Department of Computer Science and Engineering

SOUTHEAST UNIVERSITY

CSE459: Research Methodology

Research Report On

Breast Cancer Classification Using Machine Learning Algorithms

---

SUBMITTED BY

Mashruful Hasan
ID: 2020000000119
Batch: 54

Md Shahadat Hossain
ID: 2020000000117
Batch: 54

Ornab Biswass
ID: 2020000000095
Batch: 54

SUPERVISED BY

Md. Mijanur Rahman
Assistant Professor
Department of Computer Science and Engineering
Southeast University

LETTER OF TRANSMITTAL

February,15, 2024,

The Chairman,
Department of Computer Science & Engineering,
Southeast University,
Tejgaon, Dhaka.

Through: Supervisor, Md. Mijanur Rahman

Subject: Submission of Research Report

Dear Sir,

It is a great satisfaction to submit our Research Report on "**Breast Cancer Classification Using Machine Learning Algorithms**" under the course Research Methodology. We propose a algorithm for classify the breast cancer in malignant and benign. It was an honor for us to work on this topic. This research has been performed by following your instructions and meeting the requirements of Southeast University.

We have prepared this report with absolute sincerity and effort. We request your approval of this research report in partial fulfillment of our degree requirement.

Thank You.

Sincerely Yours,

Name: Mashruful Hasan
ID: 2020000000119
Batch: 54
Program: B.Sc. in CSE

Md Shahadat Hossain
Undergraduate
Department of CSE
Southeast University

Name: Ornab Biswass
ID: 2020000000095
Batch: 54
Program: B.Sc. in CSE

Supervisor:
Md. Mijanur Rahman
Assistant Professor
Department of CSE
Southeast University

CERTIFICATE

This is to certify that the research report on "**Breast Cancer Classification Using Machine Learning Algorithms**" has been submitted to the respected member of the board of examiner of the School of Science and Engineering in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science & Engineeringby the following students and has been accepted as satisfactory.

This Paper has been carried out under my guidance.

Authors-

Mashruful Hasan
Undergraduate
Department of CSE
Southeast University

Md Shahadat Hossain
Undergraduate
Department of CSE
Southeast University

Ornab Biswass
Undergraduate
Department of CSE
Southeast University

Supervisor-

Md. Mijanur Rahman
Assistant Professor
Department of CSE
Southeast University

ACKNOWLEDGEMENT

Foremost, Thanks to Almighty Allah for whose given strength makes us able to complete our research successfully.

After that, we would like to thank our honorable supervisor, **Md. Mijanur Rahman**, Assistant Professor, Department of Computer Science and Engineering at Southeast University. This research would not have been possible without the exceptional support of our supervisor. His enthusiasm, knowledge, and exacting attention to detail have been an inspiration and kept us on track from our first encounter with machine learning to the final draft of this paper.

We would also like to thank **Shahriar Manzoor**, Associate Professor & Chairman, Department of Computer Science and Engineering, Southeast University, for making Md. Mijanur Rahman, our supervisor.

We also want to thank the whole faculties of the Department of Computer Science and Engineering, Southeast University, for their encouragement and motivation. Finally, with pleasure and appreciation, we acknowledge our contributions and day-after-day hard work with proper responsibility.

In addition, we owe a sincere debt of gratitude to each and every one of our great group members for their thought-provoking debates, unwavering commitment, and endless sleepless hours spent together, which were essential to finishing this project ahead of schedule. Thanks to their unflagging encouragement and assistance, we finished this report within the allotted time.

Finally, we want to say that all of those kindnesses were indispensable to complete our research.

ABSTRACT

Breast cancer, ranking as the second most devastating cancer globally, significantly impacts women. Within the healthcare sector, the integration of Machine Learning (ML) has become increasingly prevalent, particularly in predicting fatal diseases. This study focuses on harnessing ML techniques to create a predictive model using the Wisconsin Diagnostic Breast Cancer Dataset (WDBC). The primary goal is to evaluate the efficacy of five prominent ML classifiers—Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbor (KNN), and Naive Bayes (NB)—in comparison to traditional methods. By employing hyperparameter tuning through grid search methodology, the study aims to optimize model performance metrics such as accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. Through rigorous tuning, the model achieves a notable improvement in accuracy from 94.15% to 98.83%, while the conventional method also sees an enhancement from 93.56% to 97.08%. Notably, KNN emerges as the top-performing classifier, attaining an accuracy score of 98.83%. This study highlights the effectiveness of KNN when coupled with hyperparameter tuning, suggesting its potential as a valuable tool in breast cancer prognosis. These findings emphasize the significance of ML techniques in enhancing prognostic accuracy and ultimately improving outcomes for breast cancer patients. Further exploration in this area promises to advance early detection and treatment strategies, thereby mitigating the impact of this devastating disease.

TABLE OF CONTENTS

LIST OF FIGURES & TABLES

Figures

Tables

CHAPTER 1

INTRODUCTION

Cancer refers to any one of a large number of diseases characterized by the development of abnormal cells that divide uncontrollably and have the ability to infiltrate and destroy normal body tissue. Breast cancer is the second most common cancer after lung cancer and one of the main causes of death worldwide. Women have a higher risk of breast cancer as compared to men. Thus, one of the early diagnosis with an accurate and reliable system is critical in breast cancer treatment [7]. Nearly 22.5 new instances of breast cancer per 100,000 females were re-ported in Bangladesh [1]. When compared to other typesof cancer, Bangladeshi women have the greatest occurrence rate between the ages of 15 and 44 (19.3 per 100,000). According to WHO data published in 2020, Bangladesh's death rate has reached 6808 or 0.95%. As per the report of the World Health Organization, 1,000,000 ladies are almost determined to have breast malignancy and half of them would die, since it's typically late when specialists identify the disease [30]. According to the World Health Organization (WHO), an early diagnosis and detection of breast cancer can enhance breast cancer survival [7].

To minimize the high number of superfluous breast biopsies, a few Computer supported frameworks have been proposed somewhat recently. Machine learning provides potentially large opportunities for computer aided diagnosis of a disease. Machine Learning gives possibly huge freedoms to computer helped finding of an illness [30]. The challenge with breast cancer lies in the absence of a definitive cure or flawless outpatient care, leaving doctors to focus on saving lives through surgeries that remove affected tissues. Detecting breast tumors early becomes pivotal in improving survival rates. Various machine learning (ML) techniques have revolutionized early detection by rapidly analyzing vast amounts of data to predict outcomes. This classification approach is widely applicable across different sectors, enhancing prediction and diagnosis accuracy. Strategies for breast cancer prediction have evolved significantly with ML, resulting in increasingly reliable predictions. The process involves data gathering, model selection, training, and testing. Roy et al. utilized the WDBC Dataset and employed ML algorithms such as logistic regression, K-nearest neighbors, support vector machines, naive Bayes, decision trees, and random forests, finding support vector machines and logistic regression to be the most effective methods for breast cancer prediction. SVM and LR have been shown to be the most accurate algorithms, with LR and SVM both scoring 98.245% accuracy [2]. Indeed, This study holds promise for leveraging novel methodologies and datasets to enhance its performance.

According to the analysis conducted by Chaurasiya et al. [3], when evaluating the accuracy of four popular ML classification models (LR, KNN, Random Forest Tree (RDT), and SVM) on the WDBC dataset, Random Forest Tree (RDT) emerged as the top performer, achieving an impressive accuracy rate of 95% among all classifiers. In order to enhance the conclusiveness of generalizations and minimize misclassification rates further, this study's limitations include its failure to incorporate additional classification algorithms and explore diverse and comprehensive datasets for analysis.

In their investigation, Kim et al. [4] introduced a user-friendly machine learning tool for predicting pathological Complete Response (pCR) in breast cancer survivors undergoing Neoadjuvant Chemotherapy (NAC). Utilizing the Two-class Bayes point machine technique, they constructed a training set incorporating clinical traits and gene expression patterns. The gene-based prediction model achieved an accuracy of 0.875 and an AUC of the ROC curve of 0.909. Conversely, in a different model lacking gene data, both the AUC of the ROC curve and accuracy stood at 0.800. However, the study's limitations include a small patient sample size and reliance solely on internal validation methods.

A review of the literature analyzing various approaches utilized by researchers [2]-[14] to predict

breast cancer using the WDBC dataset reveals a common emphasis on evaluating model performance through metrics such as accuracy rate, precision, recall, and F1 score. However, there is a pressing need for increased focus in this area to explore alternative methods, including data preprocessing techniques, with the goal of boosting accuracy rates. Given the severe impact of breast cancer and its rising prevalence, improving accuracy rates holds significant promise in facilitating early detection and intervention, ultimately leading to better outcomes for patients.

The primary objective of this study is to assess the performance of five machine learning classifiers- logistic regression, decision trees, random forest, K-nearest neighbors, and Naive Bayes-on the WDBC dataset for breast cancer prognosis. To optimize classifier performance and select appropriate parameters, we employ a strategic approach of hyperparameter tuning using grid search methodology. Recognizing that default settings may not yield optimal results for every dataset, hyperparameter tuning becomes imperative to enhance accuracy. Through this technique, we meticulously select the best parameters tailored to the characteristics of the dataset, aiming to achieve more precise and reliable predictions.

The structure of this paper encompasses several key sections. Following the introduction, a comprehensive review of related work is provided. Subsequently, the research methodology is delineated, covering aspects such as data collection, preprocessing, and an overview of the algorithms utilized. The fourth section presents the experimental findings derived from the analysis. This is followed by a synthesis of the overall conclusions drawn from the study, along with recommendations for future research directions. Finally, the paper concludes with acknowledgments and references, encapsulating the comprehensive scope of the research undertaken.

CHAPTER 2

LITERATURE REVIEW

Cancer stands as one of the most perilous and prevalent illnesses worldwide, predominantly affecting women. Among the various forms of cancer—such as breast, lung, ovarian, and brain malignancies—breast cancer stands out as the most devastating globally [15]. In this section, we offer a thematic overview of the advancements and characteristics of existing breast cancer prediction techniques. Researchers have developed numerous machine learning classification methods aimed at predicting breast cancer, reflecting the ongoing efforts to enhance early detection and treatment strategies for this formidable disease.

In their study utilizing the Wisconsin Breast Cancer (WBC) dataset for breast cancer identification and diagnosis, Bazazeh et al. [5] evaluate machine learning classifiers such as Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB). They conduct a comparative analysis of these classifiers based on key metrics including accuracy, precision, recall, and the Receiver Operating Characteristic (ROC) curve. The results indicate that Random Forest (RF) achieves the highest accuracy among the classifiers tested, with an impressive accuracy rate of 99.9%, surpassing SVM (96.6%) and NB (99.1%).

In their investigation, Chaurasiya et al. [3] examine the accuracy metrics of four widely recognized machine learning classification models (LR, KNN, Random Forest Tree (RDT), and SVM) concerning their performance on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Among these classifiers, Random Forest Tree (RDT) emerged as the top performer, achieving the highest accuracy rate of 95%.

In their study, Assegie [6] proposes a model for detecting breast cancer by enhancing the K-Nearest Neighbors (KNN) algorithm. To improve the accuracy of the model in detecting breast cancer, they employ hyperparameter tuning through grid search methodology to optimize the value of K. Through this approach, they achieve an accuracy of 94.35%, surpassing the accuracy obtained with the default hyperparameter value of 90.10% for KNN.

In their research, Nurul et al. [7] investigated the effectiveness of various machine learning techniques in predicting breast cancer survival. They utilized cross-validation procedures, including ten-fold, five-fold, three-fold, and two-fold, to optimize predictive performance across ML approaches such as K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and ensemble methods using the Wisconsin Breast Cancer Dataset (WBCD). Notably, AdaBoost ensemble methods achieved accuracy rates of 98.77% with ten-fold cross-validation, 98.41% with two-fold, and 98.24% with three-fold. SVM exhibited the lowest error rate and the highest accuracy rate of 98.60%, as determined by the results of five-fold cross-validation.

In their study, Gupta et al. [8] propose the utilization of both deep learning (specifically Adam Gradient Descent) and traditional machine learning algorithms (Decision Trees (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM)) for analyzing malignant and benign cells within the Wisconsin Breast Cancer (WBC) datasets. Deep learning, which combines the strengths of AdaGrad and RMSProp optimization techniques, is highlighted for its ability to achieve highly accurate results with minimal loss, reaching an accuracy of 98.24%. RMSProp demonstrates effectiveness in handling nonstationary signals, while AdaGrad is particularly well-suited for addressing computer vision tasks.

Ara et al. [9] aim to analyze the Wisconsin Breast Cancer (WBC) dataset and evaluate the performance of various machine learning classifiers for breast cancer prediction. Through their investigation, they assess the effectiveness of Decision Trees (DT), Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB) algorithms. Their findings reveal an overall accuracy rate of 96.5%, with Random Forest (RF) and Support Vector Machine (SVM) demonstrating superior performance compared to other classifiers.

In their study, Amrane et al. [10] present two distinct machine learning classifiers, namely Naive Bayes (NB) and K-Nearest Neighbor (KNN), applied to the Wisconsin Breast Cancer (WBC) dataset for breast cancer classification. The researchers utilize cross-validation to evaluate the performance of these classifiers and ascertain their accuracy. Interestingly, while the Naive Bayes classifier achieves an accuracy of 96.19%, the findings indicate that K-Nearest Neighbor (KNN) exhibits superior performance, boasting a higher accuracy rate of 97.51% and a lower error rate.

The findings of the comprehensive literature reviews are presented in **Table 1.** Reference numbers are listed in column 1, while column 2 displays the corresponding years of the studies. In column 3, the datasets utilized in each study are specified. The research algorithms employed by the authors are outlined in column 4. Lastly, column 5 provides an overview of the efficiency or performance of the algorithms used in each study.

Table 1. Comparison of publicly available prediction models.

Ref. No.	Period	Datasets	Algorithm	Accurateness (%)
[21]	2022	WDBC and BCCD	SVM, LR, KNN and EC	99.3%, 98.06%, 97.35%, and 97.61%
[5]	2022	WDBC	KNN, SVM, LR and Random Forest Tree(RFT)	91.25%, 92.5%, 93.75% and 95%
[11]	2022	Regional Oncology Center in Meknes, Morocco.	SVM, KNN, LR and NB	90.6%, 86.1%, 80.6% and 51.7%
[2]	2021	WDBC	LR, SVM, KNN, DT Classifier, RFClassifier and NB Classifier.	98.2%, 98.2%, 96.8%, 91.4%, 97.4% and 97.1%
[14]	2020	WDBC	LR and DT	94.4% and 95.1%
[15]	2020	(WBC) and (WDBC)	NB, SVM, KNN and LR,	92%, 96%, 97% and 99% (WBC) and 96%, 94%, 96% and 98% (WDBC)
[16]	2020	WBC	NB, LR, and Neural Networks (NN)	95% training and 93% testing and 98%training and 97% testing
[28]	2019	WDBC	DT and KNN	92% and 95.95%
[7]	2021	WBCD	KNN, RF, SVM, Ensemble Methods	AdaBoost (98.77%)
[17]	2019	WBCD	MLP, KNN, CART, Gaussian Naive Bayes (NB) and SVM	99.12%, 95.61%, 93.85% 94.73% and 98.24%
[18]	2019	WDBC	Kernel SVM, LR KNN, DT, NB and RF	98.24%, 96.49%, 95.61%, 88.59%, 85.09% and 92.98%
[10]	2018	WBC	NB and KNN	96.19% and 97.51%
[20]	2018	BCCD and WBCD	DT, SVM, RF, LR, NN DT, SVM, RF, LR, NN	68.3%, 76.3%, 78.5%, 73.7%, 74.8% (BCCD), 96.3%, 97.7%, 98.9%, 98.1%, 98.5% (WBCD)
[19]	2017	BCD	NB and KNN	96.19% and 97.51%
[4]	2016	WBC	SVM, Bayesian Networks (BN), and RF	96.6%, 99.2%, and 99.9%
[12]	2013	WDBC	K-SVM (Hybrid), ACO-SVM, GA-SVM and PSO-SVM	97.38%, 95.96%, 97.19% and 97.37%

CHAPTER 3

METHODOLOGY

To determine the malignancy status of tumors, we have devised a rigorous methodology aimed at obtaining reliable results and facilitating informed decision-making. This methodology is structured into several key components: Dataset Description, Data Collection, Data Pre-processing, and Feature Selection.

In **Figure 1**, the process begins with the compilation of the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Subsequently, the dataset undergoes a thorough examination to identify any duplicates or missing data. Given the absence of missing data, the next step involves partitioning the dataset into training (70%) and testing (30%) subsets. Following this, standard scaling is applied to ensure uniformity in feature scales. Finally, both traditional and hyper-tuned parameter algorithms are constructed to assess and compare their respective performances in tumor classification. Through this comprehensive approach, we aim to generate robust and trustworthy results that contribute to effective decision-making in clinical settings.

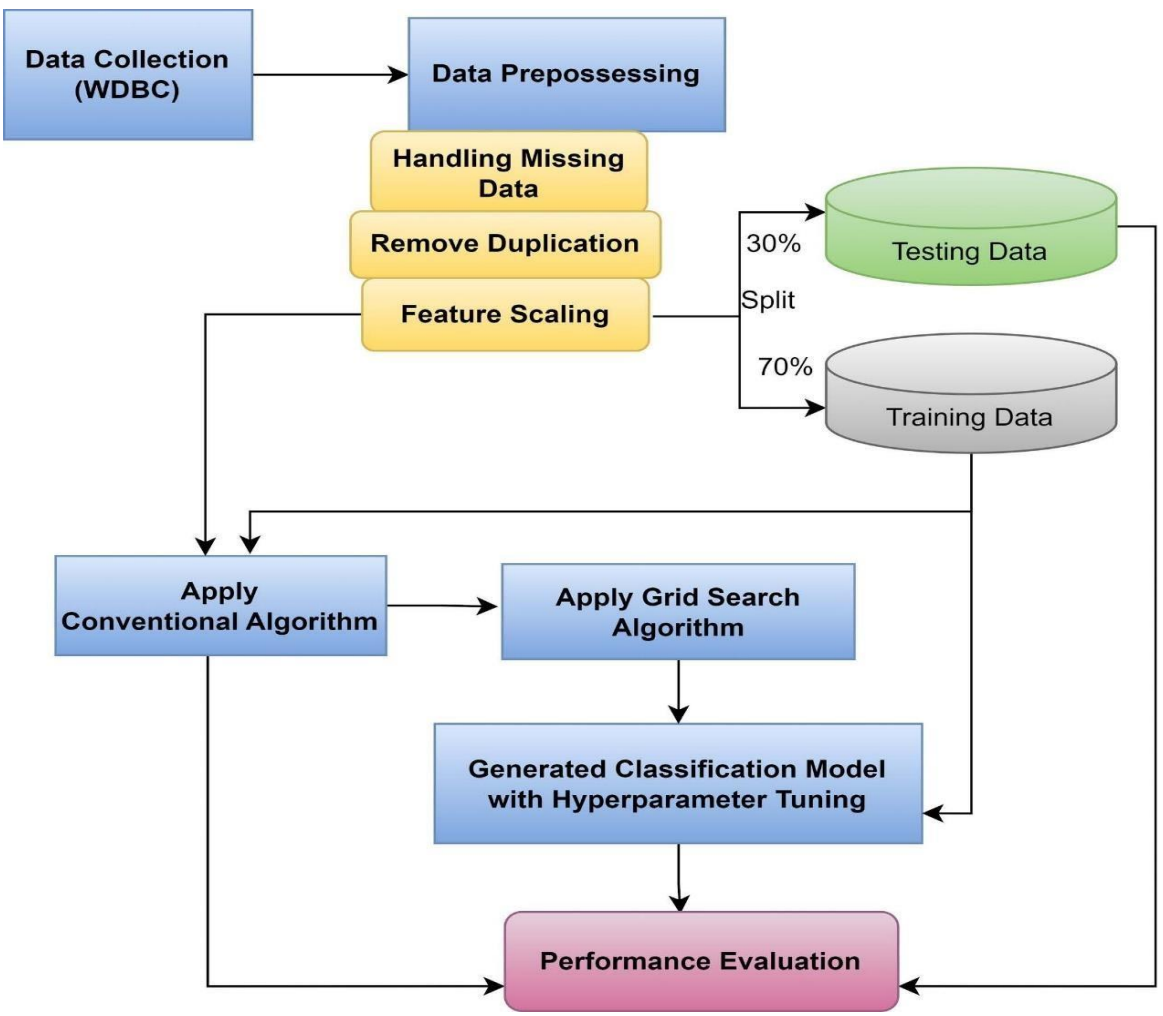

Figure 1. Model for research system.

3.1 Dataset Description

The WDBC dataset was curated by Dr. William H. Wolberg from the University of Wisconsin Hospital in Madison, Wisconsin, USA. This dataset comprises 32 columns, with the first column representing the **unique ID** of each sample, and the second column indicating the diagnosis outcome, where **0 represents benign and 1 represents malignant**. The remaining columns (3 - 32) contain measurements for 10 distinct attributes related to the size and shape characteristics of cancer cell nuclei. These attributes include **mean, standard deviation, and worst-case mean measurements, reflecting variability in nucleus qualities**. The dataset originates from biopsy samples obtained using the Fine Needle Aspiration (FNA) technique, where cells' nuclei are examined microscopically in a pathology lab to identify specific traits. All feature values are recorded with a maximum precision of 4 significant digits. Notably, no null values were observed within the dataset. Further details regarding the ten primary attributes are provided in **Table 2.**

Table 2. Description of WDBC dataset.

Feature Name	Feature Description
Radius	The average distance between the spots at the circumference's centerand edges.
Texture	Grayscale value's SD. Perimeter Gross separation exists between thesnake's points.
Perimeter	Gross separation exists at the snake's tip and between.
Area	Total amount of pixels inside the snake, plus one-half of each pixeloutside its body.
Smoothness	Measured locally by computing the length difference, the variation inradius length.

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
906564	B	14.69	13.98	98.22	656.1	0.10310
85715	M	13.17	18.66	85.98	534.6	0.11580
891670	B	12.95	16.02	83.14	513.7	0.10050
874217	M	18.31	18.58	118.60	1041.0	0.08588
905680	M	15.13	29.81	96.71	719.5	0.08320

compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
0.18360	0.14500	0.06300	0.2086	0.07406
0.12310	0.12260	0.07340	0.2128	0.06777
0.07943	0.06155	0.03370	0.1730	0.06470
0.08468	0.08169	0.05814	0.1621	0.05425
0.04605	0.04686	0.02739	0.1852	0.05294

radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se
0.5462	1.5110	4.795	49.45	0.009976	0.052440
0.2871	0.8937	1.897	24.25	0.006532	0.023360
0.2094	0.7636	1.231	17.67	0.008725	0.020030
0.2577	0.4757	1.817	28.92	0.002866	0.009181
0.4681	1.6270	3.043	45.38	0.006831	0.014270

concavity_se	concave points_se	symmetry_se	fractal_ dimension_se	radius_worst	texture_worst
0.05278	0.015800	0.02653	0.005444	16.46	18.34
0.02905	0.012150	0.01743	0.003643	15.67	27.95
0.02335	0.011320	0.02625	0.004726	13.74	19.93
0.01412	0.006719	0.01069	0.001087	21.31	26.36
0.02489	0.009087	0.03151	0.001750	17.26	36.91

perimeter_ worst	area_worst	smoothness_ worst	compactness_ worst	concavity_ worst	concave points_worst
114.10	809.2	0.1312	0.36350	0.3219	0.11080
102.80	759.4	0.1786	0.41660	0.5006	0.20880
88.81	585.4	0.1483	0.20680	0.2241	0.10560
139.20	1410.0	0.1234	0.24450	0.3538	0.15710
110.10	931.4	0.1148	0.09866	0.1547	0.06575

symmetry_worst	fractal_dimension_worst
0.2827	0.09208
0.3900	0.11790
0.3380	0.09584
0.3206	0.06938
0.3233	0.06165

3.2 Dataset Collection

The WDBC dataset, sourced from Kaggle, is employed for breast cancer prediction purposes and encompasses 569 instances, each characterized by a total of 32 features. Below is a sample excerpt of the dataset.

3.3 Data Pre-Processing

Before proceeding with the analysis, the WDBC dataset undergoes initial scrutiny to ensure data integrity. Subsequently, irrelevant features such as the ID and unnamed columns are eliminated from the dataset. As these variables do not contribute to the prediction of breast cancer, their removal aims to streamline the dataset and enhance accuracy. Following this, feature scaling is applied using standard scaling techniques to ensure uniformity in feature scales across the dataset.

3.4 Feature Selection

In the context of distinguishing between benign and malignant cells within the dataset, there are a total of **569 records, with 357 (62.7%) classified as benign and 212 (37.3%)** classified as malignant. This distribution is visually represented in **Figure 2**. Unlike employing specific feature selection techniques, we opted not to utilize any particular method in this instance. This decision was made due to the satisfactory results achieved without the need for additional feature selection strategies, such as correlation coefficient analysis. Moreover, considering the medical nature of the data, it was deemed appropriate to refrain from employing complex feature selection techniques.

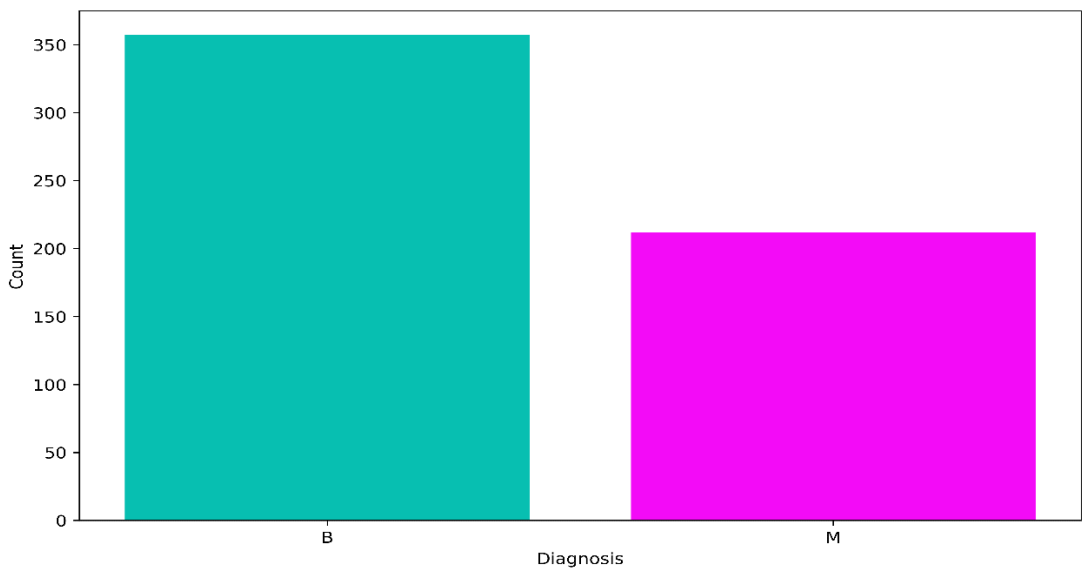

Figure 2. Benign vs malignant cells.

3.5 Algorithm Used

In this section, we explored the WDBC dataset to determine which algorithm performs best with this small dataset. In this study, five of the most well-liked ML algorithms are used, but KNN and DT performed well on small datasets while RF, NB, and LR performed well on large datasets. The paramount goal is to benchmark each approach against one another and determine the most efficient and robust technique for the WDBC dataset.

K-Nearest Neighbor (KNN): It is a straightforward classification technique commonly used in machine learning. Unlike some algorithms that learn from the dataset, KNN does not explicitly learn any underlying patterns or relationships from the data [11]. Instead, during the training phase, KNN stores the entire dataset and classifies new data points based on their similarity to neighboring data points in the feature space [24]. While KNN can be effective for smaller datasets where the underlying patterns are relatively clear, its performance may diminish when applied to larger datasets due to computational constraints and increased complexity.

Decision Tree (DT): Decision Tree is a supervised machine learning technique employed for both classification and regression tasks [25]. Its structure resembles that of a tree, with distinct entities comprising decision nodes, branches, and leaf nodes. Decision nodes utilize features of the dataset to make decisions, while branches represent decision rules. Leaf nodes correspond to the output of the decision process, typically in the form of a yes/no question and answer [2]. Decision trees are particularly effective for classification tasks with a limited number of class labels, as they can efficiently partition the feature space to classify instances into distinct categories.

Random Forest (RF): Random Forest is a powerful ensemble learning technique that improves prediction accuracy by creating multiple decision trees (DTs) on diverse subsets of the provided dataset and then averaging their predictions [26]. This approach, widely employed for classification, regression, and various other applications, effectively enhances the accuracy of predictions during training. Random Forest is particularly well-suited for handling large datasets, as it can efficiently manage high-dimensional data and mitigate overfitting by aggregating predictions from numerous trees.

Naive Bayes (NB): Naive Bayes is a widely recognized and simple classification algorithm commonly employed in predictive modeling tasks. Also referred to as a probabilistic classifier, NB is utilized for swift predictions based on the probability of a given outcome [24]. Known for its effectiveness, NB performs admirably even on large datasets, making it a versatile and powerful algorithm for various machine learning applications.

Logistic regression (LR): Logistic Regression is a machine learning technique derived from statistics, commonly employed to address classification problems [15]. Primarily suited for binary classification tasks, LR predicts a binary dependent variable by employing a logistic function. Despite its simplicity, LR demonstrates robust performance, particularly on very large datasets, making it a valuable tool for various classification applications.

CHAPTER 4

EXPERIMENTAL RESULTS

In this experimental phase, we evaluated the performance of the constructed machine learning algorithms using the test dataset, which was partitioned to contain 30% of the total dataset. The effectiveness of each method was assessed by calculating metrics such as accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC). To facilitate this evaluation, a confusion matrix (Figure 3) was generated, detailing true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) to compare actual and predicted results. The interpretation of these terms is provided below.

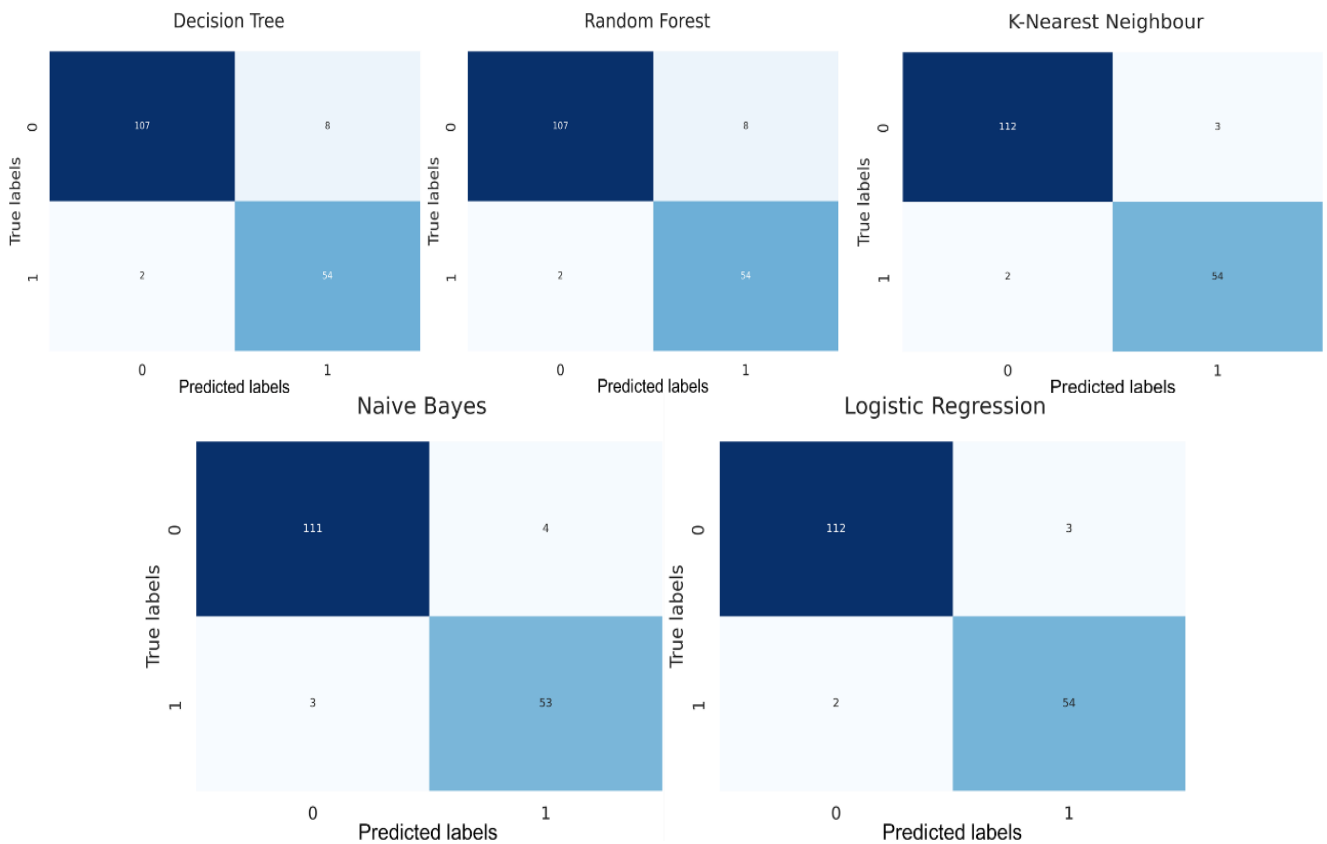

Figure 3. Confusion matrix after tuning.

TP: True Positive (Correctly Identified)
FP: False Positive (Correctly Rejected)
TN: True Negative (Incorrectly Identified)
FN: False Negative (Incorrectly Rejected)

4.1 Accuracy

Accuracy represents the overall correctness of the machine learning model's predictions. It is calculated by dividing the total number of correct predictions by the total number of instances in the dataset. It's worth mentioning that accuracy can fluctuate across different testing sets, influenced by the threshold selection of the classifier. The accuracy is determined using the formula (1).

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} * 100 \qquad (1)$$

4.2 Precision

Precision measures the effectiveness of the model in predicting a particular category accurately. It quantifies the proportion of correctly predicted positive instances among all instances predicted as positive. The mathematical expression for precision is depicted in Equation (2).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} * 100 \qquad (2)$$

4.3 Recall

Recall, also known as sensitivity, denotes the proportion of correctly predicted instances that were correctly identified or found by the model. It quantifies the model's ability to capture all relevant instances of a particular class. The mathematical expression for recall is provided in Equation (3).

$$Recall = \frac{TP}{TP+FN} * 100 \qquad (3)$$

4.4 F1 Score

The metric referred here to combines two typically contrasting variables, recall and precision, into a single measure summarizing the predictive performance of a model. This composite metric provides an overall assessment of the model's predictive capability. The mathematical expression for this metric is presented in Equation (4)

$$F1\ Score = 2 * \frac{Recall*Precision}{Recall+Precision} \qquad (4)$$

4.5 AUC & ROC Curve

The ROC curve graphically illustrates the True Positive Rate (TPR) against the False Positive Rate (FPR) for various threshold values of the model's predicted probabilities. The Area Under the Curve (AUC) is a numerical metric that quantifies the area under the ROC curve. It ranges from 0 to 1, with 0.5 indicating a random classifier and 1 representing a perfect classifier. The performance of the optimized model is depicted in **Figure 4** using the AUC and ROC curve.

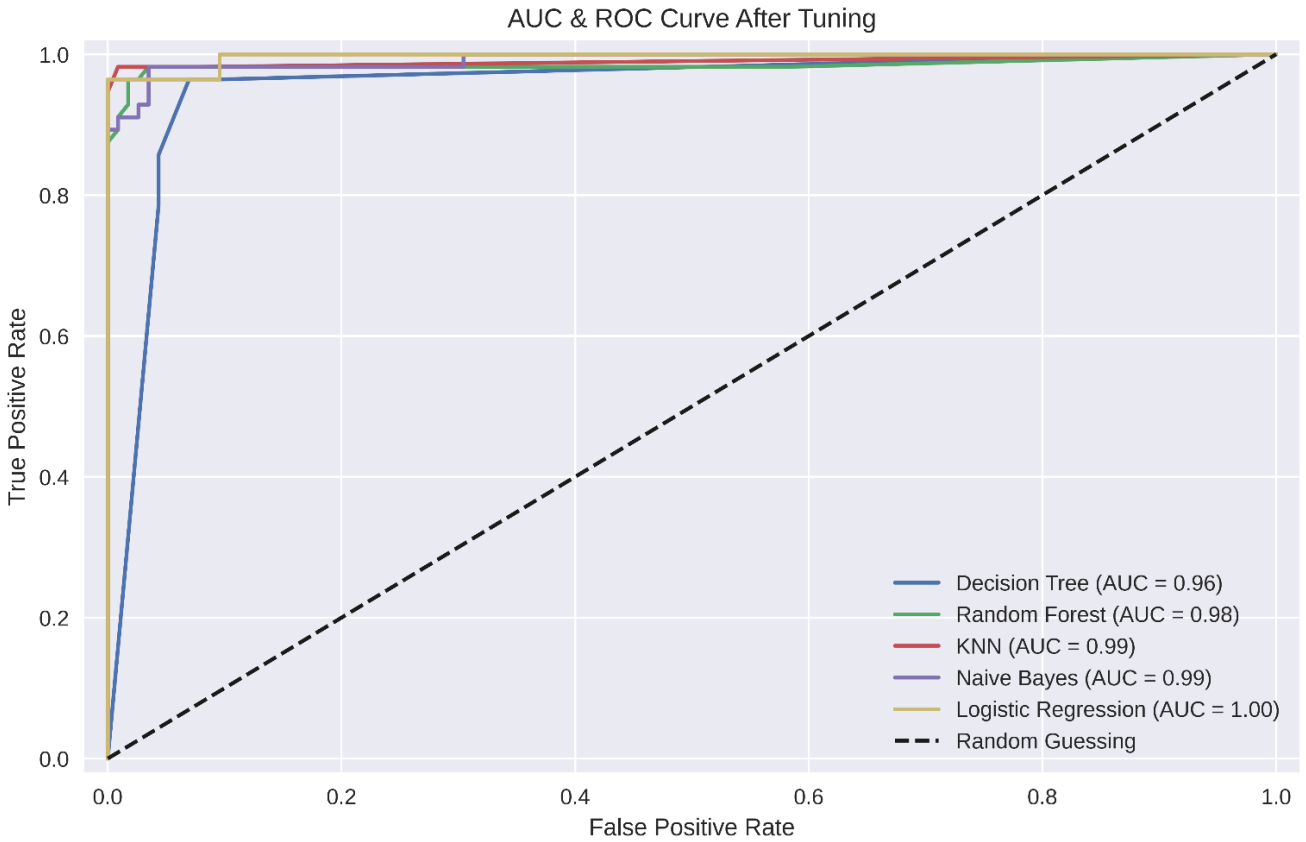

Figure 4. AUC and ROC curve after tuning.

Algorithm Names	Accuracy	Precisions	Recall	F1 Score
Decision Tree	93.56%	94%	94%	94%
Random Forest	97.08%	97%	97%	97%
K Nearest Neighbor	96.49%	96%	96%	96%
Naive Bayes	95.91%	96%	96%	96%
Logistic Regression	96.49%	96%	96%	96%

Table 3. Performance evaluation without hyperparameter tuning.

Algorithm Names	Accuracy	Precisions	Recall	F1 Score
Decision Tree	94.15%	95%	94%	94%

Random Forest	97.08%	97%	97%	97%
K Nearest Neighbor	98.83%	99%	99%	99%
Naive Bayes	95.91%	96%	96%	96%
Logistic Regression	97.08%	97%	97%	97%

Table 4. Performance evaluation with hyperparameter tuning.

The findings presented in **Table 3 and Table 4** reveal that the KNN classifier exhibits superior performance in this study following hyperparameter tuning, as evidenced by its high accuracy, precision, recall, and F1
score. The results indicate that the KNN model outperforms the other five suggested classifiers in accurately predicting breast cancer. **Figure 5** provides a visual representation for enhanced comprehension of these findings.

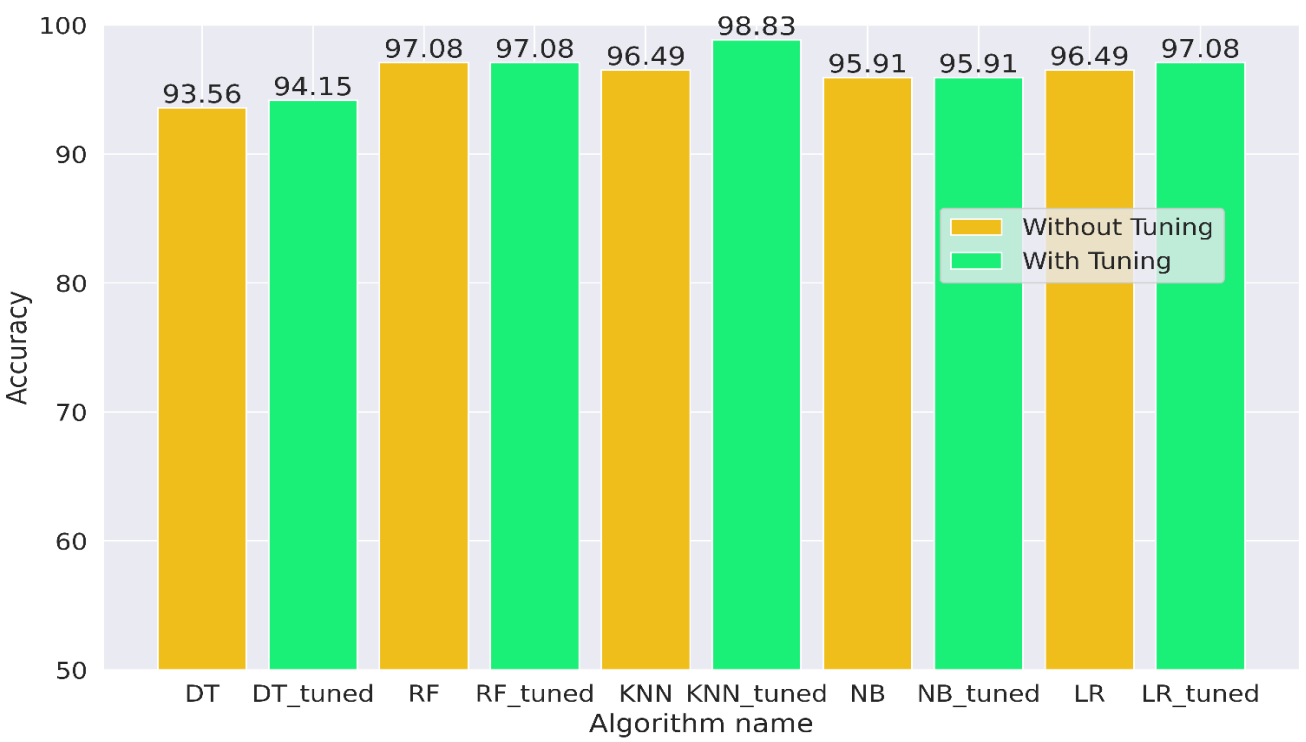

Figure 5. Result analysis on accuracy.

Table 5 contrasts the impact of our proposed model, which involves hyperparameter tuning for breast cancer prediction using the WDBC dataset exclusively, with the accuracy of the KNN classifier. Through this comparison, we conclude that our suggested method outperforms all other approaches documented in the literature. By comparing the results of KNN with those of other state-of-the-art studies listed in **Table 5**, we affirm the superior performance of our proposed approach. **Figure 6** offers a visual depiction to facilitate a clearer understanding of these

conclusions.

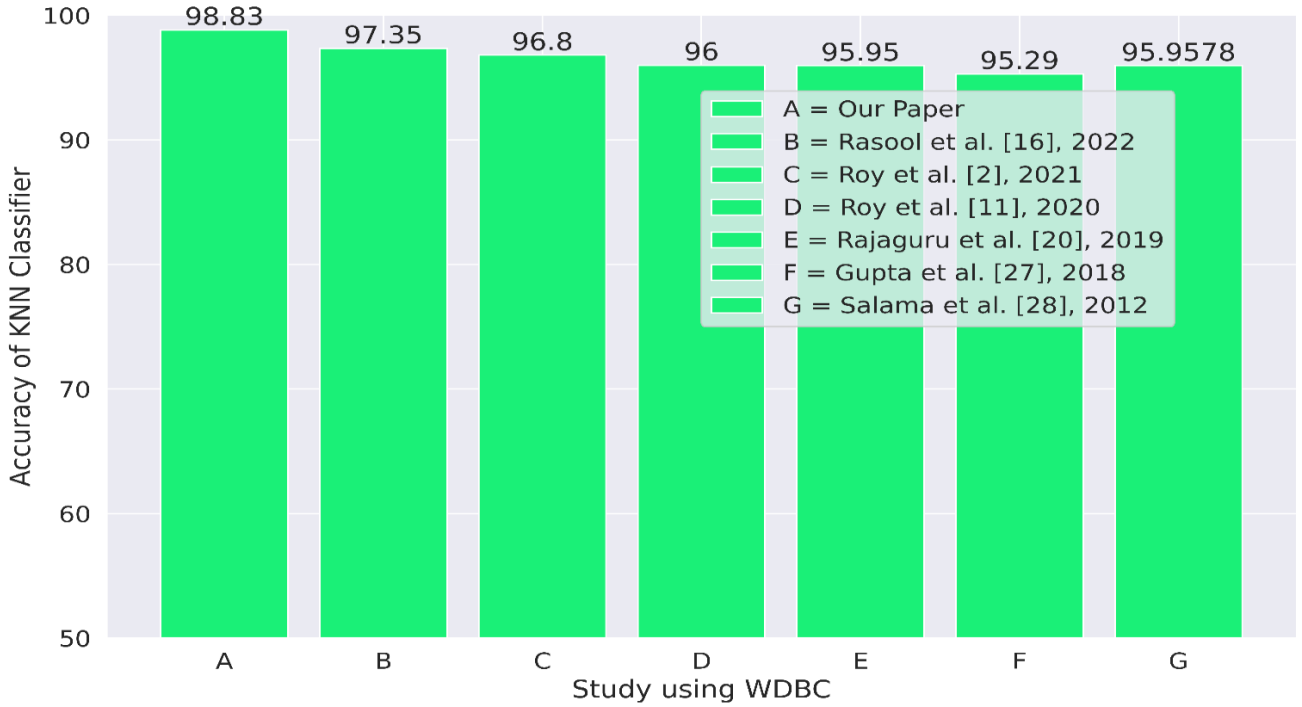

Figure 6. Result comparison with existing work.

Table 5. Result comparison with existing work.

Study using WDBC	Accuracy of KNN Classifier
Our paper	98.83%
Rasool *et al.* [16], 2022	97.35%
Roy *et al.* [2], 2021	96.8%
Roy *et al.* [11], 2020	96%
Rajaguru *et al.* [27], 2019	95.95%
Gupta *et al.* [28], 2018	95.29%
Salama *et al.* [29], 2012	95.9578% (IBK)

CHAPTER 5

CONCLUSION AND FUTURE WORK

In summary, breast cancer stands as a leading cause of mortality among women, necessitating the development of robust predictive methods. This study introduced a novel approach to forecast breast cancer by employing five distinct machine learning classifiers—LR, DT, RF, KNN, and NB—utilizing the WDBC dataset. By implementing hyperparameter tuning through grid search, we departed from conventional methods, significantly improving the accuracy rates of the classifiers. Specifically, the hyperparameter-tuned DT, RF, KNN, NB, and LR classifiers achieved accuracy rates of 94.15%, 97.08%, 98.83%, 95.91%, and 97.08%, respectively, surpassing their untuned counterparts. Notably, KNN emerged as the top-performing classifier with an accuracy of 98.83%. Looking ahead, further enhancing accuracy through expanded dataset sizes presents a promising avenue for future research. Additionally, exploring applications beyond cancer prediction, such as cancer stage detection, holds potential for advancing the field and improving patient outcomes.

CHAPTER 6

REFERENCES

[1]. Begum, S.A., Mahmud, T., Rahman, T., Zannat, J., Khatun, F., Nahar, K., Towhida, M., Joarder, M., Harun, A. and Sharmin, F. (2019) Knowledge, Attitude and Prac- tice of Bangladeshi Women towards Breast Cancer: A Cross Sectional Study. *My- mensingh Medical Journal*, **28**, 96-104. https://pubmed.ncbi.nlm.nih.gov/30755557/

[2]. Roy, S., Gawande, R., Nawghare, A. and Mistry, S. (2021) Comparative Study of Machine Learning Algorithms for Detecting Breast Cancer. *International Journal of Computer Science Trends and Technology* (*IJCST*), **9**, 103-111.

[3]. Chaurasiya, S. and Rajak, R. (2022) Comparative Analysis of Machine Learning Al- gorithms in Breast Cancer Classification. (Preprint) https://doi.org/10.21203/rs.3.rs-1772158/v1

[4]. Kim, I., Lee, K., Lee, S., Park, Y. and Lee, K. (2021) A Predictive Model for Patho- logic Complete Response in Breast Cancer Patients Treated with Neoadjuvant Che- motherapy Using Machine Learning. *Advances in Breast Cancer Research*, **10**, 141-155. https://doi.org/10.4236/abcr.2021.104012
https://www.scirp.org/journal/paperinformation.aspx?paperid=111495

[5]. Bazazeh, D. and Shubair, R. (2016) Comparative Study of Machine Learning Algo- rithms for Breast Cancer Detection and Diagnosis. *Proceedings of* 2016 5*th Interna- tional Conference on Electronic Devices, Systems and Applications* (*ICEDSA*), Ras Al Khaimah, 6-8 December 2016,1-4. https://doi.org/10.1109/ICEDSA.2016.7818560
https://ieeexplore.ieee.org/abstract/document/7818560

[6]. Assegie, T.A. (2021) An Optimized K-Nearest Neighbor Based Breast Cancer De- tection. *Journal of Robotics and Control* (*JRC*), **2**, 115-118. https://doi.org/10.18196/jrc.2363

[7]. Mashudi, N.A., Rossli, S.A., Ahmad, N. and Noor, N.M. (2021) Comparison on Some Machine Learning Techniques in Breast Cancer Classification. *Proceedings of* 2020 *IEEE-EMBS Conference on Biomedical Engineering and Sciences* (*IECBES*), Langkawi Island, 1-3 March 2021, 499-504. https://ieeexplore.ieee.org/abstract/document/9398837
https://doi.org/10.1109/IECBES48179.2021.9398837

[8]. Gupta, P. and Garg, S. (2020) Breast Cancer Prediction Using Varying Parameters of Machine Learning Models. *Procedia Computer Science*, **171**, 593-601. https://doi.org/10.1016/j.procs.2020.04.064

[9]. Ara, S., Das, A. and Dey, A. (2021) Malignant and Benign Breast Cancer Classifica- tion Using Machine Learning Algorithms. *Proceedings of* 2021 *International Con- ference on Artificial Intelligence* (*ICAI*), Islamabad, 5-7 April 2021, 97-101. https://ieeexplore.ieee.org/abstract/document/9445249
https://doi.org/10.1109/ICAI52203.2021.9445249

[10]. Amrane, M., Oukid, S., Gagaoua, I. and Ensari, T. (2018) Breast Cancer Classifica- tion Using Machine Learning. *Proceedings of* 2018 *Electric Electronics, Computer Science,*

Biomedical Engineerings' *Meeting* (*EBBT*), Istanbul, 18-19 April 2018, 1-4. https://ieeexplore.ieee.org/abstract/document/8391453 https://doi.org/10.1109/EBBT.2018.8391453

[11]. Roy, B.R., Pal, M., Das, S. and Huq, A. (2020) Comparative Study of Machine Learning Approaches on Diagnosing Breast Cancer for Two Different Dataset. *Pro- ceedings of* 2020 2*nd International Conference on Advanced Information and Com- munication Technology* (*ICAICT*), Dhaka, 28-29 November 2020, 29-34. https://ieeexplore.ieee.org/abstract/document/9333507 https://doi.org/10.1109/ICAICT51780.2020.9333507

[12]. El-Shair, Z.A., Sánchez-Pérez, L.A. and Rawashdeh, S.A. (2020) Comparative Study of Machine Learning Algorithms Using a Breast Cancer Dataset. *Proceedings of* 2020 *IEEE International Conference on Electro Information Technology* (*EIT*), Chi- cago, 31 July 2020-1 August 2020, 500-508. https://ieeexplore.ieee.org/abstract/document/9208315 https://doi.org/10.1109/EIT48999.2020.9208315

[13]. Bataineh, A.A. (2019) A Comparative Analysis of Nonlinear Machine Learning Al- gorithms for Breast Cancer Detection. *International Journal of Machine Learning and Computing*, **9**, 248-254. https://doi.org/10.18178/ijmlc.2019.9.3.794

[14]. Li, Y. and Chen, Z. (2018) Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. *Applied and Computational Mathematics*, **7**, 212-216. https://doi.org/10.11648/j.acm.20180704.15

[15]. Hashi, E.K. and Zaman, M.S.U. (2020) Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. *Journal of Applied Science & Process Engineering*, **7**, 631-647. https://publisher.unimas.my/ojs/index.php/JASPE/article/view/2639 https://doi.org/10.33736/jaspe.2639.2020

[16]. Rasool, A., Bunterngchit, C., Tiejian, L., Islam, M.R., Qu, Q. and Jiang, Q. (2022) Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. *International Journal of Environmental Research and Public Health*, **19**, Article 3211. https://www.mdpi.com/1660-4601/19/6/3211 https://doi.org/10.3390/ijerph19063211

[17]. Merouane, E. and Said, A. (2022) Prediction of Metastatic Relapse in Breast Cancer using Machine Learning Classifiers. *International Journal of Advanced Computer Science and Applications*, **13**, 176-181. https://doi.org/10.14569/IJACSA.2022.0130222

[18]. Aswathy, M.A. and Jagannath, M. (2021) An SVM Approach towards Breast Cancer Classification from H&E-Stained Histopathology Images Based on Integrated Fea- tures. *Medical & Biological Engineering & Computing*, **59**, 1773-1783. https://link.springer.com/article/10.1007/s11517-021-02403-0 https://doi.org/10.1007/s11517-021-02403-0

[19]. Sengar, P.P., Gaikwad, M.J. and Nagdive, A.S. (2020) Comparative Study of Ma- chine Learning Algorithms for Breast Cancer Prediction. *Proceedings of* 2020 *Third International Conference on Smart Systems and Inventive Technology* (*ICSSIT*), Tirunelveli, 20-22 August 2020, 796-801. https://doi.org/10.1109/ICSSIT48917.2020.9214267 https://ieeexplore.ieee.org/abstract/document/9214267

[20]. Gayathri, B.M. and Sumathi, C.P. (2016) Comparative Study of Relevance Vector Machine with Various Machine Learning Techniques Used for Detecting Breast Can- cer. *Proceedings*

of 2016 *IEEE International Conference on Computational Intelli- gence and Computing Research* (*ICCIC*), Chennai, 15-17 December 2016, 1-5. https://ieeexplore.ieee.org/abstract/document/7919576 https://doi.org/10.1109/ICCIC.2016.7919576

[21]. Kumar, A. and Poonkodi, M. (2019) Comparative Study of Different Machine Learning Models for Breast Cancer Diagnosis. In: Chattopadhyay, J., Singh, R. and Bhattacherjee, V., Eds., *Innovations in Soft Computing and Information Technolo- gy*, Springer, The Gateway, 17-25. https://doi.org/10.1007/978-981-13-3185-5_3

[22]. Sharma, A., Kulshrestha, S. and Daniel, S. (2017) Machine Learning Approaches for Breast Cancer Diagnosis and Prognosis. *Proceedings of* 2017 *International Confe- rence on Soft Computing and Its Engineering Applications* (*icSoftComp*), Changa, 1-2 December 2017, 1-5. https://ieeexplore.ieee.org/abstract/document/8280082 https://doi.org/10.1109/ICSOFTCOMP.2017.8280082

[23]. Zheng, B., Yoon, S.W. and Lam, S.S. (2013) Breast Cancer Diagnosis Based on Fea- ture Extraction Using a Hybrid of K-Means and Support Vector Machine Algorithms. *Expert Systems with Applications*, **41**, 1476-1482. https://doi.org/10.1016/j.eswa.2013.08.044

[24]. Javapoint (2023) K-Nearest Neighbor (KNN) Algorithm for Machine Learning. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[25]. XORIANT (2023) Decision Trees for Classification: A Machine Learning Algorithm. https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning- algorithm/

[26]. Wikipedia (2023) Random Forest. https://en.wikipedia.org/wiki/Random_forest

[27]. Rajaguru, H. and Chakravarthy, S.R. (2019) Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer. *Asian Pacific Journal of Cancer Prevention*, **20**, 3777-3781.

[28]. Gupta, M. and Gupta, B. (2018) A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques. *Proceedings of* 2018 *Second In- ternational Conference on Computing Methodologies and Communication* (*ICCMC*), Erode, 15-16 February 2018, 997-1002. https://doi.org/10.1109/ICCMC.2018.8487537

[29]. Salama, G.I., Abdelhalim, M.B. and Zeid, M.A. (2012) Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. *International Journal of Computer and Information Technology*, **1**, 36-43.

[30] Harinishree M. S., Aditya C. R.*, Sachin D. N.: Detection of Breast Cancer using Machine Learning Algorithms –A Survey. IEEE Xplore Part Number: CFP21K25-AR