, which are data points closest to the decision boundary. These support vectors play a crucial role in determining the optimal separating hyperplane.

VI. EXPERIMENT ENVIRONMENT

In our experiment environment for breast cancer prediction, we utilized Anaconda Jupyter Notebook, a widely-used interactive computing environment for data analysis and machine learning tasks. We implemented five different machine learning algorithms for breast cancer prediction: Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). To facilitate our experimentation, we leveraged several Python libraries, including: Scikit-learn (sklearn), Pandas, NumPy, Matplotlib and Seaborn etc. Additionally, to ensure robust evaluation of our models' performance and avoid overfitting, we employed k-fold cross-validation. We also utilized SMOTE and Standard Scalar to prevent data imbalance, by combining Anaconda Jupyter Notebook, essential Python libraries, we created a robust, efficient experiment environment for breast cancer prediction. This setup enabled us to explore the effectiveness of different machine learning algorithms and make informed decisions regarding model selection and optimization strategies.
Moreover, in our study, we ensured that the software runs smoothly by specifying certain hardware and software requirements. For hardware, we recommend having at least an Intel Dual Core processor with a clock speed of 2.0GHz or higher, a minimum of 1TB of storage space on your hard disk, and 8GB of RAM for optimal performance. On the software side, you'll need a modern web browser such as Chrome, Windows 7 Server or a later version for your operating system, and Python installed to use COLAB. These specifications were chosen to guarantee that the software functions effectively and is compatible with most systems.

VII. RESULTS

In our study, we conducted experiments to predict breast cancer outcomes using various machine learning algorithms. We employed Synthetic Minority Over-sampling Technique (SMOTE) and Standard Scalar to preprocess the training and testing data, enhancing data balance and standardizing feature scales for improved model performance.

We applied five different algorithms: Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest (RF), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Each algorithm was trained and tested on the preprocessed dataset to evaluate its predictive capabilities. Accuracy and precision metrics were calculated for every model to quantitatively assess their performance.

Accuracy measures the overall correctness of predictions, while precision indicates the proportion of true positive predictions among all positive predictions.
The fig. 10 below represents the comparison of accuracy precision and f1 score of each model. It clearly depicts that SVM has highest accuracy of and precision followed by Logistic Regression, Random forest , KNN and then decision tree.

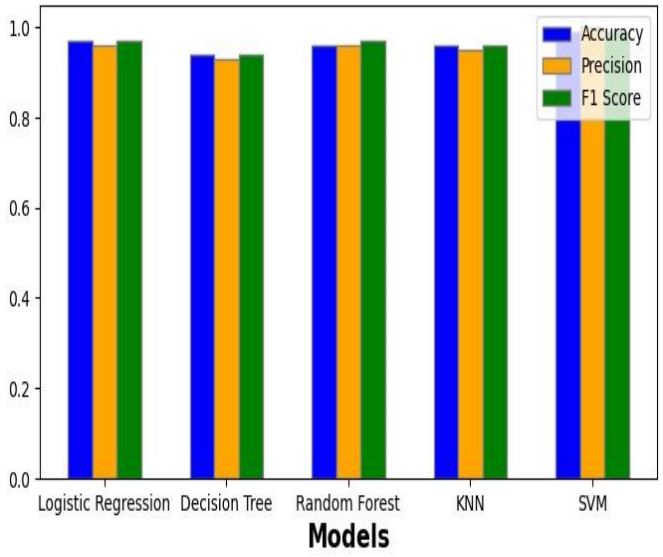

Fig.10. Comparison of metrics

VIII. DISCUSSION

Our experiments showed that the **Support Vector Machine (SVM) algorithm had the highest accuracy at 99%,** outperforming other models. SVM's success is due to its ability to distinguish between benign and malignant tumors by identifying complex decision boundaries[12]. Although SVM performed exceptionally well, other algorithms like Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest (RF), and K-Nearest Neighbors (KNN) also achieved good results, although their accuracy scores were slightly lower compared to SVM. We found that using preprocessing techniques like **SMOTE** and **Standard Scalar** helped balance the data and improve the performance of all algorithms. Overall, our research highlights the effectiveness of machine learning in predicting breast cancer outcomes with the highest accuracy being 99%.

In our study, we took several steps to refine our model and improve its accuracy. Initially, we evaluated various machine learning algorithms using our original dataset.

Then, we examined the correlation between features to identify and remove highly correlated ones, streamlining our dataset. After refining our dataset, we re-evaluated our models and noticed enhancements in accuracy. To address class imbalances, we applied techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) and standardized feature values using standard scaling. These techniques enhanced the reliability of our models.

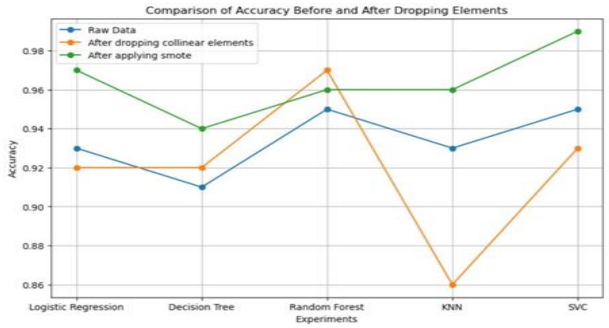

Fig.11. Comparison before and after SMOTE

To summarize our findings, we created a simple graph(Fig. 11) illustrating the evolution of accuracy throughout our analysis.

A. Area Under Curve (AUC)

Table 1 shows AUC metric calculated for each model

Table1. Area under Curve

Algorithms	AUC (%)
SVM	1.00
Random Forests	0.99
Logistic Regression	0.99
KNN	0.98
Decision Tree	0.97

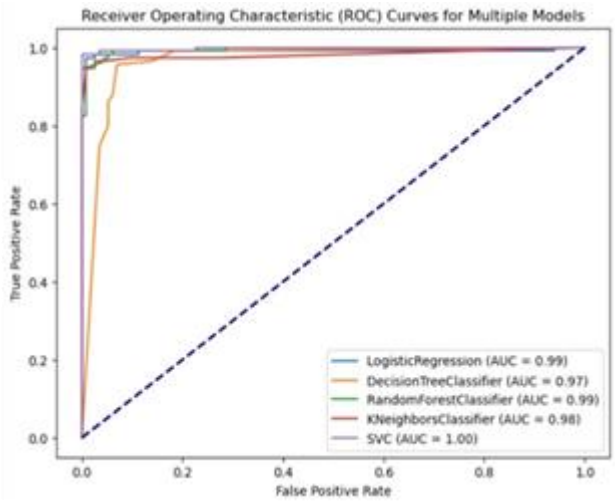

Fig.12. ROC Curve

In Fig. 12 the ROC curves of each machine learning algorithm are illustrated, providing insights into their classification performance. The Area Under the Curve (AUC) serves as a pivotal metric, where higher values denote superior classifier performance. SVM achieves the highest AUC score of 1.00%, indicating its robust discriminative ability. Conversely, the Decision Tree classifier demonstrates the lowest AUC score of 0.97%, as depicted in Table 1.

These findings emphasize the superiority of SVM in diagnosing breast cancer, as evidenced by its impressive predictive accuracy, precision, sensitivity, and AUC score.

B. Comparison of Accuracies

We've compared the testing accuracies of every model of our proposed system with the accuracies recorded in existing system in the Table2 below.

Table 2. Accuracy Comparison

Model	Existing Model Accuracy [20]	Proposed System Accuracy
SVM	97.2	99.2
Logistic Regression	95.8	97.3
Random Forest	96.5	96.5
KNN	93.7	96
Decision Tree	95.1	94.3

From table 2 we can see the spike of testing accuracies from existing model to our proposed system. SVM has the highest testing accuracy of 99.2 and decision tree stays at the bottom with 94.3. The improvement in accuracy is the result of data balancing techniques as well as eliminating the collinear features in our proposed system. Fig.13 shows the graph of comparisons.

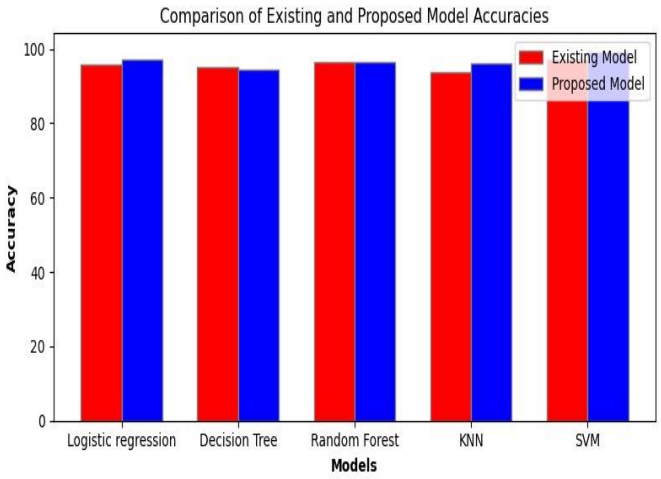

Fig.13.

C. Classifiers Performance

Precision, recall, and F-Measure percentages for SVM are notably higher (0.98%, 1.00%, and 0.99% respectively) compared to other classifiers, underscoring its efficacy in distinguishing between malignant and benign classes in the Breast Cancer Wisconsin Diagnostic dataset as depicted in Table 3.

This further solidifies SVM's position as the preferred choice among classifiers for effectively distinguishing between malignant(M)and benign(B) instances in the dataset.

Table 3. Classifiers performance

Algorithm	Precision	Recall	F-measure	Class
SVM	0.98	1.00	0.99	B
	1.00	0.98	0.99	M
Logistic Regression	0.98	0.97	0.97	B
	0.97	0.98	0.97	M
Random Forest	0.97	0.97	0.97	B
	0.97	0.97	0.97	M
KNN	0.96	0.96	0.96	B
	0.96	0.97	0.96	M
Decision Tree	0.96	0.93	0.94	B
	0.93	0.96	0.94	M

E. Confusion Matrix:

We created confusion matrices for each model using the testing data. These matrices help us see how well our models classify data by showing the number of correct and incorrect predictions. By analyzing these matrices, we can understand how accurate our models are in making predictions and identify areas for improvement. This process allows us to select the most effective model for our analysis and improve the overall reliability of our predictions. Here are the confusion matrices generated for every model.

1. Logistic Regression:

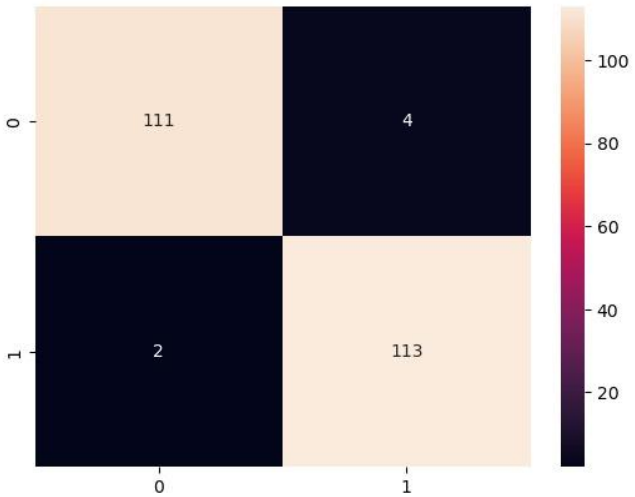

Fig.14. Logistic Regression Confusion Matrix

	precision	recall	f1-score	support
0	0.98	0.97	0.97	115
1	0.97	0.98	0.97	115
accuracy			0.97	230
macro avg	0.97	0.97	0.97	230
weighted avg	0.97	0.97	0.97	230

Fig.15.

Fig, 14 shows that the Logistic Regression model accurately identified 111 positive cases and 113 negative cases. However, it incorrectly classified 4 negative cases as positive and missed 2 positive cases and Fig. 15 shows model's performance through the report using various metrics.

2. Decision Tree

Fig. 16 shows that the Decision Tree model accurately identified 107 positive cases and 110 negative cases. However, it incorrectly classified 8 negative cases as positive and missed 5 positive cases. Fig. 17 shows model's performance through the report using various metrics.

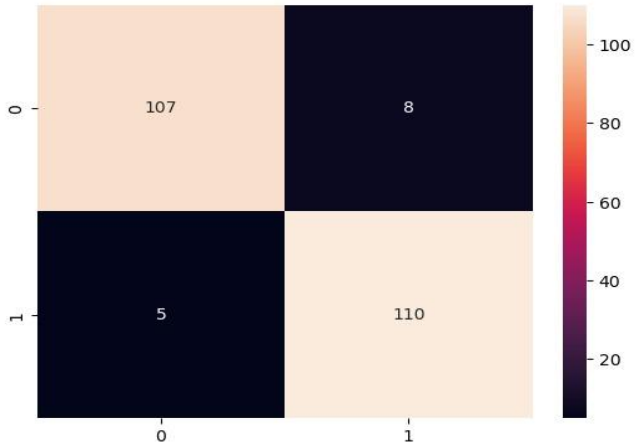

Fig. 16. Decision Tree Confusion Matrix

	precision	recall	f1-score	support
0	0.96	0.93	0.94	115
1	0.93	0.96	0.94	115
accuracy			0.94	230
macro avg	0.94	0.94	0.94	230
weighted avg	0.94	0.94	0.94	230

Fig.17.

3. Random Forest:

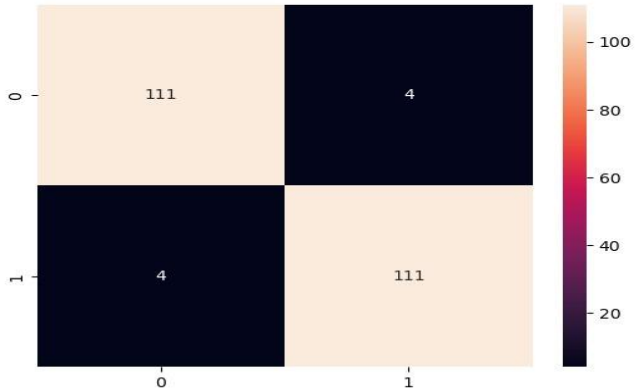

Fig. 18. Random Forest Confusion Matrix

	precision	recall	f1-score	support
0	0.97	0.97	0.97	115
1	0.97	0.97	0.97	115
accuracy			0.97	230
macro avg	0.97	0.97	0.97	230
weighted avg	0.97	0.97	0.97	230

Fig. 19.

Fig.18 shows that the Random Forest model accurately identified 111 positive cases and 111 negative cases. However, it incorrectly classified 4 negative cases as positive and missed 4 positive cases. Fig. 19 shows model's performance through the report using various metrics.

4. KNN:

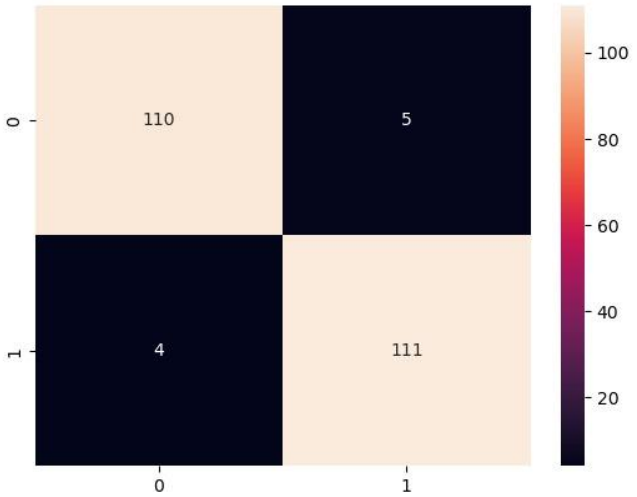

Fig. 20. KNN Confusion Matrix

Fig.20 shows that the KNN model accurately identified 110 positive cases and 111 negative cases. However, it incorrectly classified 5 negative cases as positive and missed 4 positive cases. Fig. 21 shows model's performance through the report using various metrics.

	precision	recall	f1-score	support
0	0.96	0.96	0.96	115
1	0.96	0.97	0.96	115
accuracy			0.96	230
macro avg	0.96	0.96	0.96	230
weighted avg	0.96	0.96	0.96	230

Fig. 21

5. SVM:

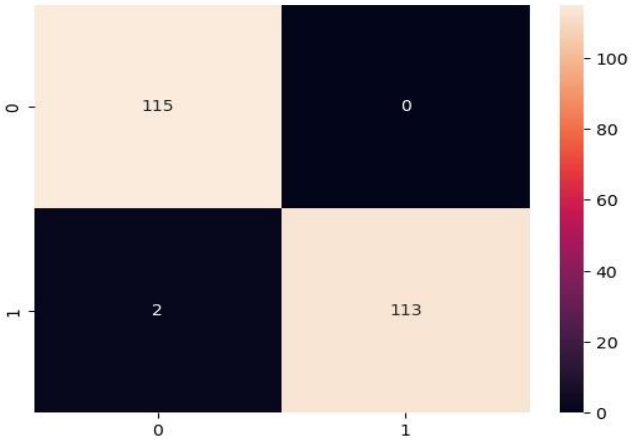

Fig. 22. SVM Confusion Matrix

	precision	recall	f1-score	support
0	0.98	1.00	0.99	115
1	1.00	0.98	0.99	115
accuracy			0.99	230
macro avg	0.99	0.99	0.99	230
weighted avg	0.99	0.99	0.99	230

Fig. 23.

Fig.22 shows that the SVM model accurately identified 115 positive cases and 113 negative cases. However, it incorrectly classified 0 negative cases as positive and missed 2 positive cases. Fig. 23 shows model's performance through the report using various metrics.

IX. CONCLUSION

We examined the WBCD dataset using five different ML algorithms: SVM, Random Forests, Logistic Regression, Decision Tree, and K-NN. Our analysis, conducted in Python with the scikit-learn library in the Anaconda environment, aimed to identify the most accurate and precise algorithm for breast cancer prediction. After thorough evaluation using metrics like confusion matrix, accuracy, sensitivity, precision, and AUC, we determined that the SVM algorithm achieved the best performance, SVM surpassed all other algorithms as it had the accuracy of 99.2% and the precision of 100% and AUC of 100% . These results highlight the SVM's effectiveness in predicting and diagnosing breast cancer, showcasing its accuracy and precision. However, our study is limited to the WBCD dataset, emphasizing the need for future research to validate these findings across various datasets

X. REFERENCES

[1] WHO, "Breast cancer," WHO, http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020)

[2] ICMR, "CONSENSUS DOCUMENT FOR MANAGEMENT OF BREAST CANCER," ICMR, https://main.icmr.nic.in/sites/default/files/guidelines/Breast_Cancer.pdf. (accessed 2016)

[3] American Cancer Society, "Breast Cancer Statistics, 2022," American Cancer Society, https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21754 (accessed 2022)

[4] Smith KR, Hanson HA, Mineau GP and Buys SS, "Effects of BRCA1 and BRCA2 mutations on female fertility," Proc Biol Sci, vol. 279, no. 1732, pp. 1389-95, Apr. 7, 2012. doi: 10.1098/rspb.2011.1697.

[5] Satpati S, Gaurkar SS, Potdukhe A, and Wanjari MB. "Unveiling the Role of Hormonal Imbalance in Breast Cancer Development: A Comprehensive Review." Cureus. July 11, 2023; 15(7): e41737. doi: 10.7759/cureus.41737.

[6] Ronckers CM, Erdmann CA and Land CE. "Radiation and breast cancer: a review of current evidence." Breast Cancer Research, vol. 7, no. 1, pp. 21-32, 2005. doi: 10.1186/bcr970. Epub 2004 Nov 23.

[7] Eriksson M, Czene K, Strand F, Zackrisson S, Lindholm P, Lång K, Förnvik D, Sartor H, Mavaddat N, Easton D and Hall P. "Identification of Women at High Risk of Breast Cancer Who Need Supplemental Screening." Radiology, vol. 297, no. 2, pp. 327-333, November 2020. doi: 10.1148/radiol.2020201620. Epub 2020 Sep 8. PMID: 32897160.

[8] Sparano JA, "Clinical and Genomic Risk to Guide the Use of Adjuvant Therapy for Breast Cancer," N Engl J Med, vol. 380, no. 25, pp. 2395-2405, Jun. 20, 2019. doi: 10.1056/NEJMoa1904819. Epub 2019 Jun 3. PMID: 31157962; PMCID: PMC6709671.

[9] Noblejas-López MdM, López-Cade I, Fuentes-Antrás J, Fernández-Hinojal G, Esteban-Sánchez A, Manzano A, García-Sáenz JÁ, Pérez-Segura P, la Hoya Md, Pandiella and Győrffy, Bl. "Genomic Mapping of Splicing-Related Genes Identify Amplifications in LSM1, CLNS1A, and ILF2 in Luminal Breast Cancer." Cancers, vol. 13, no. 16, p. 4118, 2021. doi: 10.3390/cancers13164118.

[10] Zhang H, "Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses," Nat Genet, vol. 52, no. 6, pp. 572-581, Jun. 2020. doi: 10.1038/s41588-020-0609-2. Epub 2020 May 18. PMID: 32424353; PMCID: PMC7808397.

[11] Assiri AS, Nazir S and Velastin SA, "Breast Tumor Classification Using an Ensemble Machine Learning Method," J Imaging, vol. 6, no. 6, p. 39, May 29, 2020. doi: 10.3390/jimaging6060039. PMID: 34460585; PMCID: PMC8321060.

[12] Khourdifi Y and Bahaj M. "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification." 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, Morocco, 2018, pp. 1-5. doi: 10.1109/ICECOCS.2018.8610632.

[13] Behravan H, Hartikainen JM, Tengström M, Kosma VM and Mannermaa A, "Predicting breast cancer risk using interacting genetic and demographic factors and machine learning," Sci Rep, vol. 10, no. 1, p. 11044, 2020. doi: 10.1038/s41598-020-66907-9.

[14] Deng Z, Su B and Zhang K, "Breast cancer classification based on ensemble learning," China Medical Devices, vol. 35, no. 12, 2020.

[15] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y and Xu W, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," Cancer Genomics Proteomics, vol. 15, no. 1, pp. 41-51, Jan-Feb 2018. doi: 10.21873/cgp.20063. PMID: 29275361; PMCID: PMC5822181.

[16] Kulkarni A, Chong D and Batarseh FA. "Foundations of data imbalance and solutions for a data democracy."

[17] Hand DJ and Anagnostopoulos C. "When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance." Pattern Recognition Letters, vol. 34, no. 5, pp. 492-495, 2013. ISSN 0167-8655. DOI: 10.1016/j.patrec.2012.12.004.

[18] Hall M, "Correlation-Based Feature Selection for Machine Learning," Department of Computer Science, 2000.

[19] Dataset Link
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[20] Naji, Mohammed Amine, et al. "Machine learning algorithms for breast cancer prediction and diagnosis." Procedia Computer Science 191 (2021): 487-492.