Performance Analysis of Machine Learning Techniques used for the Prediction of Breast Cancer

B. Lavanya[1] and I. Dilshad Banu[2]
[1]*Associate Professor, Department of Computer Science, University of Madras, Chennai, India.*
[2]*Mphil (Computer Science), Department of Computer Science, University of Madras, Chennai, India*

***Abstract*- Breast cancer is one of the most common malignancies in women, although it is extremely rare in men. According to the World Health Organization (WHO), cancer is characterized by the uncontrolled aberrant proliferation of cells in any organ or tissue of the human body. The probability of saving a patient with breast cancer largely depends on its detection and initiation of treatment, which leads to high survival rates and low treatment costs. Therefore, an accurate cancer prediction model that can diagnose breast cancer at an early stage is needed. Since machine learning algorithms can be used to most accurately predict breast cancer. This study presents an overview of the capabilities of many machinelearning algorithms, including Support Vector Machine (SVM), Light Gradient Boosting (LGB), Decision tree (C4.5), Naive Bayes, Random Forest, and K-Means algorithms with techniques selection of elements and without them. We used publicly available breast cancer datasets to test several approaches for autonomous tumour categories. After obtaining the results, a performance evaluation and comparison are carried out between these different classifiers. Thus, the best classification accuracy is achieved by the proposed Support vector machine algorithm, which has a maximum accuracy of 0.98% according to the experimental results.**

Keywords: **Breast cancer; Prediction; Treatment; Early diagnosis; Support vector machine**

I. INTRODUCTION

In recent years, breast cancer has surpassed prostate cancer as the most frequent malignancy in girls. Breast cell development is the first step in the treatment of breast cancer. A lump or bulk is formed when healthy cells divide and spread more slowly than unhealthy cells do. At some point in your body, these cells can go from your breast to your lymph nodes and elsewhere. Breast cancer in younger girls, especially those under thirty, is not uncommon. It demands to be resolved immediately. Organizations have been established for both systemic and local treatment of breast cancer. Radiation and surgery are specialized treatments, whereas chemotherapy and hormone therapy are systemic. These kinds of therapies are applied widely to achieve outstanding results.

During, In the last few decades, persistent research has been performed on this cancer disease. Despite the fact that breast cancer is second among the causes of death for women. The typical signs of breast cancer include protuberances in the breast, shape changes, dimpling of the skin, and flaky regions. Changes in eating habits, a lack of real wellness, chemical replacement therapy throughout the menopausal stage, and even genetic factors are some of the important causes of these manifestations. Therefore, to lower the rate of death, early cancer identification is crucial. For this, we have many different machine learning algorithms being used to predict breast cancer. One of the prime tasks is determining the most effective and appropriate algorithm for breast cancer prediction.According to Pei Liu et al., [1] The best survival assessment, which is used to project breast cancer improvement, is improved by XG. The XG boost in system learning and the CPH model in survival analysis served as the foundation for the suggested EXSA technique. The model continued to run for 5 and 10 years, finishing with C-indices of 0 and AUCs of 0 and 0.78155. According to the results of the trials, the EXSA prognostic version shows excessive discriminative functionality in forecasting the likelihood that a breast cancer condition will progress.

Also, Farhad Imani et al., [2] By examining the effects of several factors, such as age, marital status, and histological grade, on the recurrence of breast cancer using an ensemble random survival forest (RSF) approach. The RSF approach assists in the estimation and prediction of the survival function by sampling and bootstrapping into large amounts of data using an ensemble model of survival trees. The results show that breast cancer recurrence rates

range from 2 to 6, with an average of 7.And in line with Yash Mate et al., [3] critical qualities are identified by the feature choosing processes including Pearson's coefficient, Chi-square test, Logistic regression, Random Forest, RFE, and light gradient boosting. In conclusion, the Ada boost Classifier achieved a maximum accuracy of 97% with feature selection, but the Tree Classifier had a maximum accuracy of 96.2% with characteristic selection by carefully considering 20 important characteristics for breast cancer prediction.As per Devender Kaushik et al., [4] It has been suggested to employ a distributed support vector machine model that is communication-efficient in a novel way to forecast the survival of breast cancer patients after surgery. The results of this experiment highlight the SVM CoCoA algorithm's excellent performance as well as the vast variety of applications it has, particularly in the domains of radiology and medical sciences.

The goal of the paper"A comparative study onmachine learning techniques used for the prediction of Breast Cancer" is to mainly shows a comparison between the performance of five classifiers: Support Vector Machine (SVM), Light Gradient Boosting (LGB), Decision tree (C4.5), Naive Bayes, Random Forest, and K-Means algorithms. Our objective is to predict and diagnose breast cancer, using machine-learning algorithms.with accumulated data from a variety of studies linked to breast cancer. The major goal was to find out the most effective algorithm based on the performance of each classifier in terms of, accuracy, precision and sensitivity.

II.RELATED WORKS

Machine learning techniques are widely used in the detection and prediction of breast cancer. Support Vector Machine (SVM), Light Gradient Boosting (LGB), Decision tree (C4.5), Naive Bayes, Random Forest, and K-Means algorithms are a few examples of machine learning methods. Several datasets have been used by numerous researchers to conduct breast cancer research.

In the scientific literature, several methods for breast cancer prediction have been described and put into practice.[7] In this study, Histopathology statistics is used to extend the tool and done categorization using SVM and CNN styles, achieving the highest level of accuracy for the test ratio of roughly 60:40.[8] In this study, we present a machine learning strategy for creating a reliable breast tumour classifier. We analysed five classification algorithms on the Wisconsin Breast Cancer datasets such as decision tree, support vector machine, CNN, and logistic regression (WBCD). We implemented k-fold cross-validation to confirm the accuracy of the predictions and to help with fine-tuning the model hyper-parameters. The classification of benign and malignant tumours by the four algorithms worked well. In contrast, the CNN model has a maximum accuracy of 98 %. The ensemble model's average accuracy rose from 94% for the base models to 96% on average.

According to Sunanda Das et al.,[10] The goal of this work was to accurately diagnose Breast Cancer. We offer an Ensemble approach for improved performance, and our proposed system has an accuracy of 99.28 %, which is obtained by combining Ensemble voting for improved accuracy. Ensemble voting improves the stability of the system. As a result, the proposed system is more resistant to unforeseen circumstances. The ensemble-based solution is more practical, and it will give breast cancer patients with better treatment and a more exact diagnosis. [13]In this article, a Deep Neural Network has been integrated in order to evaluate the results with different categorization techniques. We specifically compared and discussed ROC curve measurements and accuracy. The study results were analysed using the data tables and graphs. As a result, Deep Neural Networks performed brilliantly in this inquiry and also produced improved outcomes in studies using images. [15]This study described the application of deep learning techniques to computer-assisted breast cancer diagnosis. Deep learning is a method of artificial intelligence where a machine tries to mimic the cognitive function of the human brain. The Wisconsin breast cancer dataset was used for the UCI analysis. The model was enhanced by early pausing and dropouts in order to prevent overfitting. According to the neural network model, the benign class had an F1 score of 98, whereas the malignant class had an F1 score of 99. The goal of this computer-assisted diagnosis approach is to enhance rather than replace the professional medical practitioners' and doctors' diagnostic abilities. Our major aim of the research focuses on evaluating these machine learning techniques to determine the most effective techniques for the early detection and diagnosis of breast cancer.

III.METHODOLOGY

The significant objective of our research is to identify an effective technique for the prediction of

breast cancer using machine learning algorithms. Therefore, we applied machine learning classifiers such as Support Vector Machine (SVM), Light Gradient Boosting (LGB), Decision tree (C4.5), Naive Bayes, Random Forest, and K-Means algorithms to the Wisconsin Breast Cancer Diagnostic datasets and evaluate the results obtained to define which model provides higher accuracy.

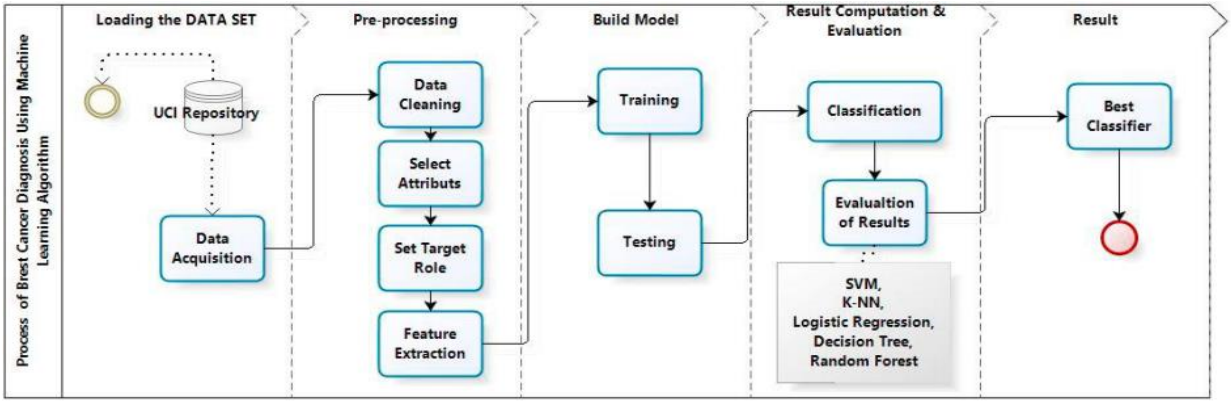

Fig. 1. Process Flow Diagram

Our approach begins with data collection and moves on to four stages of pre-processing, including data cleansing, attribute selection, target role set up and feature extraction. Machine learning methods that can be applied to predict breast cancer for a different set of measures are developed using the prepared data. We present the model with brand-new data that we have labels for to evaluate the efficacy of the strategies. This is accomplished by using the Train_Test_Split method to divide the labelled data. The training data, also known as the training set, comprises 75% of the data that we utilised to create our machine learning model. Test data or test set refers to the 25% of the data that will be used to gauge how well the model performs. The algorithms that provide the highest accuracy and are the most predictive for the detection of breast cancer are chosen after the models have been tested, and the results are compared.

TECHNIQUES USED FOR THE PREDICTION OF BREAST CANCER:

MACHINE LEARNING:

Machine Learning algorithms are derived as either supervised or unsupervised. For supervised learning algorithms, labelling provides both input data and anticipated output data. Unsupervised algorithms work with unlabelled, unclassified data. For instance, an unsupervised algorithm could group the unsorted data based on similarities and differences.

Semi-supervised algorithms are used in several machine learning approaches, including transfer learning and active learning. Transfer learning uses the knowledge gained from completing one task to assist in solving a different but related problem, whereas active learning permits an algorithm to ask the user or another source for further information. Both techniques are frequently used when labelled data is limited.

Based on the trial-and-error approach, reinforcement learning is used to identify solutions and strategies for complicated issues. One of the three methods of machine learning is also explained. In contrast to traditional learning methods, the agent doesn't need data to be trained (the learning system). AlphaGo, a game developed by Google, is a well-known example of reinforcement learning in action.

Our project uses the following machine learning algorithms:

3.1 SUPPORT VECTOR MACHINE:

Support Vector Machine (SVM) is a supervised learning method applied to regression and classification. SVM is a well-known technique widely used in the detection of cancer. Typically, this controlled classification system is used for cancer early detection and prognosis. Principally, SVM's main purpose is to classify the results by outputting data between input vectors and mapping it to a large viewpoint space. As a result, by classifying the dataset, the primary goal of SVM is to identify the best hyperplane.

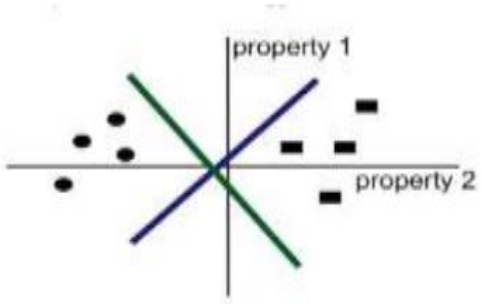

Fig.2 SVM generated hyper-planes.

3.2*LIGHT GRADIENT BOOSTING*

A light gradient boosting machine (LGBM) is a framework and a variant of gradient boosting. Like another gradient boosting Light GBM is also based on Decision tree algorithms. It is the most useful and faster algorithm in Data science. LGBM selects the leaf which produces the least error and maximum efficiency. LGBM deals with a large amount of data and consumes only less amount of memory. This method is way more helpful in reducing the error percentage. In short, it grows leaf-wise while others expand level-wise.

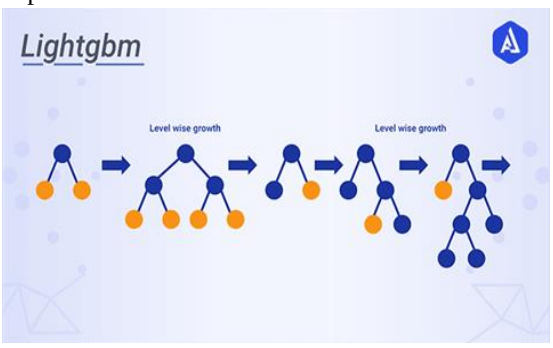

Fig 3. LGB Model

3.3 DECISION TREE:

Decision Tree C4.5 is a predictive modelling tool that can be applied across many areas. It can be built using an algorithmic method that can divide the dataset in many ways depending on various circumstances. It evaluates every potential result of a choice and follows each node to its conclusion. In each specific problem, the decision, and result are assigned a specific value (s).

3.4 NAIVE BAYES:

The Naive Bayes algorithm is frequently used to resolve classification issues. The Naive Bayes algorithm is a supervised learning technique which is based on Bayes' theorem. Large training sets are primarily utilised in text categorization. It ignores the unnecessary properties of the supplied datasets in order to forecast the outcome.

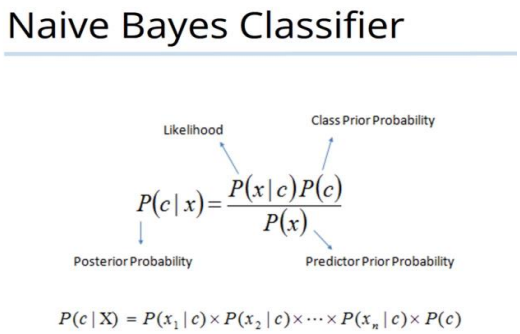

Fig 4. Naïve Bayes Classifier

3.5 RANDOM FOREST:

Random Forest is a versatile and user-friendly machine learning technique. Random forests, also known as random decision forests, are ensemble methods for classification, regression, and other tasks that function by building a large number of decision trees during the training phase and then producing the class that represents the mean of the classes (for classification) or mean prediction (for regression) of the individual trees. Most of the time this technique yields excellent results. It is an algorithm which is frequently used because of its adaptability and simplicity. For both classification and regression issues, a random forest approach can be used.

3.6 K-MEANS:

K- Means clustering is a type of unsupervised learning algorithm. It is utilised when the data, also known as unlabelled data, is not defined in groups or categories. This clustering algorithm's goal is to search for and identify groups in the data, where variable K stands in for the number of groups.

IV. DATASET ACQUISITION

In our study, we use three types of Wisconsin Breast Cancer Diagnostic datasets**.**

The first dataset has 569 instances), 2 classes (benign[0] and malignant[1]), and 33 integer-valued attributes (diagnosis, radius_mean, texture_mean, perimeter_mean**,** area_mean, smoothness_mean, compactness_mean, concavity_mean, concavepoints_mean, symmetry_mean, fractional_dimensional_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst)

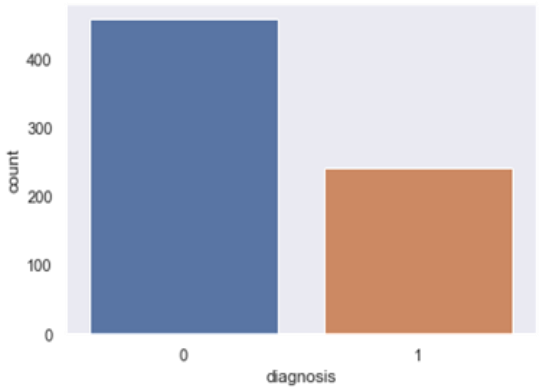

Fig.5 WISCONSIN BREAST CANCER DIAGNOSTIC DATASETS.

The second dataset has 699 instances,2 classes (benign [B] and malignant [M]),and 11 integer-valued attributes (Id, Diagnosis, clump_thickness, size_uniformity, shape_uniformity, marginal_adhesion, epithelial_size, bare_nucleoli, bland_chromatin, normal_nucleoli, mitoses)

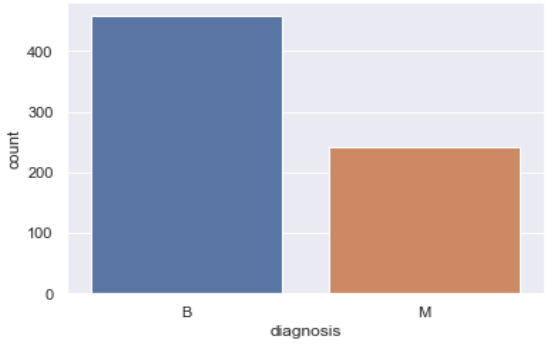

Fig.6 WISCONSIN BREAST CANCER DIAGNOSTIC DATASETS.

The third dataset has 117 instances,2 classes (Cancerous [0], Non-Cancerous [1]]), and 10 integer-valued attributes (Age, BMI, Glocose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.1, Classification)

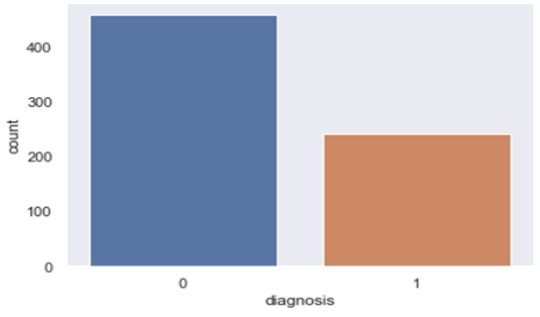

Fig.7 WISCONSIN BREAST CANCER DIAGNOSTIC DATASETS

4.2 EXPERIMENT ENVIRONMENT:

All experiments with machine learning algorithms described in this paper were performed using the Scikitlearn library and the Python programming language. For the Python programming language, Scikit-learn, sometimes referred to asScikit-learn, is a free machine learning package. It includes various Support Vector Machines, Light Gradient Boosting (LGB), Decision Tree (C4.5), Naive Bayes, Random Forest and K-Means algorithms for classification, regression and clustering. It is also built to work with the scientific and numerical Python libraries NumPy and SciPy.

V.RESULTS AND DISCUSSIONS

By applying machine learning techniques to three datasets of the Wisconsin Breast Cancer Diagnostic dataset, We used accuracy, precision, sensitivity, F1 score, and AUC as performance metrics to evaluate and compare the models to identify the best prediction algorithm for Breast Cancer. The most popular performance indicator for classification algorithms is accuracy. It is described as the proportion of all forecasts that were made that were right. The number of accurate documents retrieved by our ML model can be thought of as the accuracy employed in document retrieval. The number of positive results your ML model returned can be used to define sensitivity. The harmonic mean of accuracy and sensitivity is revealed by the F1 score. The F1 score is a weighted average of precision and sensitivity in mathematics. The table and image display the Wisconsin Breast Cancer Diagnostic data sets' accuracy percentage. We can see from the training set and test set data that each classifier has a different accuracy, but SVM consistently outperforms other classifiers with an accuracy of 0.98%.

ACCURACIES OF MACHINE LEARNING (ML) TECHNIQUES			
MACHINE LEARNING TECHNIQUE	DATA SET 1	DATA SET 2	DATA SET 3
LIGHT GRADIENT BOOSTING	0.93%	0.95%	0.56%
DECISION TREE	0.93%	0.96%	0.48
SUPPORT VECTOR MACHINE	0.98%	0.97%	0.67%
NAIVE BAYES	0.95%	0.95%	0.57%
RANDOM FOREST	0.92%	0.95%	0.56%
K-MEANS	0.94%	0.96%	0.53%

Table 1. Algorithms with their accuracies

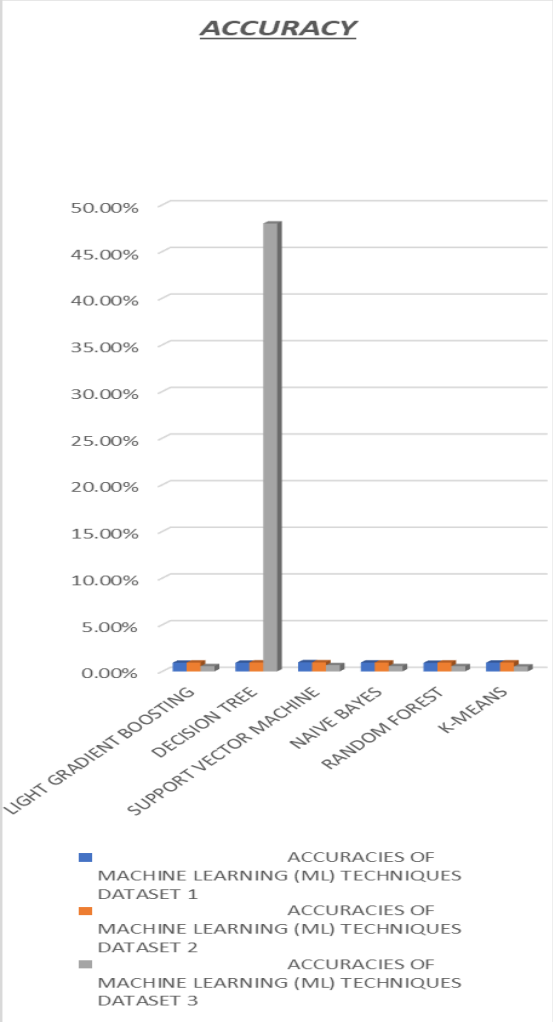

Fig 8. Comparative Graph

From the above table.1 and Fig.8, we conclude that the Support Vector Machine is better than other classificationtechniques. From the results of the table, we can see that the percentages of precision of 0.98% of SVM are higher than that of other classifiers.In terms of accuracy and precision, the Support Vector Machine has proven to be effective in the detection and prediction of breast cancer. It should be mentioned that all the results obtained are related just to the WBCD database, which can be viewed as a restriction of our work.

V.CONCLUSION AND FUTURE SCOPE

Hence, this study looks at a variety of techniques as well as a review of breast cancer diagnosis and prognosis issues.On the Wisconsin Breast Cancer Diagnostic dataset (WBCD) we applied six main techniques which are Support Vector Machine (SVM), Light Gradient Boosting (LGB), Decision tree (C4.5), Naive Bayes, Random Forest, and K-Means to compare and evaluate different results obtained based on its accuracy to identify the best machine learning algorithm that is precise, reliable and higher accuracy. All algorithms have been programmed in Python using the scikit-learn library in theAnaconda environment. After an accurate comparison between our models, we conclude that the Support Vector Machine achieved a maximum accuracy of 0.98 % over all other algorithms.

In a future study, additional clinical data on breast cancer and more follow-up data will be gathered, and it may be intended to merge neural networks with other machine learning and deep learning approaches in order to increase accuracy.

REFERENCE

[1] Pei Liu et al.,"Optimizing Survival Analysis of XGBoost for Ties to Predict Disease Progression of Breast Cancer," in IEEE Transactions on Biomedical Engineering, vol. 68, no. 1, pp. 148-160, Jan. 2021

[2] Farhad Imani et al., "Random Forest Modelling for Survival Analysis of Cancer Recurrences," 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), 2019, pp. 399-404

[3] Yash Mate et al., "Hybrid Feature Selection and Bayesian Optimization with Machine Learning for Breast Cancer Prediction," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, pp. 612-619

[4] Devender Kaushik et al., "Post-Surgical Survival Forecasting of Breast Cancer Patient: A Novel Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 37-41

[5] Abeer Saberet al., "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," in IEEE Access, vol. 9, pp. 71194-71209, 2021

[6] Yassine Amkrane et al.,"Towards Breast Cancer Response Prediction using Artificial Intelligence and Radiomics," 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), 2020, pp. 1-5

[7] AnjuYadav etal.,"Automated Detection and Classification of Breast Cancer Tumour Cells using Machine Learning and Deep Learning on

Histopathological Images," 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1-6,

[8] AsrarAlgarniet al., "Convolutional Neural Networks for Breast Tumor Classification using Structured Features," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif),2021, pp. 1-5,

[9] Shuai Liu et al., "Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall," in IEEE Access, vol. 9, pp. 24433-24445, 2021

[10] Sunanda Das et al., "Prediction of Breast Cancer Using Ensemble Learning," 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), 2019, pp. 804-808,

[11] Qasem Abu Al-Haija et al.,"Breast Cancer Diagnosis in Histopathological Images Using ResNet-50 Convolutional Neural Network," 2020 IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), 2020, pp. 1-7

[12] B. Abdikenov, et al., "Analytics of Heterogeneous Breast Cancer Data Using Neuroevolution," in IEEE Access, vol. 7, pp. 18050-18060, 2019

[13] Fabiano Teixeira et al., "An Analysis of Machine Learning Classifiers in Breast Cancer Diagnosis," 2019 XLV Latin American Computing Conference (CLEI), 2019, pp. 1-10

[14] Shailesh Kumar Verma et al., "Breast Cancer Survival Rate Prediction in Mammograms Using Machine Learning," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 169-171

[15] Sidharth S Prakash et al.,"Breast Cancer Malignancy Prediction Using Deep Learning Neural Networks," 2020 Second International Conference on Inventive Research in Computing Applications (ICIR), 2020, pp. 88-92