Machine Learning Model for Risk of Breast Cancer Relapse

Vedant Padole
[1]Student
[1]Computer Science and Engineering (sp. AIML),
[1]Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Abstract - A malignant tumor that arises in the breast cells is called breast cancer. It affects both men and women and is the second most common cancer worldwide. Although millions of new cases are recorded each year, the annual mortality toll from breast cancer varies. Survival rates have increased due to early discovery through routine screenings and treatment improvements, but the impact is still substantial, underscoring the significance of support, education, and awareness campaigns for those who are impacted. Every individual with a breast cancer diagnosis carries a certain amount of risk in the event of a recurrence, albeit this risk varies. According to research, up to 50% of patients with inflammatory breast cancer and roughly 40% of those with early-stage triple-negative breast cancer is expected to experience a recurrence.

Early identification of breast cancer, risk assessment, individualized treatment planning, prognosis prediction, and medication discovery are all aided by machine learning algorithms. These technologies employ data to improve clinical decision-making; but, in order to assure efficacy and ethical use, their deployment necessitates rigorous validation and collaboration between data scientists and healthcare professionals.

Support Vector Machines (SVM) efficiently identify data points, which helps with jobs connected to breast cancer. They are excellent at identifying intricate, non-linear patterns in tabular data, which can be used to classify situations or forecast outcomes. SVM is useful for classification problems in breast cancer research because of its capacity to handle high-dimensional spaces and identify appropriate decision boundaries. By identifying intricate patterns in tabular data, Artificial Neural Networks (ANNs) prove to be efficacious in tasks pertaining to breast cancer. They are able to categorize cases, forecast results, and identify complex relationships. ANNs are flexible and can provide insights from structured information, which advances the study and treatment of breast cancer. This report will study and examine the best suitable machine learning algorithm of the three and give a comparative- results of all with respect to principal component analysis using random forest algorithm.

Index Terms – Artificial Intelligence, Artificial Neural Network, Support Vector Machines, Breast Cancer, Logistic Regression.

I. INTRODUCTION (HEADING 1)

The World Health Organization (WHO) estimates that 685,000 women worldwide lost their lives to breast cancer in 2020. Approximately 42,000 women and 500 men in the US pass away from breast cancer every year. If there is a reoccurrence of breast cancer then the survival rate if very low. The breast cancer be further classified into localized, regional, distant. The relative 5-year survival rates for each form of recurrent breast cancer are listed by the American Cancer Society as follows: Localized: 99% Regional: 86% Distant: 30%. This study aims to build a machine learning model which can further help to predict the risk of breast cancer using machine learning algorithms and also build a web application so that the people can use it at their ease.

Support Vector Machines (SVM) are useful for complex datasets because they perform well in high-dimensional environments and show resilience to outliers. Artificial Neural Networks (ANN) are a flexible and excellent tool for automatically learning features and capturing non-linear correlations. For simple jobs, Linear Regression is an interpretable and computationally efficient solution. The prediction performance can be improved by the powerful ensemble learning capabilities of gradient boosting models (e.g., XGBoost, LightGBM). Table data is handled effectively by the MLP Regressor. The decision is based on variables such as the associations' nature, interpretability, and dataset complexity.

Working with specialists and experimenting with various models is essential to achieve the best possible outcomes when performing duties linked to breast cancer. So the objectives for the following paper are:

1) Utilizing data from Electronic Health Records (EHRs), implement and compare different machine learning (ML) models, such as SVM, ANN, Linear Regression, to evaluate how well they predict the recurrence of breast cancer.
2) Examine the effects of merging structured and unstructured EHR data to improve the prediction of breast cancer recurrence. Use the features obtained from both sources to assess the models' performance.
3) Optimize performance by fine-tuning model topologies, feature engineering strategies, and hyperparameters. Evaluate and contrast each model's, F1-score, recall, precision, and prediction accuracy.
4) Provide tools that help healthcare providers track patients after cancer and design more efficient follow-up plans by offering insights into patient risk classification.

II. LITERATURE SURVEY

González-Castro L, et. al in their study [1] investigated the use of machine learning (ML) on healthcare data to predict the recurrence of breast cancer. The study concludes that the XGB model performs best when integrating structured and unstructured data, with precision = 0.900, recall = 0.907, F1-score = 0.897, and AUROC = 0.807. The best results are obtained with structured data, but natural language processing can get results that are equivalent with less mapping work, which emphasizes the usefulness of machine learning algorithms for risk assessment and cancer survivorship tracking. The classification of Single Photon Emission Computed Tomography – Myocardial Perfusion Imaging (SPECT-MPI) for coronary artery disease using convolutional neural networks (CNNs) is the subject of the study conducted by Magboo, et. al [2]. Similar excellent performance is shown by VGG16, InceptionV3, and DenseNet121, which offer nuclear medicine doctors invaluable decision support when analyzing SPECT-MPI tests. The models have the potential to improve clinical practice in nuclear cardiology by serving as trustworthy secondary assessments and instructional resources.

Similarly K.Shailaja, et. al in her study [3] focuses on the critical role that machine learning (ML) plays in healthcare, particularly in the area of disease prediction, while highlighting the ML's widespread influence across other industries. By examining several machine learning methods, such as Decision Trees and Support Vector Machines, it fills a research void and advances the creation of effective decision support systems for use in medical settings. While E. S. Tumpa and K. Dey in their study examined the broad applications of machine learning (ML) in the fields of security, industry, and medicine with a focus on ML's ability to forecast disease. In reviewing several machine learning algorithms, the study offers insights into the approach for healthcare applications. It talks on machine learning's contribution to healthcare, particularly in the areas of illness diagnosis and cost containment, and how it might improve people's daily life in the present epidemic.

The article[5] by Kavitha S, et. al focuses on applying statistical tools to analyze consumer interest, behaviour, and product revenues in order to anticipate corporate futures. The paper compares and contrasts Linear Regression and support vector regression models for time series data forecasting, highlighting the importance of insights from recent or past data. The comparison seeks to determine which model is most suited for making precise forecasts in fields like stock markets, banking, and weather forecasting. In order to enable computer control based on inferred thoughts, the study by K. S. Aswin, et. al explores the field of brain-computer interface, or BCI. It includes signal detection, feature extraction, and brain wave classification in the capture and classification of EEG signals. The study investigates the use of brain waves to steer a wheelchair. Precision, recall, and F1-score are used to compare the performance of artificial neural networks and deep neural networks. Furthermore, a web application is developed for improved control and prediction in BCI applications.

III. METHODOLOGY

Data preparation is essential when you get your dataset. After doing data cleaning and data binning the data is ready for the Machine Learning. Later, we used gradient boosting, Support Vector Regression (SVR), Artificial Neural Networks (ANN), and Linear Regression to forecast the prediction of the breast cancer detection. Using a variety of methods improves resilience and accuracy, offering a

complete solution for trustworthy forecasts that are essential to our research projects and analytical procedures. We evaluate different machine learning models using Mean Squared Error (MSE) as our selected loss function. This metric enables thorough assessment and guarantees accuracy when quantifying the difference between expected and actual values, which is crucial for fine-tuning and optimizing model results inside our analytical framework.

1) Logistic Regression

A statistical technique called logistic regression is frequently used for tasks involving binary classification, in which the result variable might have one of two potential categorical values. Numerous disciplines, including machine learning, statistics, and epidemiology, heavily rely on this approach. The capacity of logistic regression to forecast the likelihood that a given instance falls into a specific class is one of its unique selling points. The logistic function, sometimes referred to as the sigmoid function, is used in logistic regression to represent the probability as opposed to linear regression, which predicts continuous values. When there are just two possible classes—0 and 1—for the categorical outcome variable, logistic regression is mostly utilized in binary classification situations.

2) Artificial Neural Networks

A machine learning model called an Artificial Neural Network (ANN) is modelled after the neural architecture of the human brain. It is made up of layers of connected nodes, or artificial neurons. An input layer, one or more hidden layers, and an output layer are the most common types of these layers. ANNs are utilized in many different applications, including predictive modelling, natural language processing, and image identification.

Structure: An input layer, one or more hidden layers, and an output layer make up the structure of ANNs. Every neuron in one layer of the network is connected to every other layer's neuron, allowing information to flow through the network. In order to catch intricate patterns in the data, the associated weights of the connections are modified during training. Artificial Neural Network was earlier used to predict the weather accurately [9], to categorize earthquake vibrations [10] and also for processing fingerprint image noise [11]. An Artificial Neural Network (ANN) model's architecture, comprising its input layer, hidden layers, and output layer, is shown in detail in the fig 5.1.

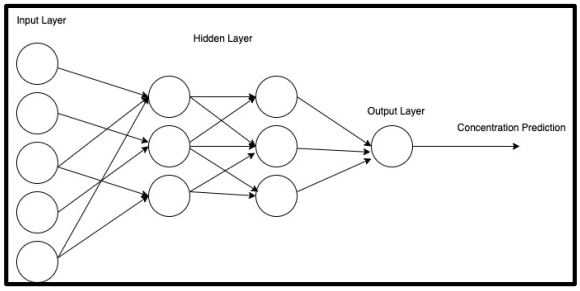


Fig 5.1 Artificial Neural Network’s structure

3) Support Vector Machines

Overview: Support Vector Regression (SVR) seeks to identify a hyperplane that maximizes the epsilon-tube error margin while fitting the training data. The goal is to keep this margin as small as feasible while making sure that the majority of the data points fall inside it. Data points that sit outside or near the margin are of particular interest in SVR because they have a big impact on the regression model.

Support Vector Regression (SVR) is illustrated in fig 5.3, where the focus is on the support vectors—that is, the important data points that have a major impact on the decision boundary of the model. An additional figure fig 5.4 demonstrates how data is transformed via the kernel function of the SVR, clarifying how non-linear mapping improves the model's performance.

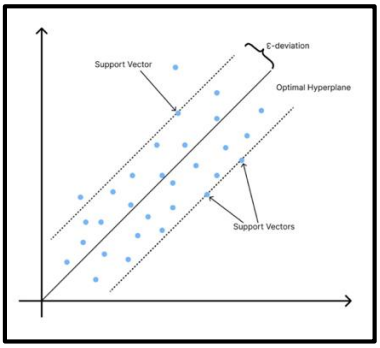


Fig 5.2 Decision boundary

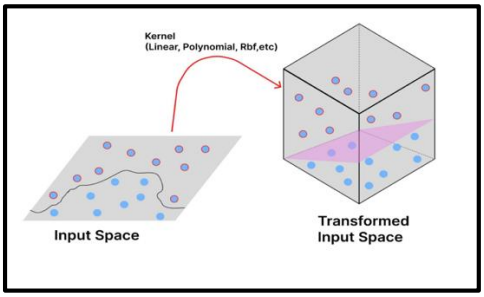


Fig 5.3 Transformed Input space and Kernel Function

IV. RESULTS AND ANALYSIS

MAP (Mean Average Precision)

A popular statistic in information retrieval and information retrieval systems, notably in the fields of computer science and machine learning, is MAP (Mean Average Precision). It is generally used to assess how well ranked list retrieval algorithms, such as those used by search engines, recommendation engines, and object detection models, perform.

By analysing how closely these ranked lists correspond to the real pertinent things, MAP evaluates the relevancy of the lists. The MAP formula is as follows: (1 / N) * (i=1 to N) (Precision at Rank i * Relevant at Rank i).
This formula reads:
The number of items that were found in the ranked list overall is N.
The percentage of pertinent items among the first i items in the list is known as precision at rank i.
The binary value "Relevant at Rank i" indicates whether or not the item at that rank is relevant.

Recall

Recall is a metric used to assess the effectiveness of classification and information retrieval systems, notably in machine learning and information retrieval applications. Recall is also known as Sensitivity or True Positive Rate. It gauges a system's capacity to accurately pick out each pertinent item from a dataset.
Recall works like this: True Positives / (False Negatives + True Positives).
This formula reads:
True Positives (TP) are instances that the system accurately classifies as positive (relevant).
False Negatives (FN) are situations that the system mistakenly categorised as negative even though they were positive (relevant).

Precision

In especially in machine learning and information retrieval applications, precision is a statistic used to assess the effectiveness of categorization and information retrieval systems. It assesses a system's capacity to appropriately select pertinent items while excluding unimportant ones.
The Precision equation is: True Positives / (True Positives + False Positives)
This formula reads:
True Positives (TP) are instances that the system accurately classifies as positive (relevant).

False Positives (FP) are situations that the system mistakenly categorised as positive even when they are truly negative (irrelevant).

A dimensionality reduction method called principal component analysis (PCA) converts high-dimensional data into a lower-dimensional representation. By lowering input characteristics while keeping crucial information, PCA improves the model's efficiency when paired with Random Forest. As an ensemble learning approach, Random Forest gains from PCA by enhancing computational efficiency and reducing the possibility of overfitting. In situations with big datasets and plenty of input variables, the combination enables faster training and prediction while preserving model accuracy with a smaller feature set.

1) Logistic Regression

One statistical technique for binary classification is logistic regression. It forecasts the likelihood that an instance will fall into a specific class. The logistic function is used to alter the output, producing numbers between 0 and 1. Based on probabilities, a threshold is used to categorise cases into the appropriate classifications.

With PCA	Mean Square Error	R-squared	Accuracy
	0.24561403508771928	-0.17352941176470593	Accuracy: 0.7544

Table 1.2 Confusion Report

189	11
59	26

	precision	recall	F1-score	support
0	0.76	0.94	0.84	200
1	0.70	0.31	0.63	85
Accuracy			0.75	285
Macro average	0.73	0.63	0.63	285
Weighted average	0.74	0.75	0.72	285

Without PCA	Mean Square Error	R-squared	Accuracy
	0.24912280701754386	-0.19029411764705895	Accuracy: 0.7509

Table 1.3 Confusion Matrix

187	13
58	27

	precision	recall	F1-score	support
0	0.76	0.94	0.84	200
1	0.68	0.32	0.43	85
Accuracy			0.75	285
Macro average	0.72	0.63	0.64	285
Weighted average	0.74	0.75	0.72	285

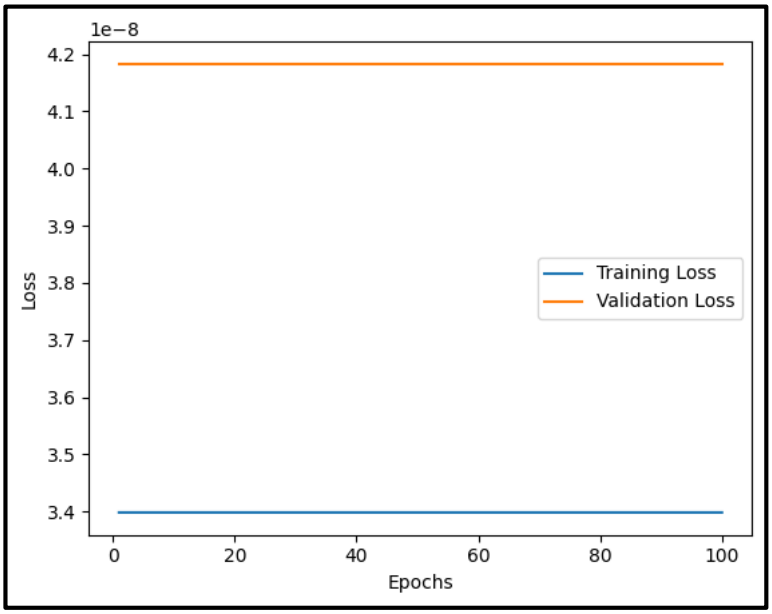
In this case we can clearly see that after applying PCA the accuracy of the system increases as it is able to determine the best fit algorithm for the same.

2) Artificial Neural Networks

Computational models that are modelled after the human brain are called artificial neural networks, or ANNs. They are made up of layers of networked nodes, or neurons. ANNs use data to learn and create classifications or predictions. Using strategies like backpropagation, the model modifies its weights during training to reduce prediction errors.

With PCA	Accuracy
	Accuracy: 0.7149

Without PCA	Accuracy
	Accuracy: 0.5702



Epochs vs Loss curve

However in ANN PCA doesn't provide us the demanded results, but it is not able to determine the intricate patterns in the dataset.

3) Support Vector Machines

Supervised learning models called Support Vector Machines (SVM) are applied to regression and classification problems. The hyperplane that best divides data into classes is found using SVM. It seeks to reduce classification mistakes while increasing the margin between classes. Through the use of kernel functions, SVM can handle non-linear interactions and is efficient in high-dimensional areas.

With PCA	Accuracy
	Accuracy: 0.65

	precision	recall	F1-score	support
0	0.65	1.00	0.79	37
1	0.00	0.00	0.00	20
Accuracy			0.65	57
Macro average	0.32	0.50	0.39	57
Weighted average	0.42	0.65	0.51	57

Without PCA	Accuracy
	Accuracy: 0.70

	precision	recall	F1-score	support
0	0.72	0.89	0.80	37
1	0.64	0.35	0.45	20
Accuracy			0.70	57
Macro average	0.68	0.62	0.62	57
Weighted average	0.69	0.70	0.70	57

When the underlying relationships in the data are extremely non-linear, PCA may not work as intended. In order to capture complex patterns, models such as Logistic Regression, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) may need access to the original feature space. These models may perform worse if PCA ignores non-linear information since they may find it difficult to understand complicated relationships. To better reflect the intricacies of the data in such circumstances, it is imperative to assess model performance without dimensionality reduction or investigate non-linear dimensionality reduction strategies.

V. Conclusions

Principal component analysis, or PCA, might face difficulties in situations when the correlations that are inherent to the data are very non-linear. Due to their reliance on the inherent structure of the input data, machine learning models like Logistic Regression, Support Vector Machines, and Artificial Neural Networks (ANN) may perform less well as a result of this constraint.

To determine the highest variance along each main component, PCA converts the data into a new set of orthogonal variables. On the other hand, the linear transformation imposed by PCA might not sufficiently capture the intricate patterns present in the data when the underlying data relationships are non-linear. This is particularly true for models that frequently depend on capturing complex non-linear interactions in order to produce correct predictions, such as Logistic Regression, SVM, and ANN.

Because they are sophisticated neural networks, ANNs can identify subtle non-linear relationships in the data. However, ANNs could find it difficult to retrieve these patterns if PCA ignores significant non-linear information during the dimensionality reduction process. This could prevent the ANN from adapting to non-linearities and result in less-than-ideal performance.

SVMs look for the best hyperplane in a high-dimensional space for classification. Reduced classification accuracy can occur if SVMs fail to sufficiently separate data points due to PCA's removal of non-linear information that is essential for determining class borders. SVMs can make use of non-linear kernel

functions; however, the kernel might not adequately reflect the complexity of the original data if PCA discards important non-linear features.

A linear decision boundary is the foundation of the linear model for binary classification known as logistic regression. In the event that PCA removes non-linear characteristics, Logistic Regression can have trouble appropriately fitting the data. It might classify cases incorrectly that call for non-linear decision boundaries, which would jeopardize the accuracy of the model.

In conclusion, care must be taken when combining PCA with ANN, SVM, or logistic regression when working with non-linear data. To better capture the complex relationships within the data and enhance overall model performance, it is imperative to assess model performance both with and without dimensionality reduction and, if appropriate, take into consideration different non-linear dimensionality reduction strategies. Therefore while using Streamlit for web deployment we need to use logistic regression in place of SVM and ANN as the accuracy is more when compared to others.

VI. REFERENCES

[1] González-Castro L, Chávez M, Duflot P, Bleret V, Martin AG, Zobel M, Nateqi J, Lin S, Pazos-Arias JJ, Del Fiol G, López-Nores M. Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records. Cancers (Basel). 2023 May 13;15(10):2741. doi: 10.3390/cancers15102741. PMID: 37345078; PMCID: PMC10216131.

[2] Magboo, V.P.C. and Magboo, Ma.S.A. (2021). Machine Learning Classifiers on Breast Cancer Recurrences. Procedia Computer Science, 192, pp.2742–2752. doi:https://doi.org/10.1016/j.procs.2021.09.044.

[3] K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918.

[4] E. S. Tumpa and K. Dey, "A Review on Applications of Machine Learning in Healthcare," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1388-1392, doi: 10.1109/ICOEI53556.2022.9776844.

[5] Kavitha S, Varuna S and Ramya R, "A comparative analysis on Linear Regression and support vector regression," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, 2016, pp. 1-5, doi: 10.1109/GET.2016.7916627.

[6] K. S. Aswin, M. Purushothaman, P. Sritharani and A. T. S, "ANN and Deep Learning Classifiers for BCI applications," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1603-1607, doi: 10.1109/ICICICT54557.2022.9917834.

[7] Y. Shimakura et al., "Short-term load forecasting using an artificial neural network," [1993] Proceedings of the Second International Forum on Applications of Neural Networks to Power Systems, Yokohama, Japan, 1993, pp. 233-238, doi: 10.1109/ANN.1993.264285.

[8] M. Huang, "Theory and Implementation of linear regression," 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 2020, pp. 210-217, doi: 10.1109/CVIDL51233.2020.00-99.

[9] D. V. Rayudu and J. F. Roseline, "Accurate Weather Forecasting for Rainfall Prediction Using Artificial Neural Network Compared with Deep Learning Neural Network," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICECONF57129.2023.10084252.

[10] F. A. Tasa, Istiqomah, M. A. Murti and I. Alinursafa, "Classification of Earthquake Vibrations Using the ANN (Artificial Neural Network) Algorithm," 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), BALI, Indonesia, 2022, pp. 102-107, doi: 10.1109/IAICT55358.2022.9887421.

[11] K. Han, "Artificial Neural Network for Processing Fingerprint Image Noise," 2022 23rd ACIS International Summer Virtual Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Summer), Kyoto City, Japan, 2022, pp. 9-14, doi: 10.1109/SNPD-Summer57817.2022.00011.