Data Mining Project: Breast Cancer Wisconsin (Diagnostic) Dataset

Md Moinul Islam
MSc (Technology) in CSE
University of Oulu
moinul.islam@student.oulu.fi
2307268

Efta Khairul Bashar
MSc (Technology) in CSE
University of Oulu
ebashar23@student.oulu.fi
2307235

Abdullah Reza
MSc in Business Analytics
University of Oulu
abdullah.reza@student.oulu.fi
2310691

Anthony Korsah
MSc in Business Analytics
University of Oulu
anthony.korsah@student.oulu.fi
student number

***Abstract*—The global prevalence of breast cancer requires technological advances in diagnostic procedures to improve detection and treatment outcomes. The present investigation uses the Breast Cancer Wisconsin (diagnosis) dataset to increase breast cancer diagnosis accuracy through the use of machine learning techniques. On a dataset enriched with cell nuclei measurements from fine needle aspirates of breast cancers, this research evaluates several classifiers, including K-Nearest Neighbors (KNN), Support Vector Machine, Random Forest, XGBoost, and Multilayer Perceptron (MLP). The primary objective of this research is to evaluate and compare the performance of machine learning models that can consistently distinguish between benign and malignant tumors and optimize these models through pre-processing techniques, such as feature scaling and outlier removal. Multilayer Perceptron outperforms other models in performance metrics, such as accuracy, recall, and F1-score, suggesting that it may be useful in clinical applications.**

***Index Terms*—Breast cancer, Neural network, Machine learning algorithms, Random oversampling, Isolation forest, Recursive feature elimination with cross-validation**

I. Introduction

The Breast Cancer Wisconsin (Diagnostic) dataset [1], which was released to the scientific community on October 31, 1995, is a critical resource in the field of medical informatics, notably for the investigation and categorization of breast cancer. This dataset, housed at the University of California, Irvine's Machine Learning Repository, is a multivariate collection particularly developed for classification tasks in the health and medicine fields.

The drive for investigating machine learning applications in breast cancer detection stems from the disease's global burden, which remains one of the major causes of cancer-related death among women globally. Early identification and proper categorization of breast cancer considerably enhance survival rates and treatment results. Advancements in machine learning provide powerful tools for improving diagnostic precision, addressing a key demand in oncological treatment.

Our interest in this research stems from the promise of machine learning to transform established diagnostic approaches by delivering faster, more reliable, and cost-effective solutions. The combination of artificial intelligence and medical diagnostics has the potential to revolutionize patient outcomes by personalizing medicine and minimizing subjective interpretation of medical imaging.

The Breast Cancer Wisconsin (Diagnostic) dataset is used in this study, which is a well-curated collection of characteristics extracted from fine needle aspirates of breast tumors. The collection contains measurements of cell nuclei properties extracted from digitized pictures, making it an important resource for constructing and testing prediction algorithms. Our objective is to accurately classify tumors as benign or malignant based on these features by utilizing a variety of machine learning techniques.

The scope of this project includes preparing the dataset to remove any potential skew or bias in feature distributions, developing several classification models, and assessing their performance using measures such as accuracy, precision, recall, and F1-score. This extensive study seeks to determine the most efficient model that makes reliable diagnostic predictions, therefore, adding vital insights to current cancer research and diagnostics efforts. This project aims to deliver the following,

1) This study aims to improve breast cancer diagnostic accuracy by implementing and evaluating different machine learning techniques. The enhanced prediction models created by this effort can assist in discriminating between benign and malignant tumors with more precision than standard diagnostic approaches.
2) By thoroughly testing each machine learning model's performance using metrics (accuracy, precision, recall, F1-score), this initiative creates a complete baseline for future research.

The rest of this paper is organized as follows: Section II discusses the relevant studies in the current domain. Section III provides the objectives and relevant research questions of the project. Section IV and V represent the dataset and the methodology with its architecture and training parameters. Section VI provides an overview of the experimental evaluation, findings and results. Section VII presents discussions on those results and summarizes our research and provides some recommendations for further research. Finally, section VIII provides the reflections on the group work.

II. Related work

To offer a complete overview of the literature relevant to the application of machine learning in breast cancer diagnostics utilizing the Breast Cancer Wisconsin (Diagnostic) dataset,

the following significant research and their conclusions are reviewed. This overview represents a compilation of at least 10 papers that demonstrate the scope of study in this area.

The research by [2] describes a unique machine learning-based framework for breast cancer prediction that employs Random Forest, Gradient Boosting, Support Vector Machine, Artificial Neural Network, and Multilayer Perceptron techniques. It employs a hybrid Multi-layer Perceptron Model with 5-fold cross-validation. Connection-based feature selection approaches are used to enhance classification, resulting in a high accuracy of 99.12% on the dataset. The study in [3] created four machine learning models to improve breast cancer diagnosis by utilizing data exploratory methods (DET) such as feature distribution, correlation, removal, and hyperparameter optimization. These models, implemented on the WDBC and BCCD datasets, were evaluated using confusion matrices and K-fold cross-validation, yielding high accuracies: polynomial SVM (99.3%), logistic regression (98.06%), KNN (97.35%), and ensemble classifier (97.61%). Through the use of data exploratory methods (DET), including feature distribution, correlation, removal, and hyperparameter optimization, the study in [4] constructed four machine learning models to improve breast cancer diagnosis. These models were evaluated using confusion matrices and K-fold cross-validation on the WDBC and BCCD datasets. The models with the highest accuracy were the polynomial SVM (99.3%), the logistic regression (98.06%), the KNN (97.35%), and the ensemble classifier (97.61%). A deep learning-based method based on the Wisconsin Breast Cancer Database is proposed for breast cancer diagnosis in [5]. Following the preparation of the data using methods such as Normalizer, StandardScaler, and Label Encoder, a neural network model was trained and successfully achieved 99.67% accuracy. The study uses 11 specific characteristics from the dataset for diagnosis, highlighting the superiority of deep learning over conventional machine learning methods. The study [6] evaluates the efficacy of two well-known machine learning methods for categorizing the Wisconsin Breast Cancer dataset using ROC area, accuracy, precision, recall, and recall metrics. According to the study, the Support Vector Machine method produces the best accuracy, highlighting its potential for early cancer detection and prediction.

The study in [7] examines three machine learning models: Random Forest, Decision Tree, and Logistic Regression. With five important indicators, the Random Forest model produced the best results, with a cross-validation score of 93% and a maximum accuracy of 95%, demonstrating its efficacy in predicting breast cancer. A study in [8] investigates a hybrid machine learning method for breast cancer detection. With the use of linear discriminant analysis for feature reduction and Support Vector Machine, Artificial Neural Networks, and Naïve Bayes combined, it achieves 99.07% precision, 98.41% recall, 98.82% accuracy, and an AUC score of 0.9904. A study in [9] evaluates six machine learning algorithms, showcasing the Multilayer Perceptron (MLP) for its exceptional test accuracy of around 99.04%. Over 90% accuracy was attained by all algorithms, with particular attention to sensitivity and specificity measures. To identify breast cancer using the dataset, a study in [10] assesses machine learning algorithms including SVM, Decision Tree, Naive Bayes, k-NN, Adaboost, XGboost, and Random Forest. XGboost has the greatest accuracy of 98.24%. The study in [11] investigates the use of Particle Swarm Optimisation (PSO) combined with Decision Stump, J48 pruned tree, and Naive Bayes classifiers to improve the detection accuracy of breast cancer. The technique selects the best features from the dataset to increase accuracy, precision, and Kappa statistics, among other performance indicators.

III. Objectives

The overarching objectives of this project utilizing the Breast Cancer Wisconsin (Diagnostic) dataset are rooted in enhancing the accuracy and efficiency of breast cancer diagnosis through machine learning techniques. The specific objectives, research questions, and expected outcomes driving this study are as follows:

A. Objectives

1) To evaluate and compare the performance of various machine learning algorithms including but not limited to Support Vector Machine (SVM), XGBoost, Random Forest, K-Nearest Neighbours and Neural Network in classifying breast cancer tumors as benign or malignant.
2) To optimize these models through preprocessing techniques such as feature selection, and hyperparameter tuning to achieve the highest possible accuracy.
3) To develop a prediction model that can be incorporated into clinical processes and provide real-time diagnostic help to healthcare providers.

B. Research Questions

1) Which machine learning algorithm provides the highest accuracy for diagnosing breast cancer using the provided dataset?
2) How do different preprocessing techniques affect the performance of these machine learning models in terms of accuracy, sensitivity, and specificity?

C. Expected Results

1) We expect that through rigorous testing and optimization, we will identify one or more machine learning models that significantly outperform traditional diagnostic methods in terms of accuracy, sensitivity (true positive rate), and specificity (true negative rate).
2) The study is expected to provide insights into which features are most indicative of malignancy in breast cancer, and how different models handle the complexity and variability in data.
3) By the end of this project, we anticipate establishing a robust model that can be proposed for real-world clinical testing, potentially culminating in a tool that enhances diagnostic processes and patient outcomes in breast cancer care.

IV. Data

This study used the Breast Cancer Wisconsin (Diagnostic) dataset [1], which was provided by Dr. William H. Wolberg of the University of Wisconsin Hospitals in Madison. The dataset, which was compiled on October 31, 1995, is intended to enable the development and validation of machine learning algorithms for breast cancer diagnosis. It is publicly available through the UCI Machine Learning Repository [12] and is normally in CSV format, allowing for easy integration with a variety of analytical applications.

This dataset comprises 569 cases, each tagged with an ID and a diagnostic label ('M' for malignant or 'B' for benign), as well as 30 feature variables generated from fine needle aspiration (FNA) samples of breast masses. These features include ten primary measurements—radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension—each of which is quantified in three ways: mean, standard error, and the average of the three largest values, providing a detailed foundation for predictive modelling in oncological diagnostics.

A. Exploratory Data Analysis

The correlation matrix, especially when presented as a heat map in Figure 1, provides a vivid representation of the correlations between variables in the dataset. High correlations (0.9+) between dataset variables like radius, area, and perimeter indicate derived connections. Texture_mean and texture_worst are highly correlated (0.98), showing that texture_worst reflects extreme values. Other properties, such as compactness, concavity, and concave tips, exhibit strong dependency (0.7-0.9), emphasizing their importance in forming cellular structures.

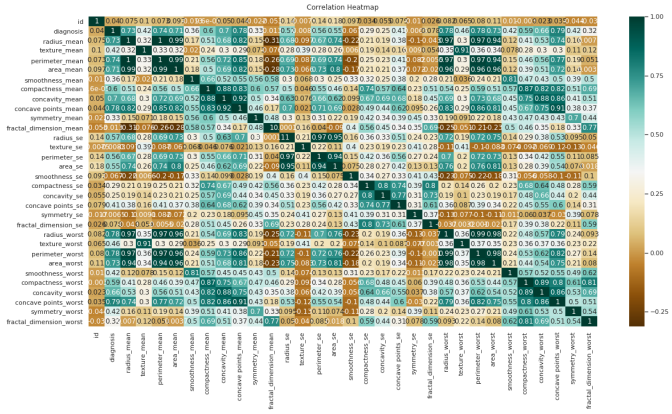

Fig. 1. Correlation Heatmap of the Features of Wisconsin Breast Cancer Dataset

Pair plots visualize the distribution and relationships between dataset variables, using 2D scatter and KDE plots. Diagonal plots reveal a clear distinction between malignant and benign data points in Figure 2, with the upper and lower triangles of the matrix mirroring each other.

We gained considerable insights into data distribution and variance by analyzing the dataset with box plots and violin plots as shown in Figure 3. These illustrations clearly distinguish between benign and malignant instances, notably in aspects such as radius, perimeter, and area, emphasizing their diagnostic value. Outliers across all parameters and overlapping distributions in texture_se highlight the importance of rigorous data preparation for proper modeling.

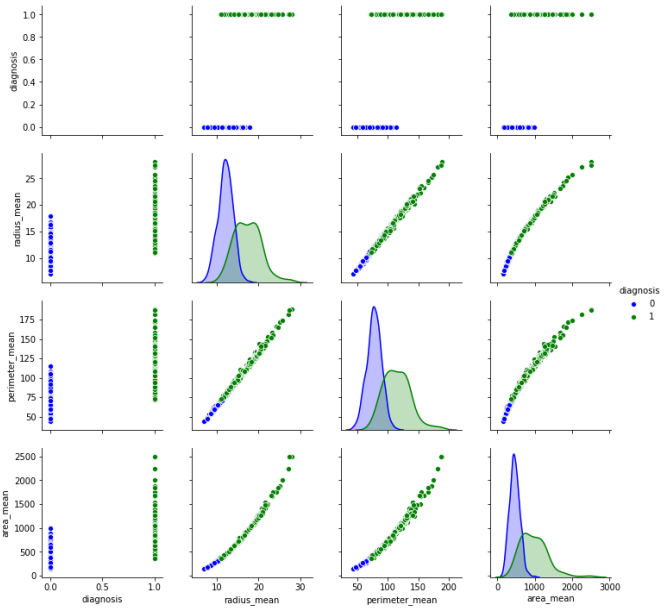

Fig. 2. Pair plot between target (diagnosis) and other features

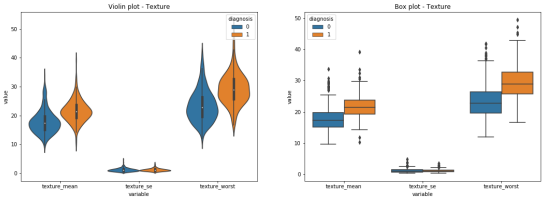

Fig. 3. Violin plot and Box plot of different features

B. Data pre-processing

Several important preprocessing measures were done to maintain the integrity and utility of the Breast Cancer Wisconsin (Diagnostic) dataset for machine learning research. Throughout the process, a thorough understanding of the data's characteristics was established, and specific method was used to handle the issues given by outliers.

1) Data Cleaning: Initially, the dataset was checked for missing values and discrepancies. Given the robustness of the data collecting method, no missing values were found, allowing us to proceed without imputation.

2) Feature Scaling: To normalize the range of the continuous input variables, StandardScaler was used for feature scaling. This is critical because machine learning algorithms that compute distances between data points, such as Support Vector Machines (SVM) and k-nearest neighbors (KNN), perform

better with scaled data. Scaling guarantees that features with longer ranges do not outperform those with shorter ranges.

3) Outlier Detection and Removal: During the exploratory data analysis, outliers were found in various characteristics. Outliers can distort findings by changing the dataset's mean and standard deviation. To address this, Isolation Forest [13] algorithm based machine learning approach was used. This approach is useful for high-dimensional datasets since it isolates abnormalities rather than profiling regular data points. By recursively splitting the data, the Isolation Forest discovers cases that require more splits to isolate—these are likely to be outliers. The 3D scatter plot as shown in Figure 4 visualizes data points labeled as inliers (green) and outliers (red) across three principal component axes (PC1, PC2, PC3) using Principal Component Analysis (PCA).

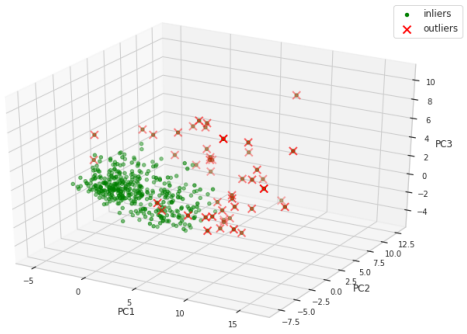

Fig. 4. 3D Scatter Plot of Data Points with Outliers

4) Class Balancing: While preparing the dataset for further analysis, it was observed that the data was not balanced. This imbalance between the classifications (benign and malignant) may result in biased findings, with machine learning models favoring the majority class. To address this issue, Random Oversampling was used. This method involves replicating instances from the minority class (malignant tumors) until the number of instances in both classes is equal. This strategy helps to keep the model from being biased towards the more common class (benign tumors) as shown in Figure 5.

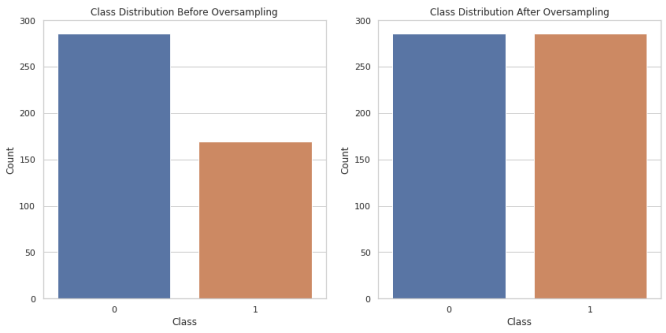

Fig. 5. Class Balancing using Random OverSampling

If the dataset's initial imbalance and the existence of outliers are not handled, they may result in biased models that favor the majority class and skewed forecasts owing to extreme values. We balanced class distribution using Random OverSampling, however this posed the danger of overfitting, which might make the model overly customized to the training data. Similarly, while outlier elimination via Isolation Forest improves model generalizability by removing extreme data points, it may result in the loss of potentially useful, yet uncommon observations. These preprocessing decisions demand rigorous model evaluation to guarantee robustness and generalizability to real-world data, stressing the importance of measures that indicate both sensitivity and specificity.

C. Data Splitting

Finally, to assess how well machine learning models work, the dataset is finally divided into training and testing sets. One of the insights obtained from this step is to make sure that the data is shuffled randomly to prevent biases in the split.

V. Methods

The overview of the proposed methodology as shown in Figure 6 illustrates an organized strategy for analyzing breast cancer using machine learning techniques. It starts with collecting the dataset. The first phase, feature analysis, includes thoroughly investigating the dataset using histogram, distribution visualization, and the computation of descriptive statistics to acquire insights from the features.

The methodology applying different classifiers - KNN (k-Nearest Neighbors), XGBoost, Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) - on the Wisconsin Breast Cancer Diagnostic Dataset reflects a comprehensive approach to identifying the most effective algorithm for diagnosing breast cancer from FNA images. KNN [14] is a simple yet efficient classification algorithm that assigns a class based on the majority vote of the closest data points in the feature space. It's especially handy when the data is clearly split into multiple kinds. However, its performance may suffer with a high number of features due to the curse of dimensionality. Extreme Gradient Boosting (XGBoost) [15] is an ensemble approach that makes use of a gradient boosting framework. It is resistant to overfitting and is well-known for its good performance in classification tasks, including the effective handling of feature interactions in high-dimensional data. Random Forest [16] is another ensemble approach that constructs several decision trees and outputs the mode of the classes for classification purposes. It delivers good performance without requiring extensive parameter adjustment and and is less likely to overfit compared to individual decision trees. SVM [17] is an effective classifier that identifies the best hyperplane to optimize the margin between classes. It works well in high-dimensional spaces and is adaptable since it can modify the feature space using a variety of kernel functions. Furthermore, SVM is included because of its efficacy in distinguishing non-linearly separable classes by changing the feature space. Finally, the multilayer perceptron [18] method is

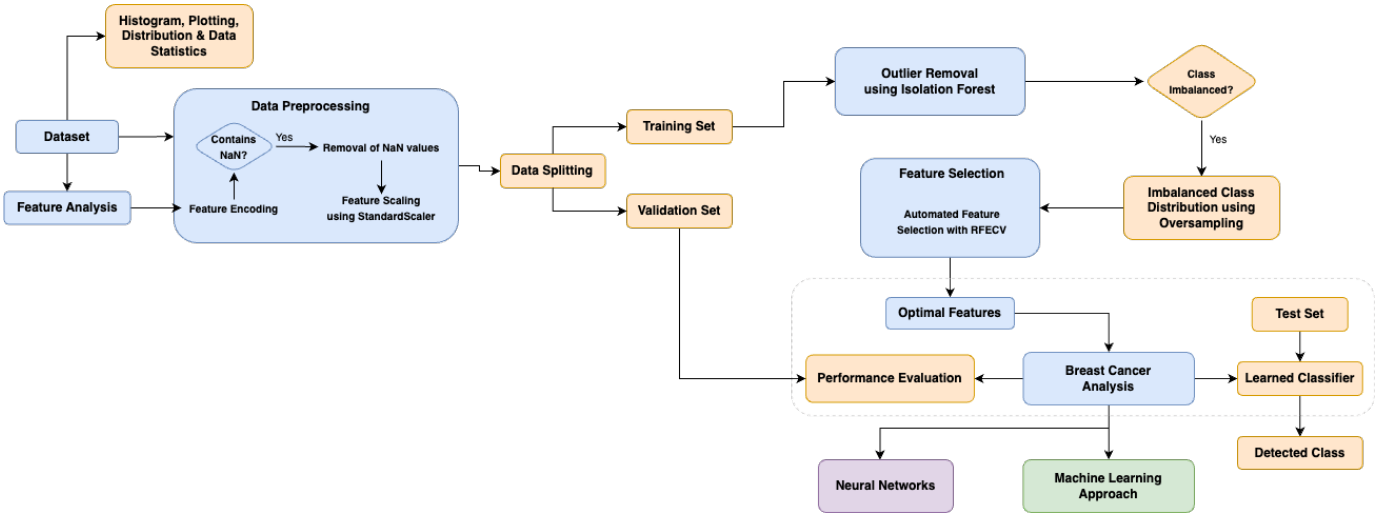

Fig. 6. Overview of the Proposed Methodology

a form of neural network designed to capture complex patterns and non-linear correlations in data. It can learn sophisticated data structures, but parameters must be carefully tuned, and it is prone to overfitting on short samples. The chosen classifiers also provide a balance of simplicity and complexity, ranging from the simple implementation of KNN to the more complicated neural network design of MLP, enabling a diverse range of learning dynamics to capture various features of the data. Given the dataset's absence of missing values, the classifiers concentrate on identifying the underlying patterns without the requirement for imputation, allowing for a more accurate evaluation of their performance.

VI. Results

We initially trained several classifiers — KNN, XGBoost, Random Forest, Support Vector Machine, and Multilayer Perceptron — with and without considering outliers. All classifiers achieved test accuracies over 95%, with the MLP model reaching the highest test accuracy at 98.25%. We then trained the same classifiers on the dataset after removing the outliers from the training data. Most of the results were quite similar to the previous ones. The performance metrics of these methods are represented in Table I. This table compares the performance metrics of five distinct machine learning approaches when used on a breast cancer diagnostic dataset, both before and after outlier elimination. The algorithms tested are K-Nearest Neighbours (KNN), XGBoost, Support Vector Machine (SVM), Random Forest, and Multi Layer Perceptron (MLP). Performance measurements include Precision (PRE), Recall (REC), F1-Score, and Accuracy (ACC), with an emphasis on the macro average to show overall performance across 2 classes.

Before outlier elimination, the MLP outperforms all measures, showing an excellent fit to the data. The SVM and Random Forest perform similarly well, with precision, recall, and F1-scores of 0.97, as well as accuracy of 0.97. Following outlier removal, SVM and MLP had the excellent performance across all metrics, suggesting robustness in the presence of outliers. The training and validation loss and accuracy curves for Multilayer Perceptron before outlier removal show notable swings, indicating probable overfitting, as seen in Figure 7. Figure 8 indicates that after removing outliers, validation measures are less unstable, indicating better model stability and generalization.

TABLE I
Performance Metrics of Different Methods on Breast Cancer Diagnostic Dataset

Methods	**Overall (Macro Average)**			
	PRE	REC	F1-Score	ACC
Before Outlier Removal				
K-Nearest Neighbours	0.96	0.96	0.96	0.96
XGBoost	0.96	0.95	0.95	0.96
Support Vector Machine	0.97	0.97	0.97	0.97
Random Forest	0.97	0.97	0.97	0.97
Multi Layer Perceptron	**0.99**	**0.98**	**0.98**	**0.98**
After Outlier Removal				
K-Nearest Neighbours	0.96	0.96	0.96	0.96
XGBoost	0.96	0.95	0.95	0.96
Support Vector Machine	**0.97**	**0.97**	**0.97**	**0.97**
Random Forest	0.96	0.96	0.96	0.96
Multi Layer Perceptron	**0.97**	**0.97**	**0.97**	**0.97**

The ROC curve in Figure 9 shows high Area Under the Curve (AUROC) values for all classifiers before outlier removal, indicating excellent classification performance. The ROC curve in Figure 10 demonstrates similarly high AUROC values after outlier removal, with slight variations in performance metrics, suggesting that outlier removal had a minimal impact on the discriminative ability of the classifiers.

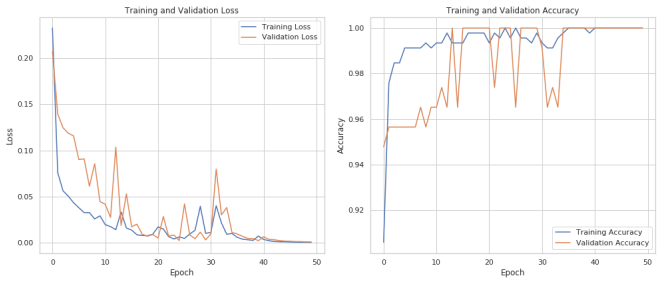

Fig. 7. MLP Loss and Accuracy Curves before Outlier Removal

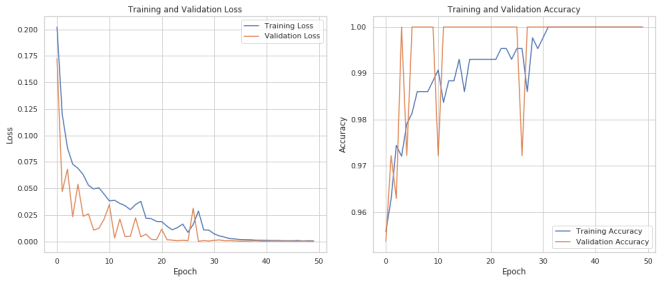

Fig. 8. MLP Loss and Accuracy Curves after Outlier Removal

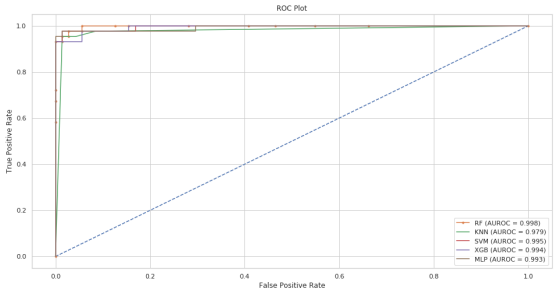

Fig. 9. ROC Curve of Different Methods Used before Outlier Removal

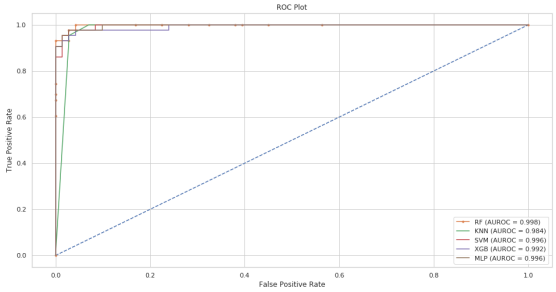

Fig. 10. ROC Curve of Different Methods Used after Outlier Removal

VII. Discussion

The study's findings demonstrate the usefulness of several machine learning classifiers in classifying breast cancer using the Wisconsin Breast Cancer Diagnostic Dataset. The evaluated classifiers, K-Nearest Neighbours (KNN), XGBoost, Support Vector Machine (SVM), Random Forest, and Multi Layer Perceptron (MLP), achieved good accuracy, precision, recall, and F1-scores both before and after outlier removal. The Multilayer Perceptron (MLP) model, in particular, achieved the best performance metrics, suggesting a significant capacity to detect complex patterns in data.

These findings are consistent with previous research, as several studies have also reported high accuracies using similar methodologies on the same dataset. For instance, the studies cited in [7], [10], [11] found machine learning models capable of achieving accuracies over 95% with the use of advanced algorithms and preprocessing techniques. This consistency supports the robustness of machine learning models in the application of breast cancer classification. Compared to previous studies, our research expands on the findings by doing a comparison analysis of several classifiers with an emphasis on outlier influences. Notably, the performance of SVM and MLP remained steady even after outliers were removed, indicating that they can handle data anomalies. This aspect of our research gives a better understanding of the models' generalizability and reliability in real-world scenarios where data anomalies occur often.

While our research gives insightful perspectives, it is important to acknowledge its limits. The key issue is the dependence on a single dataset, which may not capture all variances of breast cancer incidences worldwide. The study's emphasis on technical performance indicators may ignore the clinical applicability and interpretability of the models, which are critical for real-world medical application. Future research could expand on these findings by using multi-center data to improve the models' generalizability across populations and different timeline. Finally, investigating the effects of newer and more effective machine learning or deep learning methodologies may result in additional improvements in the diagnostic accuracy of breast cancer classification models.

VIII. Reflection on Group Work

It was a good experience working on a data mining project in a group. We tried to understand the data separately first. Then we discussed our findings with one another and a few confusions were resolved during this session. The most interesting part was when we compared the classifier results with and without outlier removal. Initially, our assumption was the result of the classifiers would improve drastically without the outliers, which was not the case. We learned that even with a simple dataset like this one there is a lot to explore both in terms of the relations of the features and training the models with proper settings.

References

[1] W. Wolberg, W. Street, and O. Mangasarian, "Breast cancer wisconsin (diagnostic)," *UCI Machine Learning Repository*, vol. 414, p. 415, 1995.

[2] S. Aamir, A. Rahim, Z. Aamir, S. Abbasi, M. Khan, M. Alhaisoni, M. A. Khan, K. Khan, and J. Ahmad, "Predicting breast cancer leveraging supervised machine learning techniques," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022.

[3] A. Rasool, C. Bunterngchit, T. Luo, M. R. Islam, Q. Qu, and Q. Jiang, "Improved machine learning-based predictive models for breast cancer diagnosis," *International Journal of Environmental Research and Public Health*, vol. 19, 2022.

[4] R. Soni, S. Zaina, and D. M. Latha, "Unlocking the potential of machine learning for accurate diagnosis of breast cancer," *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pp. 1–8, 2023.

[5] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using deep learning algorithm," *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 98–103, 2018.

[6] E. A. Bayrak, P. Kırcı, and T. Ensari, "Comparison of machine learning methods for breast cancer diagnosis," *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pp. 1–3, 2019.

[7] Y.-G. Wei, D. Zhang, M. Gao, Y. Tian, Y. He, B. Huang, and C. Zheng, "Breast cancer prediction based on machine learning," *Journal of Software Engineering and Applications*, 2023.

[8] D. A. Omondiagbe, S. Veeramani, and A. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," *IOP Conference Series: Materials Science and Engineering*, vol. 495, 2019.

[9] A. F. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," pp. 5–9, 2017.

[10] M. Mangukiya, "Breast cancer detection with machine learning," *International Journal for Research in Applied Science and Engineering Technology*, 2022.

[11] A. Sen, "Application of particle swarm optimization on wisconsin diagnosis breast cancer dataset," *Journal of Scientific Enquiry*, 2022.

[12] [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[13] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422.

[14] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer, 2003, pp. 986–996.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[16] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[17] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[18] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classifiaction," 1992.