RESEARCH PAPER

AI-Based Classification and Prediction of Breast Cancer

Deepti Rani Pattanaik[1], Monalisha Pattnaik[2*], Deepak Kumar Routray[3] , Kailash Chandra Nayak[4] , Chandra Sekhar Jena[5]

[1,2 & 5] Department of Statistics, Sambalpur University, Sambalpur, Odisha, India

[3]Department of Statistics, Ispat Autonomous College Rourkela, Odisha, India

[4]Department of Business Administration, United School Of Business Management, Odisha, India

Corresponding Author: Monalisha Pattnaik

Email: monalisha_1977@yahoo.com

Abstract

Cancer in Breast is still among the most predominant and lethal disease globally, predominantly among females. Primary and precise analysis plays a vital role in improving patient survival rates and enhancing treatment approaches. This study explores the application of "Machine Learning" (ML) and "Deep Learning" (DL) techniques to enhance breast cancer detection and classification. We use more than 699 patients worth of data with 9 predictors using a random split of 70/30 for train/test using "Root Mean Squared Error" (RMSE) for measuring the performance in "Decision Tree" (DT) analysis. Furthermore, using the DT, we found that only 6 risk factors namely "uniformity of cell size", "uniformity of cell shape", "clump thickness", "normal nucleoli", "bare nuclei" and "marginal adhesion", was adequate to detect breast cancer of the patients with "RMSE" of 0.1717. By analyzing risk factors and employing advanced algorithms such as DT, "Principal Component Analysis" (PCA), "Linear Discriminant Analysis" (LDA),"Artificial Neural Networks" (ANNs), hybrid DT-ANN, Deep Neural Networks (DNNs), hybrid DT-DNN and TabNet, we aim to identify significant predictors of breast malignancy. TabNet model has shown finest results for diagnosing breast cancer with accuracy of 0.99863 and 0.99057 for train and test data set respectively which outperforms other models.

Keywords: TabNet model, DNN, ANN, Decision tree, Hybrid model, Breast cancer

How to cite this article: Pattanaik DR, Pattnaik M, Routray DK, Nayak KC, Jena CS. AI-Based Classification and Prediction of Breast Cancer. Int J Drug Deliv Technol. 2026;16(20s): 86-98. DOI: 10.25258/ijddt.16.20s.11

1. Introduction

One of the most often diagnosed malignancies in women is breast cancer.. Patients can acquire an initial diagnosis of breast cancer and its metastases based on a dependable technique, increasing their chances of survival to over 86%. When malignant tumor lumps begin to form in the breast cells, breast cancer starts. It is possible for doctors to mistakenly classify benign (non-cancerous) tumors as malignant ones. So CAD, or computer-aided detection, is necessary. Machine learning techniques are utilized by certain systems to provide precise breast cancer diagnoses. These computer-aided detection (CAD) systems play a key role in early detection, which is vital for refining treatment options and growing survival rates. Early diagnosis is essential to lowering the high mortality rate associated with cancer. Breast cancer can develop in female of any age after adolescence in any country, but the risk increases with age. In countries with very high Human Development Indexes (HDIs), one in twelve females may be detected with breast tumor during their lifetime, and one in seventy-one may die from it. Globally, it was the greatestshared cancer amid womenfolk in 157 out of 185 countries in 2022. Additionally, breast cancer affects men in approximately 0.5–1% of cases. Even with early uncovering and the development of innovative treatments, nearby 50% of affected role will advancereserved metastases throughout their continuation period. Conferring to evaluations from World Health Organization, 1.5 million Indian patients are spotted with breast cancer annually, and the illness sued the lives of 500,000 females in 2015 and around 9.6 million in 2018. The significant discrepancy between incidence and death suggests that early breast cancer performance needs to be enhanced. As a result, improvements in existing techniques are required for the early prediction of breast malignance.

There exist two distinct forms of breast cancer, Non-cancerous Benign Tumours and Malignant or Cancerous Tumours. If the cells are non-cancerous, the tumor is considered benign.It does not encroach upon adjacent tissues or disseminate to other areas of the body.. While generally less dangerous, a benign tumor can still cause harm if it exerts pressure on nearby tissues, blood vessels, or nerves. Examples of benign tumors include adenomas, chondromas, fibromas (or fibroids), and hamartomas, among others. On the other hand, malignant tumors consist of

cancerous cells that can riseoverpoweringly and conquerneighbouringnerves. These tumorcubicles are often abnormal and significantly different from the surrounding healthy tissue. Common types of malignant tumours include carcinomas, sarcomas, leukemia, and lymphomas.

A mixture of environmental, behavioural, and genetic variables are the most significant risk factors for breast cancer. Some risk factors can be controlled, but others are beyond of our control. Some Non-modifiable Risk Factors includes; Gender, Age, Family History, Gene Mutation, Early Menstruation, Later Menopause, Hormonal Replacement Therapy etc. Similarly Modifiable Risk Factors includes Reproductive History, Alcohol Consumption, Obesity, Radion Therapy, Oral Contraceptives, Physical Inactivity, Diet etc.

A mammography-based breast cancer risk model has been developed, outperforming established models. The model uses patient questionnaires and electronic medical records to assess risk factors within 5 years. Three models were developed: RF-LR, DL, and hybrid DL. Comparisons were made to an established model by (Yala et al., 2019). Comparison of various neural networks for classifying breast cancer tumours was undergone and it was found that the Probabilistic Neural Network (PNN) demonstrated the best detection results, outperforming other networks verified on the “Wisconsin Breast Cancer Database” (Khan et al., 2021).Astudy investigated three black-box interpretation techniques, “Feature Importance”,“Partial Dependence Plot”, and “LIME” applied to Multilayer Perceptron and Radial Basis Function Networks for breast cancer diagnosis. The findings revealed, limited LIME enlightenments were consistent with universal elucidations (Hakkoum et al., 2021). Deep learning approach conducted using “CNN”typical for manifold breast cancer classification, achieving high processing presentations with 95.4% accurateness compared to high-tech models on histopathological pictures (Nawaz, et al., 2018). Using Machine learning algorithm, classification was done on triple negative and non-triad negative breast cancer affected role based on genetic factor countenance statistics. They evaluated four different classification models..The “SVM” algorithm outperformed others, achieving higher accuracy and fewer misclassification errors in distinguishing between the two types of breast cancer (Wu & Hicks., 2021). An evaluation of various machine learning system was conducted using the Wisconsin Breast Cancer dataset from Machine learning repository. The goal was to assess the algorithms performance in classifying data based on “accuracy”, “precision”, “sensitivity”, and “specificity”. The outcomes showed that “SVM” achieved the peak precision at 97.13%, along with the lowermost error amount (Asri et al., 2016). Additionally, comparison of machine learning algorithms such as “Random Forest”, “k-NN”, and “Naive Bayes” for breast cancer forecast was performed using Wisconsin Diagnostic Breast Cancer dataset, yielding competitive results for detection and treatment (Sharma et al., 2018). Machine learning procedures were evaluated for breast malignance recurrence prediction using a dataset from 1997-2014. Results showed challenges in obtaining a representative dataset, no consensus on best predictors, high accuracy but compromise in sensitivity, and ineffective performance metrics. Future directions include combining different techniques and defining standard predictors for better results (Aberu et al., 2016). A detailed survey focusing on breast cancer screening techniques was conducted, offering a comprehensive analysis of their advantages and limitations. The study also discovered the pertinency of deep learning methods in breast cancer exposure, appraising various performance metrics and datasets used in this domain. Additionally, future research directions related to breast cancer detection were examined. The paper objects to offer a systematic understanding of this field while encouraging innovative research efforts (Rautela et al., 2022). A study investigated the correctness and effectiveness of “Machine learning (ML)” and “Deep learning (DL)” methods for primary breast cancer detection using alphanumeric mammography metaphors and microelectronic well-being histories. It identified 48% of false-negative cases and attained an “AUC of 0.91”, with a “specificity of 77.3%” and a “sensitivity of 87%” (Ballin et al., 2019). Additionally, a hybrid DL model was developed for the automatic detection of “invasive ductal carcinoma (IDC)” using the PCam Kaggle dataset. This model, which combines CNN and GRU architectures, delivered superior performance by surpassing pathologist-level accuracy, addressing misclassification issues, and outperforming other models (Wang et al., 2022). An augmented “Deep recurrent neural network (RNN)” model was developed using Keras-Tuner optimization. The architecture included input, hidden, dropout, and output layers. By utilizing feature-selection techniques, the optimized RNN outperformed five

traditional machine learning models (Saleh et al., 2022). A review of publicly available datasets for breast cancer diagnosis explored current deep learning approaches, introduced code repositories, and highlighted challenges and future opportunities in DL-based diagnosis (Iqbal et al., 2022).A separate study introduced a deep learning framework aimed at identifying breast anomalies by utilizing standard data. This framework incorporated pre-processing and feature extraction with a MobileNetV2 pre-trained model, achieving superior performance compared to previous methods and excelling in anomaly detection on the IN breast and MIAS datasets (Alloqmani et al., 2023). Machine learning algorithms were also employed to classify crucial risk influences for prime attacking breast cancer in the Iranian inhabitants. Significant predictors included post-menopausal status, intentional weight loss, abortion history, chest X-ray exposure, employment, menarche age, education level, second-hand smoke exposure, age at first delivery, and breastfeeding duration. The Random Forest model demonstrated the highest AUC among the methods tested (Nasab et al., 2023). In addition, various machine learning techniques were used to classify breast cancer tumours. Using a dataset split into “60% training”, “20% validation”, and “20% testing”, the classifiers achieved precision, recall, and F1 scores of 0.72, 0.80, 0.81, 0.82, and 0.82, respectively. Among the methods evaluated, the Exception prototypical achieved the maximum recital, with a “precision”, “recall”, and “F1 score” of 0.90, outstripping additional methods (Yadavendra & Chand, 2020).

Identifying risk factors is a critical step in the primary uncovering and dealing of breast cancer. Conventional statistical approaches for determining breast cancer risk factors typically involve the analysis of extensive datasets to explore the relationships between various potential risk factors and the probability of developing the disease. These approaches aim to quantify the associations between risk factors and outcomes, as well as to uncover patterns that can guide public health recommendations and individual risk evaluations. Commonly employed traditional statistical techniques in breast cancer research include logistic regression, survival analysis (Cox models), Chi-Square tests, ANOVA, Pearson correlation, multivariate regression analysis, factor analysis, meta-analysis, and descriptive statistics. These methods yield essential insights into how different demographic, genetic, and lifestyle factors influence the likelihood of breast cancer development. By applying these techniques, researchers can prioritize interventions and preventive strategies for individuals at elevated risk, ultimately enhancing outcomes in breast cancer detection and prevention. However, while these traditional statistical methods are widely utilized and beneficial, they also present certain limitations, such as assumptions inherent in statistical models, multicollinearity, model overfitting and underfitting, limited capacity to handle complex interactions, and a lack of flexibility for non-linear relationships. These confines can distress the “accuracy”,“generalizability”, and “interpretability” of the findings.

To address the limitations inherent in traditional statistical methods, researchers frequently adopt more sophisticated statistical approaches, together with different AI models. These advanced techniques are adept at managing non-linear relationships, extensive datasets, and intricate interactions with greater efficacy. By employing these methods, researchers can more effectively arrest the complexities and restraints of data, resulting in enhanced accuracy, scalability, and personalized insights regarding breast cancer risk. Machine learning and deep learning present numerous advantages over conventional statistical techniques on identification of breast cancer risk aspects. They are particularly proficient in processing large, complex datasets, performing feature extraction, recognizing non-linear relationships, and delivering more precise, individualized risk evaluations. These methodologies can reveal patterns and interactions that traditional approaches might overlook, thereby serving as essential tools for the advancement of breast cancer research and the enhancement of clinical decision-making.

In this research endeavour, our primary aim is to ascertain the most significant risk factors associated with breast cancer. Subsequently, we will classify breast cancer based on the characteristics of various risk factors by employing a range of MLand DL techniques, as well as PCA, LDA, ANN, DNN, a hybrid model combining Decision Trees with Artificial Neural Networks (DT-ANN), and a model that integrates Decision Trees with Deep Neural Networks (DT-DNN).

The remainder of this work is structured as follows: The metrics for "performance assessment," "prerequisite data analysis," and "datasets" are provided in Section 2. The detailed conception of several AI-based models is covered in Section 3. Section 4 displays the experimental evaluation and

results for each model. The results, practical implications, and concluding remarks were covered in detail in Section 5.

2. Data and Preliminary Analysis

This breast cancer databases wasgained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. In this study the breast cancer disease dataset is extracted from the Kaggle repository: "https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)" to conduct the analysis. This dataset comprises of 699 observations and 11 variables for the patients: "Sample code number id number", "Clump Thickness 1 – 10", "Uniformity of Cell Size 1 – 10", "Uniformity of Cell Shape 1 – 10", "Marginal Adhesion 1 – 10", "Single Epithelial Cell Size 1 – 10", "Bare Nuclei 1 – 10", "Bland Chromatin 1 – 10", "Normal Nucleoli 1 – 10", "Mitoses 1 – 10", "Class (2 for benign, 4 for malignant)" out of which the first variable "Sample code number id number" is ignored for analysing purpose. There is a binary classification "Class" variable revealing if the patient is diagnosed for breast cancer disease or not and 9 risk or causal variables. The "Class", a categorical variable with 2 categories namely "Benign", and "Malignant" is recoded as "0 and1" respectively and is the response variable with the 9 measurement variables, all of which are integers, and potential explanatory variables for AI-based model. There are 9 predictors to diagnose breast cancer in patients. Table 1 shows the description of the variables related to breast cancer disease of the patients. Figure 1 shows the box plot of all the variablesassociatedwithbreast cancer disease of the patients.

Table 1: Descriptive Statistics of the Variables of Breast Cancer

Sl. No.	Variables	Variable Types	Minimum	Q1	Median	Mean	Q3	Standard Deviation	Maximum	Skewness	Kurtosis
1	Clump. Thickness	Integer	1.000	2.000	4.000	4.418	6.000	2.8157	10.000	0.5915	2.3721
2	Uniformity of Cell Size	Integer	1.000	1.000	1.000	3.134	5.000	3.0514	10.000	1.2304	3.0895
3	Uniformity of Cell Shape	Integer	1.000	1.000	1.000	3.207	5.000	2.9719	10.000	1.1593	2.9983
4	Marginal Adhesion	Integer	1.000	1.000	1.000	2.807	4.000	2.8553	10.000	1.5211	3.9723
5	Bland Chromatin	Integer	1.000	2.000	3.000	3.438	5.000	2.4383	10.000	1.0976	3.1747
6	Single Epithelial Cell Size	Numeric	1.000	2.000	2.000	3.216	4.00	2.2142	10.000	1.7084	5.1450
7	Bare. Nuclei	Integer	1.000	1.000	1.000	3.567	6.000	3.6165	10.000	0.9735	2.1957
8	Normal Nucleoli	Numeric	1.000	1.000	1.000	2.867	4.000	3.0536	10.000	1.4192	3.4623
9	Mitoses	Numeric	1.000	1.000	1.000	1.589	1.000	1.7150	10.000	3.5530	15.558
10	Class	Numeric	0.000	0.000	0.000	0.341	1.000	0.4746	1.0000	0.6665	1.4442

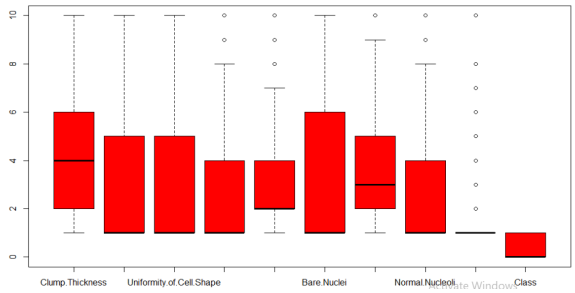

Figure 1: The Box Plot of All the VariablesRelated to Breast Cancer

2.1 Performance Measurement Metrics

The "RMSE" is implementedto measure performance ofseveralforecast models with the medical data. For this performance assessmentsystem of measurement, the "formulae" is expressed as follows:

$$"RMSE" = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(w_i - \widehat{w}_i)^2}$$

Where,"n"denotesthe number of data points in the time series,"w_i" is true value, while "$\widehat{w}_i$" is projected value. The performance metric's least value shows the improved performance of the model.

3. Methodology

TabNet(Arik & Pfister, 2021)is aadvanced deep learning model preciselyplanned for tabular data, making it a strong candidate for binary classification tasks like differentiating between Alzheimer's and Parkinson's diseases. By assimilation the interpretability of decision trees with the capabilities of deep learning, "TabNet" can professionally achieve composite datasets. "TabNet" is a valuable tool for medical data analysis and decision-making because of its capacity to provide insights about feature importance, which further improves the interpretability of classification findings..We have employed many machine learning models like "decision tree","ANN","DNN","hybrid decision tree" and "ANN (DTANN)","decision treeand DNN" (DTDNN),"boosting algorithms" and "TabNet". The suggested "TabNet" models outperform all the studied models.

3.1 Decision Tree

The "Decision Tree"procedure is a highly adaptable and effective instrument in the dominion of machine learning models, applicable to both classification and regression problems. Its straightforward design enhances its interpretability. In the context of breast cancer classification, a Decision Tree (DT) algorithm can be employed to ascertain whether a tumor is malignant (cancerous) or benign (non-cancerous) by analysing various characteristics of the tumor. A data mining approach utilizing decision trees is introduced for the early identification of breast cancer, achieving optimal accuracy and thereby facilitating patient recovery. Breast cancers are divided into benign tumors, which do not invade surrounding tissues, and malignant tumors, which possess the ability to penetrate adjacent tissues (Tarawnehet al., 2022).

3.2 Principal Component Analysis

"Principal Component Analysis (PCA)" is a statistical procedureapplied to decrease dimensionality while keeping the supreme variance existing in the dataset. By removing redundant and correlated variables, PCA simplifies the feature set, thereby enhancing the effectiveness and speed of classification task. The efficacy of 3 popular data mining algorithms, "Naïve Bayes (NB)", "K-nearest neighbour (KNN)", and "Decision tree(DT)"was evaluated using principal component analysis (PCA) as the feature extraction method to predict the recurrence of breast cancer, as reported by (Zain et al., 2018).

3.3 Linear Discriminant Analysis

"Linear Discriminant Analysis (LDA)" is also a supervised machine learning procedure primarily used for classification responsibilities and dropping "dimensionality". This technique is especially advantageous when the separation of classes is of utmost importance. LDA seeks to identify linear combinations of features that optimally differentiate the data into separate categories. In a study involving breast cancer images processed with autoencoders, features identified by convolutional neural network models were utilized. Subsequently, the most significant features were determined through the application of ridge regression, and classification was performed using linear discriminant analysis (Toğaçar et al., 2020).

3.4 Artificial Neural Network

Breast cancer prediction represents a vital field of inquiry within healthcare, wherein machine learning models, including "Artificial Neural Networks (ANNs)", are employed to assess the probability of breast cancer occurrence built on medical information. Artificial Neural Networks (ANNs) excel with handling intricate tasks like "Classification","Regression", "Image recognition", and "Natural language processing". A method was proposed utilizing a Back Propagation Neural Network (BPPN) for the automated classification of images related to breast cancer diagnosis by (Kaymak et al.,2017).

3.5 Deep Neural Network

The “Deep Neural Networks (DNNs)” are a commanding class of machine learning depictions that have reformed fields such as “Computer vision”, “Natural languageprocessing” and “Speech recognition”. Their capability to acquire hierarchical facts representations make them highly effective for complex tasks. A DNN classification model was employed alongside “Recursive feature elimination (RFE)” for feature selection. With its manifold processing layers, the DNN achieved superior classification accuracy compared to support vector machines (SVM). Consequently, the researchers opted for a deep learning approach to classify hyper-spectral data discussed by (Karthik et al., 2018).

3.6 TabNet

A deep learning architecture called “TabNet” was created for tabular data., merging the strengths of decision trees and neural networks. Unlike traditional gradient-boosting methods, TabNet uses an interpretable attention-based mechanism to select relevant features at each decision step, improving both accuracy and explainability (Arik & Pfister, 2021). It employs sequential attention to focus on different subsets of input features, reducing unnecessary computation and enhancing efficiency. TabNet is particularly effective in supervised and self-supervised learning tasks, making it suitable for applications in finance, healthcare, and fraud detection. Additionally, it provides built-in feature importance insights, aiding in model transparency and interpretability.

4. Experimental Evaluation and Results of Diagnosing Breast Cancer

Recognizing the characteristics and features of the healthcare datasets is a thought-provoking task. The current study concentrates on diagnosing the breast cancer patients using the AI-based models such as DT, PCA-LDA,ANN, DT-ANN,DNN, DT-DNN and TabNet. The AI-based model of TabNet outperformed in comparison to other related models for prediction and classification of this disease. It is the best to inform the scientificspecialists and management or public-policy makers, and for dealing and controlling the breast cancer disease. Table 2 showsthe accuracy of prediction and classification of diagnosing the breast cancer patients.

4.1 Decision Tree Analysis for Diagnosing Breast Cancer

DT models are influential classifiers that use a tree data structure to represent the relationships among potential outcomes and key features. When the final choice can be reached, the tree is concluded with leaf nodes (or terminal nodes) that indicate the action to be taken based on the sequence of decisions. For predictive models, these leaf nodes offer the anticipated outcome given the series of events processed through the tree.

The decision to use a decision tree model for identifying significant input variables from 9 variables namely “clump thickness”, “uniformity of cell size”, “uniformity of cell shape”, “marginal adhesion”, “single epithelial cell size”, “bare nuclei”, “bland chromatin”, “normal nucleoli”, and “mitoses” for diagnosing breast cancer stems from the model's simplicity, ease of interpretability, and high accuracy. We applied an optimal decision tree model to a dataset comprising 699 patients’ data points to identify potential causal variables associated with breast cancer prediction. The DT was applied using the “rpart” package in R software, with a control parameter “minsplit” set to 10 percent of the total data. We evaluated the model's predictive performance evaluation using RMSE. The optimal DT, built with 6 input variables and a “minsplit” of 5, was created with the same costs for each variable. Figure 3 shows the variable importance list, and Figure 2 presents the fitted tree. From the decision tree analysis it is concluded that, six out of the 9 potential input variables were found to be highly significant. But captivatingly, we captured six essential causal variables like “uniformity of cell size”,‘uniformity of cell shape”,“clump thickness”,“normal nucleoli”,“bare nuclei”and “marginal adhesion” can be managed to diagnose the patients with breast cancer.

By using the “rpart” package, we can evaluate the range of cost complexities. For comparing the errors for each of the cost complexity value, rpart performs cross-validation with 10-fold by computing the errors on the validation data. Figure 2 displays the optimal decision tree, which has 8 internal and 9 terminal nodes respectively, and partitions the data based on six variables to develop its model. This tree can be expanded to a full tree with 9 terminal nodes by setting CP=0 shown in Figure 2. In Figure 4, the Y-axis represents cross-validation error, the upper X-axis indicates the number of terminal nodes and the lower X-axis displays cost complexity values. Beyond 9 terminal nodes, the reduction in error becomes less significant as the tree grows deeper. For predicting the breast cancer, the decision tree identified six variables out of the 9 variables: “uniformity of cell size”, “uniformity of cell shape”, “clump thickness”,

"normal nucleoli", "bare nuclei"and "marginal adhesion". All 699 patients' data points across 9 variables are processed through the decision tree illustrated in Figure 2, where each data point is evaluated at a node, proceeding left for a "Yes" answer and right for a "No" answer.Firstly, 699 data points of diagnosis of the breast cancer in comparison to all data points of 9 variables where uniformity of cell size is less than 3 go to left branch, all other data points of diagnosing breast cancer patients proceeds to the right branch.In case of patients having uniformity of cell size less than 3, clump thickness less than 7 and normal nucleoli less than 5 will have 59% chance to be diagnosed with no breast cancer. All the patients having uniformity of cell size less than 3, clump thickness less than 7 and normal nucleoli greater than or equal to 5 will have 1% chance to be diagnosed with breast cancer. All the patients having uniformity of cell size less than 3, clump thickness greater than or equal to 7 will have 1% chance to be diagnosed with breast cancer. All the patients having uniformity of cell size greater or equal to 3, uniformity of cell shape less than 3 and clump thickness less than 6 will have 3% chance to be diagnosed with no breast cancer. All the patients having uniformity of cell size greater or equal to 3, uniformity of cell shape less than 3 and clump thickness greater than or equal to 6 will have 1% chance to be diagnosed with breast cancer. All the patients having uniformity of cell size greater or equal to 3 and less than 5, uniformity of cell shape greater than or equal to 3, bare nuclei less than 3 and marginal adhesion less than 4 will have 2% chance to be diagnosed with no breast cancer. All the patients having uniformity of cell size greater or equal to 3 and less than 5, uniformity of cell shape greater than or equal to 3, bare nuclei less than 3 and marginal adhesion greater than or equal to 4 will have 0% chance to be diagnosed with breast cancer. All the patients having uniformity of cell size greater or equal to 3 and less than 5, uniformity of cell shape greater than or equal to 3, and bare nuclei greater than or equal to 3 will have 8% chance to be diagnosed with breast cancer. All the patients having uniformity of cell size greater or equal to 3, and uniformity of cell shape greater than or equal to 3 will have 25% chance to be diagnosed with breast cancer. With continuing the splitting process, evaluating all the variables with each split, until the data points of breast cancer are divided into 9 partitions. The predicted diagnosis of breast cancer of the patients is (0, 1, 1, 0, 1, 0, 1, 1, 1),based solely on six variables. The RMSE result of this tree model for evaluating predictive performance is 0.1717.

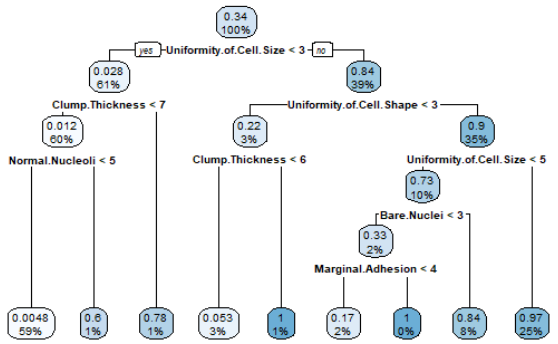

Figure 2: Decision Tree for Classification of Breast Cancer

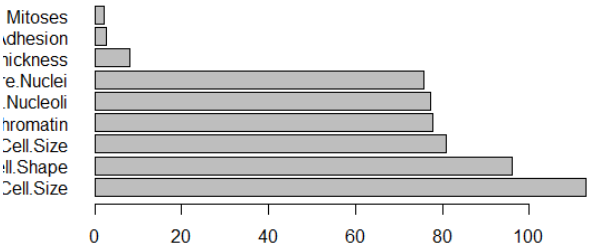

Figure 3: Variable Importance for Classification of Breast Cancer

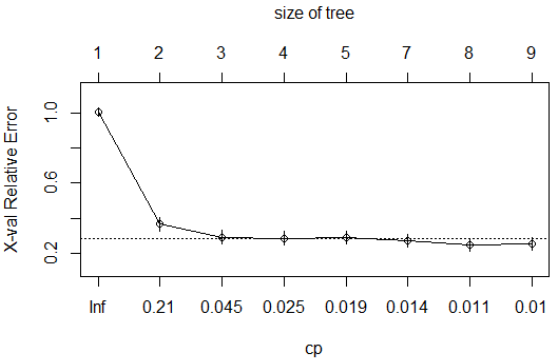

Figure 4: Complex Parameter for Classification of Breast Cancer

4.2 PCA-LDA for Diagnosing Breast Cancer

By smearing "Principal Component Analysis" (PCA) with altogethernine causal variable quantity for classifying the target variable "Class" of the patients diagnosing breast cancer disease. From this study it is detected that, only the topmost1 PC with Eigen value more than one, which is giving nearly 65.37 percentage of variance, explained. For finding the influence of the studied variables their loadings are taken into consideration. It investigates that 1 PC is derived from 9 studied variables for which the Eigen valueis greater than 1. Considering all the 1 PC result as the response variables we have used "Linear Discriminant Analysis" (LDA) to predict and classify

the target variable “Class” of the patients diagnosing breast cancer disease. The PCA-LDA model is applied to study the influence of causal variable on diagnosing the breast cancer disease of the patients. The total dataset of 699 patients is splited into the ratio of 70:30, with the purpose of using 70% of the data at our disposal to train the model and the residual 30% to test the model. One PCA-LDA model is developed with one input. Table 2 shows the accuracy level of both the train and test dataset of this model. It shows that this model has the better accuracy values for the training and test dataset compared to all models under study with [0.9507187, 0.9481132] for train and test data respectively to classify the patients diagnosing under breast cancer disease. Figure 5 shows the correlation plotting of the variables related to breast cancer. Figure 6 shows the scree plot with a red dashed line. Figure 7 shows the biplot of rows and columns variables and Figure 8 shows the biplot of variables with respect to PCs. Figure 9 and 10 show the plotting of AUC and discriminant analysis for classification and prediction of breast cancer of the patients.

4.3 Neural Networks Models for Diagnosing Breast Cancer

Analysis is conducted to diagnose breast cancer of the patients. In the first ANN model, 9 input variables are feed into a three-layered feed-forward neural network having 8 hidden neurons and with one output neuron. Similarly, second ANN model uses 9 input variables in a four-layered feed-forward neural network; it includes two hidden layers with 8 and 4hidden neurons respectively, and with one output neuron. The third hybrid decision tree and ANN (DT-ANN) model uses DT and 6 input variables in a three-layer feed-forward neural network, it includes one hidden layer with 4 neurons for diagnosing breast cancer and similarly, fourth hybrid decision tree and ANN (DT-ANN) model uses DT and 6 input variables in a four-layered feed-forward neural network, includes two hidden layers with 4 and 2 hidden neurons for diagnosing breast cancer. Five different DNNs, hybrid decision tree and DNN (DT-DNN) models, and TabNet DL model are also developed to diagnose breast cancer.

4.3.1 ANN Models for Diagnosing Breast Cancer

ANN disease prediction models can diagnose the breast cancer of the patients.The entire dataset is split into 70 percent and 30 percent to trainand testthe modelsrespectively. Two ANN models are developed: one with a single hidden layer of 8 neurons and another with two hidden layers of 8 and 4hidden neurons respectively. In Figures 11 and 12 show the best ANN and DT-ANN models with two hidden layers for diagnosing breast cancer of the patients with respect to the 9 and 7 inputs respectively. Similarly, hybrid model with single layer of 4 neurons of predicted value of decision tree model and six identified risk factors namely “uniformity of cell size”, “uniformity of cell shape”, “clump thickness”, “normal nucleoli”, “bare nuclei” and “marginal adhesion” constituted decision tree and ANN (DT-ANN) prediction models can diagnose the breast cancer. Table 2 shows the accuracy level for both train and test data to diagnose breast cancer of the patients. It is investigated that second ANN and DT-ANN model show better accuracy results with [0.9585879 and 0.9129747] and 0.9752577 and 0.9579439] for train and test data respectively to diagnose breast cancer of the patients.

4.3.2 DNN Models for Diagnosing Breast Cancer

Deep Neural Network (DNN) disease prediction models can diagnose the breast cancer of the patients. We create 5 instances of the model one can diagnose breast cancer disease accurately. 70 percent of the total dataset is used for training this model, while the remaining 30 percent is used for testing it. Activation functions namely ReLU and Sigmoid are used for hidden layer and output respectively. Adam optimizer and binary cross entropy loss are applied as the matrix to train this hybrid model. The architecture of the five different DNN models are with [32, 2], [32, 16, 2], [64, 2], [64-0.4, 32-0.3, 16-0.1, 2] and [100, 2] hidden layers and output respectively with respect to 9 inputs and with 200, 32 and 0.2 “number of epochs”,“batch size” and “validation split” respectively. Table 2 shows the accuracy level for both train and test data to diagnose breast cancer of the patients. It is examined that second DNN model shows better correctness results with [0.99381441 and 0.9719626] for train and test data to diagnose breast cancer of the patients.Similarly, hybrid model of predicted value of decision tree model and six identified risk factors namely“uniformity of cell size”,“uniformity of cell shape”,“clump thickness”,“normal nucleoli”,“bare nuclei”and “marginal adhesion”constituted decision tree and DNN (DT-DNN) prediction models can diagnose the breast cancer. 70 percent of the total dataset is used for training this model, while the remaining 30 percent is used for testing it. Activation functions namely ReLU and Sigmoid are used for hidden layer and output respectively. Adam optimizer and binary cross entropy loss are applied as the matrix to train this hybrid model. The architecture of the five

different DT-DNN models are with [32, 2], [32, 16, 2], [8, 2], [8, 4,2] and [16, 8, 2] hidden layers and output respectively. Table 2 shows the accuracy level for both train and test data to diagnose breast cancer of the patients. It is inspected that fifth DT-DNN model shows improved precision results with [0.9731985 and 0.9813084] for train and test data to diagnose breast cancer of the patients. In Figures 13 and 14 show the best DNN and DT-DNN models for diagnosing breast cancer of the patients with respect to the 9 and 7 inputs respectively.

4.3.2 TabNet Model for Diagnosing Breast Cancer

TabNet is a powerful deep learning model well-suited for breast cancer prediction, leveraging its unique attention-based feature selection mechanism. Given input features of “Clump Thickness”,“Uniformity of Cell Size”,“Uniformity of Cell Shape”,“Marginal Adhesion”,“Single Epithelial Cell Size”,“Bare Nuclei”,“Bland Chromatin”,“Normal Nucleoli”, and “Mitoses”, the model can effectively learn patterns to classify tumors as benign (0) or malignant (1). Since these are numerical features, normalization is performed to ensure that all values are on a similar scale. The dataset is then split into training and testing sets, typically using a 70/30 ratio. Unlike traditional machine learning models like Random Forest or XGBoost, TabNet dynamically selects the most relevant features at each decision step, improving interpretability and efficiency. This ability is particularly useful in medical applications where understanding feature importance is crucial for trust and transparency. Additionally, TabNet's ability to handle missing values without imputation makes it highly suitable for real-world clinical datasets, where missing data is common.

For training TabNet Classifier on the breast cancer prediction dataset, the following hyperparameters can be used: a learning rate of 0.02, batch size of 128, virtual batch size of 32, n_d and n_a (decision and attention layer neurons) set to 8, n_steps of 3, gamma of 1.5, lambda_sparse of 1e-3, momentum of 0.02, scheduler step of 10, epochs set to 100, and the AdamW optimizer for stability. The training procedure involves normalizing numerical values, splitting the dataset into 70% training and 30% testing, and initializing TabNet Classifier with these hyperparameters. The model is trained using binary cross-entropy loss, optimized with Adam W, and a learning rate scheduler for better convergence. Performance is evaluated using RMSE and MAE, where RMSE captures the squared error magnitude, and MAE provides a more interpretable absolute error measure. Additionally, TabNet's feature importance analysis helps identify the most influential predictors for breast cancer classification, offering both accuracy and interpretability in medical diagnostics. Table 2 shows the accuracy level for both train and test data to diagnose breast cancer of the patients. It is inspected that TabNet model shows the best precision results with [0.99863 and 0.99057] for train and test data to diagnose breast cancer of the patients.

Table 2: Performance Measure of Different Models for Diagnosing Breast Cancer of the Patients

Model	Inputs	Neurons in Input-Hidden-Output Layers	Accuracy		Error	
			Training	Testing	Training	Testing
PCA-LDA	1		0.9507187	0.9481132	0.0492813	0.0518868
ANN	9	8,2	0.9735772	0.9468599	0.02642276	0.0531401
ANN	9	8,4,2	0.9752577	0.9579439	0.02474227	0.04205607
DT-ANN	7	4,2	0.9731959	0.9672897	0.02680412	0.03271028
DT-ANN	7	4,2,2	0.9690722	0.9813084	0.03092784	0.01869159
DNN	9	32,2/200-32-0.2	0.97525775	0.97663552	0.02474225	0.02336448
DNN	9	32,16,2/200-32-0.2	0.99381441	0.9719626	0.00618559	0.0280374
DNN	9	64,2/200-32-0.2	0.98556703	0.9719626	0.01443297	0.0280374
DNN	9	64-0.4,32-0.3,16-0.1,2/200-32-0.2	0.98762888	0.9626168	0.01237112	0.0373832
DNN	9	100,2/200-32-0.2	0.9896907	0.9626168	0.0103093	0.0373832
DT-DNN	7	32,2/200-32-0.2	0.96701032	0.9719626	0.0329896	0.0280374
DT-DNN	7	32,16,2/200-32-0.2	0.98556703	0.96728975	0.01443297	0.03271025
DT-DNN	7	8,2/200-32-0.2	0.97113401	0.98130840	0.02886599	0.0186916
DT-DNN	7	8,4,2/200-32-0.2	0.9690722	0.98598129	0.0309278	0.01401871
DT-DNN	7	16,8,2/200-32-0.2	0.97319585	0.98130840	0.02680415	0.0186916

TabNet	9	--	0.99863	0.99057	0.008749	0.01056

The outcomes of all models are exhibited in Table 2. A TabNet deep learning model for two diseases shows higher values of accuracy both for training and testing than other models like ANNs, DNNs, DT-ANN, and DT-DNN.

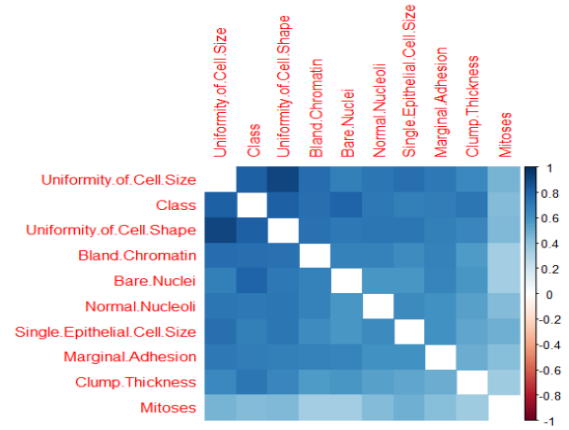

Figure 5: Correlation Plot of the Variables related to Breast Cancer

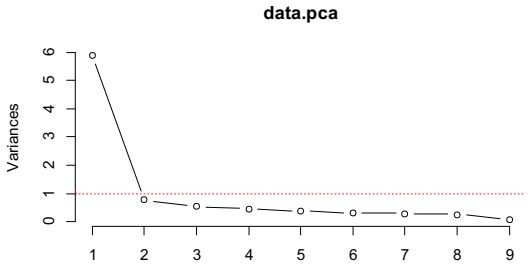

Figure 6: The Scree Plot with a Red Dashed Line

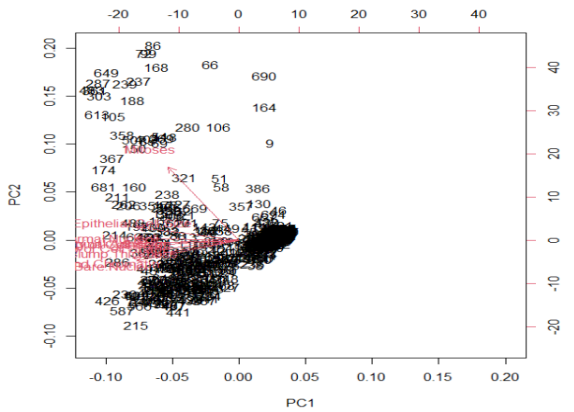

Figure 7: The Biplot of Rows and Columns Variables

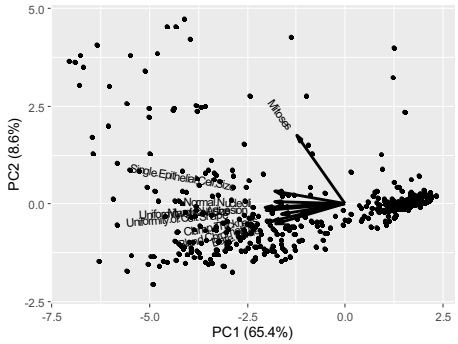

Figure 8: The Biplot of Variables related to Breast Cancer with respect to PCs

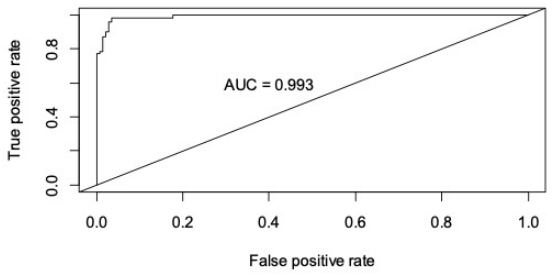

Figure 9: AUC of Disriminant Analysis for Classification of Breast Cancer

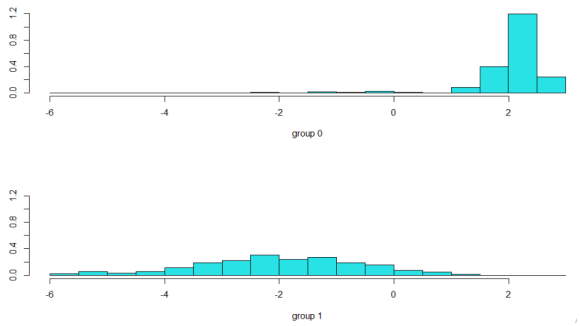

Figure 10: The Plot of Linear Discriminant Analysis for Classifying the Breast Cancer Disease

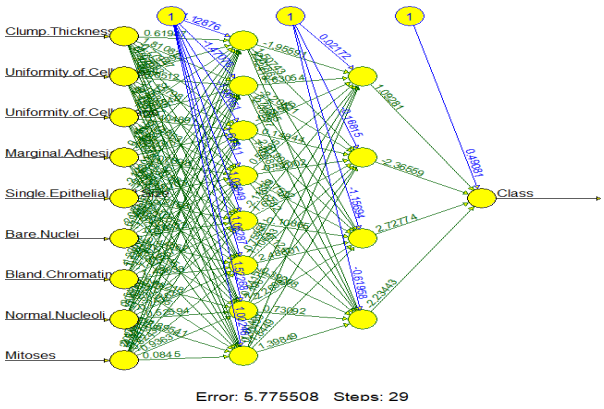

Figure 11: ANN with Two-Hidden Layer Model to Diagnose Breast Cancer of the patients

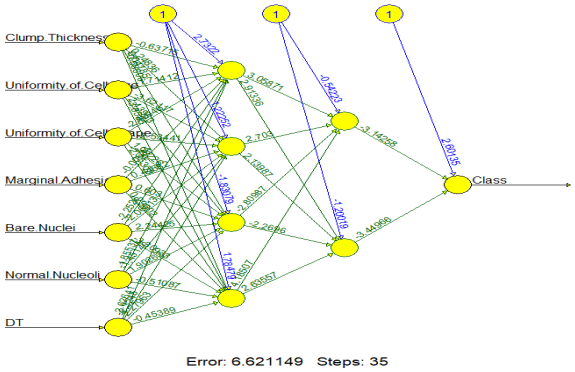

Figure 12: DT-ANN with Two-Hidden Layers Model to Diagnose Breast Cancer of the Patients

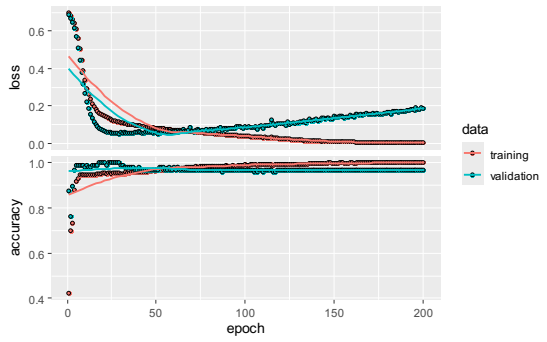

Figure 13: Second DNN Model to Diagnose Breast Cancer of the Patients

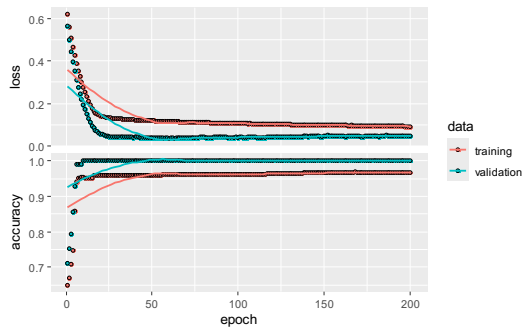

Figure 14: Fifth DT-DNN Model to Diagnose Breast Cancer of the Patients

5. Conclusion

This study highlights the effectiveness of machine learning and deep learning procedures in breast cancer analysis and classification. Traditional statistical methods, though widely used, have limitations in handling large, complex datasets and identifying non-linear relationships. In contrast, AI-based models demonstrate superior accuracy in distinguishing between benign and malignant tumors, making them invaluable tools for early detection. Among the various algorithms assessed, deep learning-based models, particularly hybrid approaches, showed remarkable performance in breast cancer classification. The study also emphasizes the importance of recognizing key risk factors for breast malignancy, aiding in preventive healthcare strategies. Forthcoming research should emphasis on refining these models, integrating genetic and imaging data, and addressing challenges such as data imbalance and interpretability to further enhance breast cancer diagnostics.This study presents our experimentation with different AI-driven models that we use to classify breast cancer. Among the different AI models, the TabNet deep learning-based prediction and classification for diagnosing the breast cancer offers a promising approach of the patients. The system increases prediction accuracy and operational efficacy. We use 699 patients'records using random split of "70/30" for train/test using "RMSE" for performance in "DT" model. From the decision tree study, it is resolved that, only 6 risk factors namely "uniformity of cell size", "uniformity of cell shape", "clump thickness", "normal nucleoli", "bare nuclei"and "marginal adhesion", was sufficient to diagnose breast cancer. Effectiveness of finding of this disease is tested by "decision tree", "PCA", "LDA", "ANNs", "hybrid DT-ANN", "DNNs", "hybrid DT-DNN" and "TabNet". The "TabNet" deep learning model has revealedpremium results for this disease diagnosis with accuracy 0.99863 and 0.99057 with train and test data set, respectively which outperforms other models. Results showed improved accuracy in prediction and classification of the diagnosis of the breast cancer compared to all other models. The strategic diagnosis of breast cancer disease ensured that the model received comprehensive data, enhancing its predictive capabilities. Challenges included managing communication overhead and ensuring constant model apprises. Limitations like

data variability for feature identification were addressed by adaptive algorithms and routine system inspections.

Ethics approval and consent to participate: 'Not applicable'

a. **Consent for publication:** All authors read and approved the final manuscript for publication.

b. Availability of data and materials: "https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)"

c. **Funding:** This research is not supported by any funding agency.

d. **Competing Interests Declarations:** The authors declare that they have no competing interests in this study.

e. Authors' Contributions: "Monalisha Pattnaik" framed concepts, analysis and results. "Deepti Rani Pattanaik" explained the interpretation of the results. "Sudev Kumar Padhi" explained TabNet and analysed this model. "Susmita Smrutirekha" wrote the Introduction and framed all the tables. "Aryan Pattnaik" wrote the conclusion and references. "Alipsa Pattnaik" update the text and interprete some results.

References

1. Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., & Silva, D. C. (2016). Predicting Breast Cancer Recurrence Using Machine Learning Techniques. ACM Computing Surveys, 49(3), 1–40. https://doi.org/10.1145/2988544
2. Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., Koren, G., Goldschmidt, Y., Shalev, V., Rosen-Zvi, M. and Guindy, M. (2019). Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms. Radiology, 292(2), pp.331–342. https://doi.org/10.1148/radiol.2019182622
3. Alloqmani, A., Abushark, Y.B. and Khan, A.I. (2023). Anomaly Detection of Breast Cancer Using Deep Learning. Arabian Journal for Science and Engineering, [online] pp.1–26. https://doi.org/10.1007/s13369-023-07945-z
4. Arik, S. O., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. The Association for the Advancement of Artificial Intelligence.
5. Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia Computer Science, 83, 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224
6. Dianati-Nasab, M., Salimifard, K., Mohammadi, R., Saadatmand, S., Fararouei, M., Hosseini, K. S., Jiavid-Sharifi, B., Chaussalet, T., &Dehdar, S. (2024b). Machine learning algorithms to uncover risk factors of breast cancer: insights from a large case-control study. Frontiers in Oncology, 13. https://doi.org/10.3389/fonc.2023.1276232
7. Hakkoum, H., Idri, A., &Abnane, I. (2021). Assessing and comparing interpretability techniques for artificial neural networks breast cancer classification. Computer Methods in Biomechanics and Biomedical Engineering Imaging & Visualization, 9(6), 587–599. https://doi.org/10.1080/21681163.2021.1901784
8. Iqbal, M.S., Ahmad, W., Alizadehsani, R., Hussain, S. and Rehman, R. (2022). Breast Cancer Dataset, Classification and Detection Using Deep Learning. Healthcare, [online] 10(12), p.2395. https://doi.org/10.3390/healthcare10122395
9. Karthik, S., Srinivasa Perumal, R., & Chandra Mouli, P. V. S. S. R. (2018). Breast Cancer Classification Using Deep Neural Networks. Knowledge Computing and Its Applications, 227–241. https://doi.org/10.1007/978-981-10-6680-1_12
10. Kaymak, S., Helwan, A., & Uzun, D. (2017). Breast cancer image classification using artificial neural networks. Procedia Computer Science, 120, 126–131. https://doi.org/10.1016/j.procs.2017.11.219
11. Khan, M. H., Boodoo-Jahangeer, N., Dullull, W., Nathire, S., Gao, X., Sinha, G. R., &Nagwanshi, K. K. (2021). Multi- class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN). PLoS ONE, 16(8), e0256500. https://doi.org/10.1371/journal.pone.0256500
12. Nawaz, M., A, A., & Hassan, T. (2018). Multi-Class Breast Cancer Classification using Deep Learning Convolutional Neural Network. International Journal of Advanced Computer Science and Applications, 9(6). https://doi.org/10.14569/ijacsa.2018.090645
13. Rautela, K., Kumar, D., & Kumar, V. (2022). A Systematic Review on Breast Cancer Detection Using Deep Learning Techniques. Archives of

Computational Methods in Engineering. https://doi.org/10.1007/s11831-022-09744-5

14. Saleh, H., Abd-elghany, S.F., Alyami, H. and Alosaimi, W. (2022). Predicting Breast Cancer Based on Optimized Deep Learning Approach. Computational Intelligence and Neuroscience, [online] 2022, p.e1820777.https://doi.org/10.1155/2022/1820777
15. Sharma, S., Aggarwal, A., & Choudhury, T. (2018). Breast Cancer Detection Using Machine Learning Algorithms. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). https://doi.org/10.1109/ctems.2018.8769187
16. Tarawneh, A. H., Faris, H., Al-Zoubi, A. M., Alqatawna, J., & Jararweh, Y. (2022). Deep learning models for breast cancer detection and diagnosis: A comparative study. International Journal of Advanced Computer Science and Applications, 13(4), 478–485. https://doi.org/10.14569/IJACSA.2022.0130478
17. Wang, X., Ahmad, I., Javeed, D., Zaidi, S., Alotaibi, F., Ghoneim, M., Daradkeh, Y., Asghar, J. and Eldin, E. (2022). Intelligent Hybrid Deep Learning Model for Breast Cancer Detection. Electronics, 11(17), p.2767. https://doi.org/10.3390/electronics11172767
18. Wu, J., & Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. Journal of Personalized Medicine, 11(2), 61. https://doi.org/10.3390/jpm11020061
19. Yadavendra, & Chand, S. (2020). A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. Machine Vision and Applications, 31(6). https://doi.org/10.1007/s00138-020-01094-1
20. Yala, A., Lehman, C., Schuster, T., Portnoi, T., & Barzilay, R. (2019). A deep learning mammography-based model for improved breast cancer risk prediction. Radiology, 292(1), 60–66. https://doi.org/10.1148/radiol.2019182716
21. Zain, Z. M., Alshenaifi, M., Aljaloud, A., Albednah, T., Alghanim, R., Alqifari, A., & Alqahtani, A. (2020). Predicting breast cancer recurrence using principal component analysis as feature extraction: an unbiased comparative analysis. International Journal of Advances in Intelligent Informatics, 6(3), 313. https://doi.org/10.26555/ijain.v6i3.462