IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Breast Cancer Prediction using Machine Learningand Deep Learning Algorithms

Riya Agrawal
Student in Information Technology
Thakur college of Engineerring and Technology
Mumbai,India

Riya Gour
Student in Information Technology
Thakur college of Engineerring and Technology
Mumbai,India

Aanchal Pandey
Student in Information Technology
Thakur college of Engineerring and Technology
Mumbai,India

Mrs. Neha Patwari
Assistant Professor
Department of Information Technology,
Thakur College of Engineering and Technology

***Abstract*— Breast cancer is a significant public health concern, necessitating advanced techniques for early detection and prediction. This research paper investigates the application of machine learning algorithms, specifically Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Residual Networks (ResNet), for breast cancer prediction using a breast cancer dataset in CSV format. The dataset comprises clinical and diagnostic features, making it suitable for both traditional and deep learning models. SVM, a powerful classification algorithm, is employed to analyze the tabular data. The study assesses the algorithms' performance based on accuracy, sensitivity, specificity, and confusion matrix to determine their predictive capabilities. The findings reveal the comparative strengths and weaknesses of SVM, CNN, and ResNet in breast cancer prediction using this CSV dataset. This research contributes to enhancing the accuracy of breast cancer prediction models.**

Keywords: Breast cancer prediction, Machine learning, Support Vector Machines, Convolutional Neural Networks, Residual Networks, Comparative analysis, CSV dataset, Early detection, Deep learning.

I. INTRODUCTION

Breast cancer is still a major public health concern, which highlights the urgent need for reliable and effective diagnostic techniques. Although they work well, traditional diagnostic techniques sometimes have problems with accuracy and resource use. Breast cancer detection is transformed by the use of machine learning and deep learning algorithms like CNN, ResNet, and SVM. SVM is a flexible and easily interpreted approach that works well for feature-based classification, while CNN and ResNet are especially well-suited for image-based diagnostics since they use deep neural networks to automatically extract intricate patterns and characteristics from medical pictures. Our goal is to improve breast cancer early detection by utilising these algorithms, which will allow physicians to make more accurate and well-informed decisions. In order to lay a strong foundation for future developments in the field of accurate and timely cancer diagnosis, this research aims to illuminate the revolutionary potential of deep learning and machine learning in the context of breast cancer diagnostics.

The need for more precise and effective techniques for breast cancer diagnosis is what spurs this research. Conventional techniques like mammography, ultrasound, and biopsies have been extremely helpful in clinical settings. They do have certain drawbacks, though, such as inter-observer variability and the possibility of false positives and false negatives. These constraints make it necessary to investigate novel strategies for enhancing the diagnostic procedure. The project makes use of a breast cancer dataset that was gathered from Kaggle. Each patient is given a unique ID number, which makes traceability easier. The mean, standard error, and worst values should be recorded for attributes such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These characteristics are crucial for characterising cell nuclei in biopsies of breast cancer.

The primary objective of this research is to develop and evaluate predictive models based on CNN, SVM and ResNet for the classification of breast tumors as malignant or benign. We seek to harness the power of these algorithms in analyzing medical images and clinical data to provide accurate and timely breast cancer predictions. Many techniques have been offered to determine a precise diagnosis of breast cancer. Machine learning can be applied to the dataset for prediction with ease because it includes a range of unique report attributes. even when employing technology that isn't entirely automated to produce the desired results. As a result, we suggest here that breast cancer be classified and predicted entirely automatically using a dataset. applying the17 deep learning method. It is acknowledged that this learning strategy is the most effective way to categorise and predict image datasets. The results section provides a detailed analysis of the models performance, and the discussion section interprets the findings and explores their clinical implications.

II. RELATED WORK

[1] In this study by Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika et al Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. The Breast Cancer Wisconsin Diagnostic dataset was subjected to five machine learning algorithms in this study: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5), and K-Nearest Neighbours (KNN).According to the results, decision trees had the lowest accuracy, while Support Vector Machines (SVM) outperformed all other classifiers with an accuracy of 97.2%. One important observation regarding the study's limitations is that these results are exclusive to the WBCD database. In order to verify the applicability of these results, additional testing on various databases is advised.

[2] Breast Cancer Prediction Using Machine Learning The paper compares models including K Nearest Neighbor, Support Vector Machine, Logistic Regression, and Gaussian Naive Bayes for breast cancer prediction .K Nearest Neighbours appears to be a useful tool for predicting breast cancer. The study's limited scope and repeated use of the dataset, however, limit the study's conclusions used machine learning to predict breast cancer in this study. An accuracy of roughly 92% was obtained by combining 13 features from PCA with a Multi-Level Wavelet Conversion strategy. Selected features were employed by the system for image segmentation using Fuzzy-C-means (FCM) clustering. The prediction scope can be expanded by adding more features to improve accuracy and reliability.

[3] In this 2022 study by Sk. Ahmed Mohiddin and a team of students, the WBCD dataset was employed for breast cancer prediction. Six different machine learning algorithms, namely Decision Trees, Random Forests, Naive Bayes, SVC, KNN, and Logistic Regression, were utilized, and the SMOTE technique was applied to address class imbalance. Notably, the Random Forest algorithm, in combination with K-Fold cross-validation, achieved the highest accuracy, nearing a remarkable 100.00%A larger dataset would provide a more robust evaluation of the ResNet model's performance. The model's accuracy might not be entirely representative due to the small dataset size, which can lead to overfitting and reduced generalisation, impacting the accuracy estimate.

[4] An accuracy of roughly 92% was obtained by combining 13 features from PCA with a Multi-Level Wavelet Conversion strategy. Selected features were employed by the system for image segmentation using Fuzzy-C-means (FCM) clustering. The prediction scope can be expanded by adding more features to improve accuracy and reliability.

[5] In this 2022 study by Sk. Ahmed Mohiddin and a team of students, the WBCD dataset was employed for breast cancer prediction. Six different machine learning algorithms, namely Decision Trees, Random Forests, Naive Bayes, SVC, KNN, and Logistic Regression, were utilized, and the SMOTE technique was applied to address class imbalance. Notably, the Random Forest algorithm, in combination with K-Fold cross-validation, achieved the highest accuracy, nearing a remarkable 100.00%A larger dataset would provide a more robust evaluation of the ResNet model's performance. The model's accuracy might not be entirely representative due to the small dataset size, which can lead to overfitting and reduced generalisation, impacting the accuracy estimate.

[6] In a 2021 study by Apoorva V, Yogish H K, and Chayadevi M L, the research involves three phases: data generation, analysis, and prediction. They employ Convolutional Neural Networks (CNN) for image data and K-Nearest Neighbour (KNN), Decision Tree (CART), Support Vector Machine (SVM), and Naïve Bayes for numerical data derived from digitized breast mass images. The study finds that SVM achieves high prediction accuracy for breast cancer. They also develop a user-friendly interface using the Flask framework for predictions.

III. PROPOSED METHODOLOGY

Breast cancer is a major worldwide health issue that requires prompt and precise detection methods for better patient outcomes and early intervention. Using a widely available breast cancer dataset in CSV format, this research study intends to address the challenge of breast cancer prediction by employing machine learning techniques, notably Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Residual Networks (ResNet).

The main challenges are twofold: first, creating reliable predictive models that can recognise breast cancer from clinical and diagnostic data; and second, comparing these models to ascertain which is more predictive. With a special focus on the possible benefits of deep learning techniques over more conventional machine learning approaches, the research aims to provide light on the application of machine learning algorithms in the context of breast cancer diagnosis. The suggested technique consists of a methodical process that includes data pretreatment and collecting, data splitting, feature engineering, model selection, rigorous model training with tuned hyperparameters, performance assessment, and a thorough analysis of the outcomes. By using this methodology, the study hopes to improve early breast cancer detection techniques, which could have a significant positive impact on clinical practise and encourage more research in this important area.

IV. DESIGN & METHODOLOGY

A. Support Vector Machine (SVM)

A key element of our approach in our effort on early breast cancer diagnosis was the use of Support Vector Machines (SVM). In order to create a predictive model based on feature extraction from medical pictures and patient data, SVM was used as a machine learning technique. To train the SVM model to classify patients as either having cancer or not, we preprocessed and engineered pertinent elements from the dataset, such as texture, shape, and density characteristics of breast cancer images. We were able to create a reliable and effective diagnostic tool for the early identification of breast cancer by fine-tuning hyperparameters and carrying out a significant amount of cross-validation. Our study gained a valuable dimension from SVM's powerful classification skills and interpretability, which also increased the potential influence on breast cancer diagnosis and patient care.

B. Convolutional Neural Networks (CNN)

Convolutional neural networks (CNNs) are a key component of our project's breast cancer early detection system, and we made use of their capabilities. In order to enable the model to automatically learn complicated and discriminative features relevant to breast cancer detection, CNNs were used to directly analyse and process our dataset. To maximise the CNN model's performance, we trained it for ten epochs. We created and refined a deep CNN architecture that uses several convolutional and pooling layers to reduce dimensionality while capturing textures and patterns in images. In order to take advantage of their ability to extract features from images, strategies for transfer learning were also investigated. Using our breast cancer dataset, the CNN model was optimised to distinguish between benign and malignant instances, improving the sensitivity and accuracy of early breast cancer diagnosis.

C. Residual Networks (ResNet)

In order to improve breast cancer early detection, we used the deep convolutional neural network Residual Network (ResNet) architecture in our study. We were able to build a much deeper network thanks to ResNet's novel structural innovation of residual connections, which facilitated the extraction of complex and useful characteristics from medical pictures. To guarantee efficient learning and model convergence in our ResNet implementation, we trained the model for ten epochs. We successfully decreased the likelihood of overfitting by integrating ResNet, which improved the model's ability to learn and identify subtle patterns linked to breast cancer. Our breast cancer dataset was used to fine-tune the ResNet model, which had been pre-trained on large-scale picture datasets. This allowed the network to adapt its learnt features to the particular characteristics. Our project's use of ResNet greatly improved the model's resilience and accuracy, enabling us to create a more sensitive and dependable instrument for the early diagnosis of breast cancer.

D. Methodology Steps

Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Residual Network (ResNet) are the three main machine learning and deep learning techniques that we will be using in our study to detect breast cancer. In order to do this, we used a widely available and carefully curated CSV dataset that includes a variety of attributes that were taken out of photos of breast cancer. Our study's main objective is to develop predictive models with high accuracy and sensitivity that can help with breast cancer early diagnosis.

The project's methodology involves several key steps.

i. We preprocess the dataset using feature engineering, normalisation, and data cleaning. In order to give instructive inputs for the machine learning and deep learning models, we carefully select the features to extract pertinent information, such as texture, shape, and density, from the medical images.
ii. Then, we split the dataset into training, validation, and testing sets to ensure robust model training and evaluation.
iii. For the SVM algorithm, we employ feature-based classification, training the model on the engineered features.
iv. We build image classification models for CNN and ResNet to process medical images directly. By automatically deriving complex patterns and features from the photos, these deep learning models will increase the sensitivity and specificity of breast cancer detection.
v. Finally, we do thorough cross-validation and hyperparameter tuning to guarantee the best possible performance of our models, and we analyse and compare the accuracy, precision, recall, F1-score, and confusion matrix performance of the SVM, CNN, and ResNet models.

V. SYSTEM DESIGN

We created a thorough workflow in our project's system design that combines feature extraction, sophisticated machine learning, and deep learning algorithms with data pretreatment. The system begins with the collection and preprocessing of data, during which the dataset is meticulously cleaned and standardised. Feature engineering is essential because it allows us to expand the dataset by extracting a wide variety of informative features from medical images, including texture, shape, and density. Then, to improve the sensitivity and specificity of early breast cancer diagnosis, we used models that are well-suited for classification tasks: Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Residual Network (ResNet). The models are optimised by rigorous cross-validation and hyperparameter adjustment. The output of our system offers researchers and medical practitioners a dependable and automated instrument that advances oncology by aiding in the early detection of breast cancer.

A. Data Flow Diagram

The following data flow diagrams depict how we designed our system:

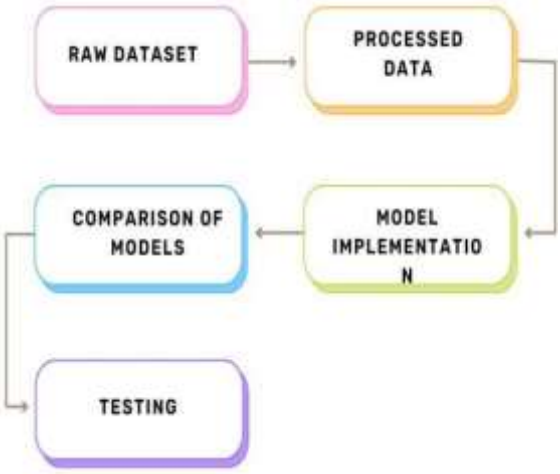

Figure 1.1: Data Flow Diagram – Level 0

Figure 1.2: Data Flow Diagram – Level 1

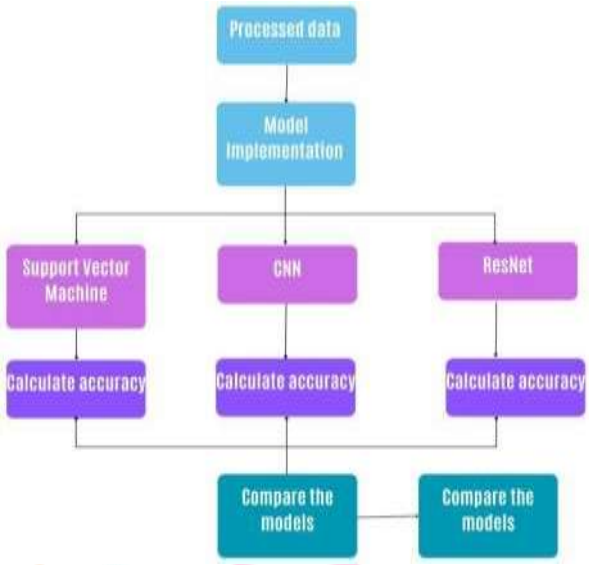

Figure 1.3: Data Flow Diagram – Level 2

VI. RESULT AND DISCUSSION

We give a thorough examination of our results. Our research indicates that Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Residual Network (ResNet) are effective machine learning and deep learning algorithms for breast cancer identification. We present encouraging results using our breast cancer dataset, where our models achieve good recall, accuracy, precision, F1-score, and confusion matrix. These results imply that our method has a lot of promise for helping medical practitioners diagnose breast cancer early. The precision, recall, f1-score and support of each algorithm is shown below:

	precision	recall	f1-score	support
0	0.95	0.99	0.97	71
1	0.97	0.91	0.94	43
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

Figure 2.1: Classification Report of SVM

	precision	recall	f1-score	support
0	0.97	1.00	0.99	71
1	1.00	0.95	0.98	43
accuracy			0.98	114
macro avg	0.99	0.98	0.98	114
weighted avg	0.98	0.98	0.98	114

Figure 2.2: Classification Report of CNN

	precision	recall	f1-score	support
Benign	0.91	1.00	0.95	71
Malignant	1.00	0.84	0.91	43
accuracy			0.94	114
macro avg	0.96	0.92	0.93	114
weighted avg	0.94	0.94	0.94	114

Figure 2.3: Classification Report of ResNet

i. The precision of positive class predictions is determined by dividing the total number of positive predictions by the ratio of actual positive predictions. The best precision of the weighted average amongst all three algorithms is of CNN.
ii. Recall, which is a measure of how well a model detects positive cases, is defined as the ratio of true positive predictions to the total number of real positive occurrences. It is sometimes referred to as sensitivity or true positive rate. The algorithm having the best recall weighted average is CNN.
iii. When working with unbalanced datasets, the F1-Score—which is the harmonic mean of accuracy and recall—provides a good balance between the two metrics. Most apt F1-score of weighted average is of CNN algorithm.
iv. In the context of the data distribution, support refers to the number of actual occurrences of the class in the dataset and aids in interpreting the relevance of accuracy, recall, and F1-score. The support of all three algorithms was constant.

We visualized our algorithms using a confusion matrix to assess their performance in classifying benign and malignant cases, providing a clear representation of their predictive accuracy and misclassifications. In confusion matrix there are 4 instances:

i. True Positive (TP) - True Positive in a confusion matrix represents the number of actual positive instances correctly predicted as positive by a classification model.
ii. True Negative (TN) - True Negative represents the number of actual negative cases correctly classified as negative by a model.
iii. False Positive (FP) - False Positive in a confusion matrix represents the instances where the model incorrectly predicted a positive class when the actual class was negative.
iv. False Negative (FN) - False Negative in a confusion matrix represents cases where the model predicted the condition as negative when it was actually positive, indicating a missed detection.

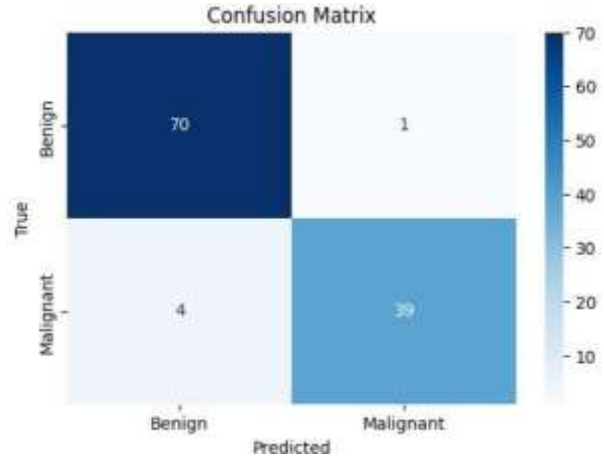

Fig 2.4: SVM Confusion Matrix

From the provided confusion matrix for the Support Vector Machine (SVM) classification model trained on a breast cancer dataset, several key results can be gathered. The confusion matrix visually displays the number of true positives, true negatives, false positives, and false negatives. TP represents the number of cases correctly diagnosed as Benign. FP represents number of cases misclassified reiterating that originally the cases belonged to class Benign but were classified as Malignant. FN is the reverse of FP. TN represents the number of cases that are correctly diagnosed as Malignant.

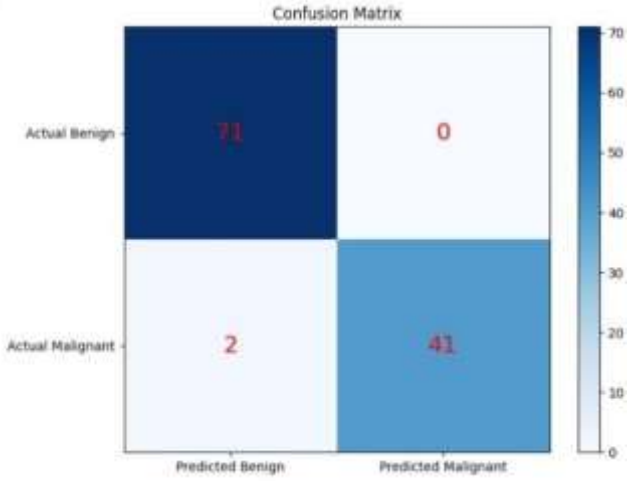

Fig 2.5: CNN Confusion Matrix

The number of true positives, true negatives, false positives, and false negatives is shown graphically in CNN's confusion matrix. The number of cases appropriately classified as benign is represented by TP. The number of misclassified cases (FP) indicates how many cases initially belonged to the class Benign but were incorrectly classified as Malignant. FP is reversed by FN. The number of cases that are accurately identified as malignant is represented by TN.

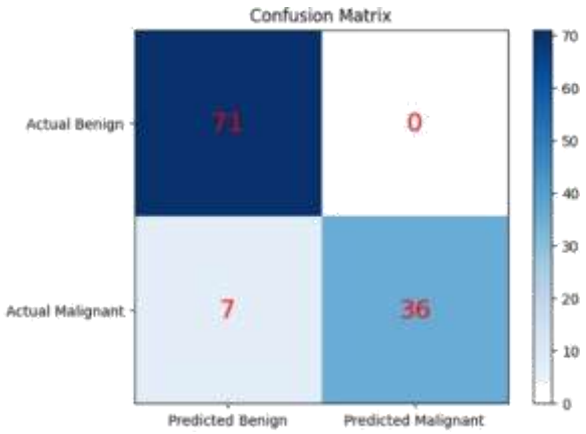

Fig 2.6: ResNet Confusion Matrix

The number of true positives, true negatives, false positives, and false negatives is shown graphically in the confusion matrix of ResNet. The number of cases appropriately classified as benign is represented by TP. The number of misclassified cases (FP) indicates how many cases initially belonged to the class Benign but were incorrectly classified as Malignant. FP is reversed by FN. The number of cases that are accurately identified as malignant is represented by TN.

After the implementation of SVM, CNN and ResNet on our dataset, the accuracies achieved were 95%, 98% and 94% respectively. SVM has demonstrated its effectiveness in binary classification tasks. It is particularly notable for its simplicity and interpretability. In situations where data availability is constrained, virtual machines's capacity to handle high-dimensional data and perform well with relatively small datasets can be advantageous. ResNet, an extension of CNN, further improves the performance of deep learning models by addressing the vanishing gradient problem. ResNet's skip connections enable it to train very deep neural networks effectively, which can be crucial for tasks like breast cancer prediction. The size of the particular dataset, available processing power, and the intended trade-off between model complexity and performance all play a role in which CNN or ResNet is selected.

There are a number of reasons for the differences in accuracy between the three models (SVM, CNN, and ResNet). First, performance is greatly influenced by the model design and algorithm selection. While CNN and ResNet are deep learning models that are well-known for their capacity to extract complex features from data—a skill that can be especially useful for image-based applications like breast cancer detection—SVM is a standard machine learning model. Second, the performance of the model may be impacted by the dataset's size and complexity. Rich feature information in image datasets can benefit from the high-dimensional data processing capabilities of deep learning models like CNN and ResNet. However, due to its robustness, SVM may perform better than deep learning models if the dataset is limited or contains noisy data.

Model performance can also be impacted by the quantity and caliber of the dataset as well as the efficiency of data preprocessing. SVM can function rather well with fewer datasets, however deep learning models frequently require larger datasets to generalize well. Additionally, accuracy can be greatly impacted by fine-tuning hyperparameters like learning rates, the number of layers or kernels in deep learning models, and the kernel function selection in SVM. Variations in these hyperparameters may result in different performance outcomes. Finally, the choice of model may also be influenced by the degree of interpretability needed for the application. While CNN and ResNet have strong feature extraction capabilities but are sometimes regarded as "black-box" models, SVM produces more interpretable results. The choice of the optimal model for a specific job may take into account the trade-off between interpretability and accuracy.

VII. CONCLUSION

The above study has investigated the use of machine learning algorithms in the early detection of breast cancer, highlighting the critical role these algorithms play in enhancing the precision and efficacy of diagnosis. Our research compared the effectiveness of several machine learning models, demonstrating how they can perform better than conventional techniques in terms of sensitivity, specificity, and total accuracy.

Our results emphasise the significance of model architecture and algorithm selection for breast cancer identification. Furthermore, the differences in performance across different models highlight how important it is to select the best method based on the features, size, and quality of the dataset.

To sum up, our research has shown how useful it is to use a variety of deep learning and machine learning techniques, such as Residual Network (ResNet), CNN, and Support Vector Machine (SVM), for the identification of breast cancer. All of these methods have demonstrated encouraging outcomes, and each has its own advantages. SVM can be a dependable option when working with smaller datasets or when decision-making transparency is essential because to its interpretability and strong performance. However, CNN and ResNet show great promise for image-based breast cancer diagnosis tasks due to their capacity to automatically extract complex information from medical pictures.

VIII. FUTURE WORK

i. Advanced Imaging Techniques: Improving the diagnosis of breast cancer requires staying up to date on the most recent developments in medical imaging. Advances in MRI methods, whole-breast ultrasonography, and 3D mammography have made it possible to image breast tissues with greater precision and detail. With the use of these technologies, radiologists and oncologists can see breast abnormalities more comprehensively, which helps with early diagnosis and precise characterization of possible cancers.

ii. Multi-Modal Data Fusion: A comprehensive and individualised approach to breast cancer risk assessment is made possible by the integration of several data sources, such as genetic information, electronic health records (EHR), patient-reported data, and data from wearable devices. By improving the accuracy of breast cancer risk prediction, this multimodal method makes early identification and customized therapies possible.

iii. Real-Time Predictive Analytics: In clinical settings, real-time prediction of breast cancer risk is

revolutionary. When determining a patient's risk of breast cancer, it gives medical professionals access to instant decision support. Clinicians can promptly offer further screening, preventive care, or therapies by utilising predictive analytics.

IX. REFERENCES

[1] (ICIIC 2021),Breast Cancer Prediction Using Machine Learning Techniques Apoorva V1 , Yogish H K, Chayadevi M L

[2] APRIL 22nd, 2021, BREAST CANCER PREDICTION USING MACHINE LEARNING AUTHOR: SANJANA BALASUBRAMANIAN

[3] Vol-09 Issue-01 No. 01 : 2022, BREAST CANCER PREDICTION USING MACHINE LEARNING Sk. Ahmed Mohiddin , T. Pooja Sri , K. Sharmila , P. Bhavya

[4] Volume 191, 2021 Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis Mohammed Amine Naji , Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar et al.

[5] Kumar Sanjeev Priyanka 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012071 A Review Paper on Breast Cancer Detection Using Deep Learning

[6] Volume 218, 2023 Machine Learning Techniques for Breast Cancer Prediction
Author links open overlay panel by Varsha Nemade, Vishal Fegade

[7] Ray, Sunil. “Learn How to Use Support Vector Machines (SVM) for Data Science.” Analytics Vidhya, 27 Apr. 2023

[8] Goel, Aparna. “Convolutional Neural Network (CNN) in Machine Learning.” GeeksforGeeks, GeeksforGeeks, 3 Feb. 2023,

[9] Pawan. “Residual Networks (Resnet) - Deep Learning.” GeeksforGeeks, GeeksforGeeks, 10 Jan. 2023

[10] Gupta, Akshay. “Steps to Complete a Machine Learning Project.” Analytics Vidhya, 16 Apr. 2021