**Ilorin Journal of Computer Science and Information Technology**
**Vol. 6, No. 2 (2023)**

**ISSN: 2141-3959** *(print)*

# A comparison of Boosting techniques for Classification of Microarray data

Ronke Seyi BABATUNDE[1], Akinbowale Nathaniel BABATUNDE[2], Bukola Fatimah BALOGUN[3], Ibrahim Aliyu YAKUBU[4], Roseline Oluwaseun OGUNDOKUN[5], Kolawole Yusuf OBIWUSI[6] and Emmanuel UMAR[7]

[1, 2, 3] *Department of Computer Science, Kwara State University, Malete, PMB 1530, Ilorin, Kwara State, Nigeria; ronke.babatunde@kwasu.edu.ng*[*], *akinbowale.babatunde@kwasu.edu.ng*, *bukola.balogun@kwasu.edu.ng*
[4] *Department of Computer Science, Bingham University, Nasarawa State, Nigeria; talktoibro80@gmail.com*,
[5]*Department of Computer Science, Landmark University, Omu-Aran, Kwara State, Nigeria; ogundokun.roseline@lmu.edu.ng*
[6]*Department of Computer Science, Summit University, Offa, Kwara State, Nigeria; obiwusi.kolawole@summituniversity.edu.ng*,
[7]*Tai Solarin University of Education, Ogun state; umaremmanuel5@gmail.com.*
***** *akinbowale.babatunde@kwasu.edu.ng*

**Abstract**
**Context:** The advancements in technology, particularly microarrays, have played a pivotal role in enhancing crucial aspects within the domains of genomics and bioinformatics. These advancements have significantly contributed to the enhancement of illness diagnosis, evaluation of therapy response in patients, and advancements in cancer research. Microarray data often exhibits a substantial likelihood of encompassing extraneous and duplicative factors, hence introducing noise into the dataset. Consequently, the process of scrutinizing the data to identify significant patterns for diagnosis can be quite daunting when employing conventional statistical approaches. Numerous studies are currently being conducted to enhance the analysis of microarray data, with the aim of enhancing performance and prediction accuracy at an accelerated pace. Most of these earlier methods are limited in their predictive capacity and are characterised by high computational time and algorithm complexity. **Objective:** This research addresses some of these issues by implementing the classification of microarray data using Boosting algorithms. **Method:** Benchmarked on a publicly available dataset, the microarray data was cleaned, normalised and salient features carrying essential information were obtained. Three state of the art boosting algorithms; AdaBoost, Gradient Boost, and XGBoost were used in classifying the microarray data and the performance result of each was compared. **Results:** The experimental findings indicate that XGBoost demonstrates superior performance compared to other boosting approaches, with a classification accuracy rate of 98.18% and training time of 11seconds. **Conclusions:** The novelty of the experiment compared to earlier work is evident in the training time reported which is an information not frequently explicit in other report of findings.
**Keywords:** Classification, Boosting techniques, and Microarray data

## 1. Introduction

The field of biotechnology has experienced significant advancements, enabling the accumulation and storage of extensive quantities of biological data through the utilization of cutting-edge technologies like microarrays. Microarrays facilitate the quantification and retrieval of data pertaining to gene expressions, which encompass the molecular processes by which genes are transcribed and translated into functional protein products (Yeang et al., 2021; Prajapati et al., 2023). The analysis of gene expressions can be employed for the purpose of investigating the impact of therapies or identifying diseases. This is achieved through the comparison of the gene expression profiles of healthy individuals with those individuals who exhibit altered or infected gene expressions as a result of the therapy or disease (Risky et al., 2018). Technological advancements, particularly in the form of microarrays, have played a pivotal role in enhancing several crucial aspects within the domains of genomics and bioinformatics. These advancements have notably contributed to the enhancement of illness diagnosis, assessment of therapy response in patients, and advancements in cancer research (Manikandan, et al., 2023). According to Tan & Gilbert (2022), a single tissue sample obtained from a microarray has the potential to encompass a multitude of gene expressions, reaching up to tens of thousands. Consequently, the process of scrutinizing the data to identify significant patterns can be exceedingly daunting when employing conventional statistical approaches. Classification techniques play an important role in micro array data analysis, and this have been demonstrated by numerous researchers with different techniques and challenges. (Sung et al., 2021).

The nascent microarray technology enables researchers to concurrently assess the expression level of several genes inside a given tissue sample. Frequently, a crucial element of the analysis of a set of microarray experiments entails class prediction, wherein an algorithm utilizes the outcomes of these tests to deduce a guideline for forecasting characteristics of a tissue sample, relying on its expression profile. (Dudoit et al., 2022).

Performing statistical analysis on microarray data poses challenges due to the presence of systematic bias in the microarray output, which is characterized by sparsity and

high dimensionality (Kumar et al., 2022). The issue of high dimensionality arises when the number of genes exceeds the number of samples, resulting in a condition known as sparsity, wherein a significant proportion of these genes are deemed useless for analysis. Due to the escalating intricacy in the quantity of samples and genes, there has been a rising preference for computationally costly statistical methods, such as machine learning algorithms, in the analysis of microarrays. According to Tan & Gilbert (2022), Machine learning algorithms are frequently utilized in the domains of genomics and bioinformatics for predictive tasks, such as the identification of tumor tissues or the diagnosis of various cancer kinds in patients. Yifan et al., (2021). In this particular context, the objective of machine learning is to acquire knowledge regarding the characteristics of samples or patients that pertain to a specific group, and subsequently categorize novel samples or patients into these respective categories. The application of machine learning has proven to be effective in the classification of cancer patients and non-cancer patients based on the analysis of microarray profiles. Various techniques have been employed in the classification of Microarray data, including Support Vector Machines (SVMs) (Brown et al., 2020), the k-nearest neighbor classifier (Yeang et al., 2021), the C4.5 decision tree (Manikandan et al., 2023), rule-based classification methods (Yeang et al., 2021), ensemble methods such as Bagging and boosting (Tan & Gibert, 2022. Support Vector Machines (SVMs), decision trees, and ensemble algorithms are often employed techniques in the classification of Microarray data. Support Vector Machines (SVMs) and Random Forests (RFs) are two well-established machine learning techniques that have gained significant interest in recent years (Tan & Gilbert, 2022). They have been employed in several types of research and their widespread usage has resulted in their ready accessibility through numerous software programs. Logistic regression, a conventional classification approach, has demonstrated its usefulness in scenarios involving high-dimensional data when combined with regularization approaches, as exemplified by the work of (Varshal and Vishal, 2023). Therefore, it would be intriguing to conduct a comparative analysis of prevalent machine learning techniques in order to ascertain potential significant disparities among them.

Various methods for variable reduction can address the challenges posed by high-dimensionality and sparsity. The process of reducing variables in microarray analysis is commonly referred to as gene selection, as it involves selecting specific gene expression coefficients as the variables of interest (Varshali et al., 2023). Microarray data sometimes exhibits a substantial likelihood of encompassing extraneous and duplicative variables, hence introducing noise into the algorithms and leading to suboptimal performance and reduced accuracy in predictions. To enhance the analytical quality, it is frequently advantageous to decrease the number of characteristics inside the dataset. The reduction of variables can be achieved through the implementation of a variable selection technique during the data preparation phase. (Kurian & Jyothi, 2022). This technique aims to identify and retain the most significant variables (or genes) in order to enhance the accuracy of the model. Similar to machine learning techniques, there exist numerous approaches for eliminating unnecessary variables, with the trade-off lying between complexity and runtime.

The concept of boosting algorithms involves the amalgamation of multiple heuristic guidelines into a more precise collective classification prediction rule (Freund & Schapire, 2019). The functioning of these algorithms involves the iterative application of a sub algorithm, commonly referred to as a base learner, on a given dataset. Prior to each application, the examples undergo a process of reweighting, wherein the significance of the examples that have been frequently misclassified by the rules generated by prior applications of the base learner is amplified. Upon reaching a predetermined number of iterations, the class prediction rules generated are combined through a voting mechanism. Each rule is assigned a weight, and a tissue sample is predicted to be a tumor if the cumulative weight of the rules favouring a tumor classification surpasses the cumulative weight of the rules favouring a normal classification. (Yifan et al., 2021). The objective of this study is to compare the performance of state of the art boosting techniques on classification of microarray data and find out the best. The study additionally assesses the performance of the system in terms of accuracy, precision, sensitivity, and computing time.

**Contribution:** The novelty of this research is the evaluation of the performance of boosting techniques based on the computation time which is an important metrics not frequently reported by previous researchers.

## 2. Related work

Numerous studies have been conducted to compare machine learning algorithms, employing various methodologies and diverse evaluation techniques. The majority of comparative research adhere to a standardized approach, wherein many datasets are selected from previously published studies on microarray analysis. Subsequently, machine learning algorithms are employed to assess performance based on a predetermined criterion. In the majority of instances, the primary criterion employed to assess performance is the level of accuracy in predictions. Several studies lack precise information regarding the computing time of the algorithms employed.

The study conducted by Prajapati et al. (2023) centers on the categorization of microarray data through the utilization of machine learning techniques, specifically Decision Tree (DT), k-nearest neighbor (KNN), Support Vector Machine (SVM), logistic regression (LR), and random forest (RF) algorithms. The study examines the comparative accuracy of various algorithms, finding that

KNN, SVM, LR, and RF demonstrate accuracies above 80% and, in certain instances, reaching 100%. Nevertheless, the research fails to take into account the computational cost or scalability of the algorithms when applied to bigger microarray datasets.

In their study, Gongora et al. (2022) examine the application of the Support Vector Machine (SVM) classifier in the categorization of microarray data that is both high-dimensional and imbalanced. The research emphasizes the sensitivity of SVM classifiers, as well as other models, to imbalanced class distributions. It suggests the use of the Synthetic Minority Oversampling Technique (SMOTE) as a preprocessing strategy to address this issue. SMOTE involves boosting the amount of samples from the minority class in order to balance the class distribution. The study lacks a comprehensive examination of the limits associated with the SMOTE Boost technique, including potential downsides and situations in which its effectiveness may be limited.

The study conducted by Ben-Dor et al. (2020) represents an early instance of a comparative analysis of machine learning algorithms. The study of classification rates on samples derived from gene expression data was conducted using several methods of assessment. The research encompassed the utilization of three distinct algorithms over three diverse datasets. The study's findings indicate that the effectiveness of the algorithm is affected by the attributes of the dataset, and datasets containing numerous irrelevant features may result in suboptimal classification outcomes. However, all of the techniques exhibit equal performance in terms of classification accuracy.

In a study with a comparable design to that conducted by Ben-Dor et al. (2020), Dudoit et al. (2022) similarly examined the performance of three algorithms across three distinct gene expression datasets. Their investigation, like the former, centered on the categorization of cancerous tumors. The investigation revealed that the classifiers encountered minimal difficulty in handling two of the datasets, however the third dataset posed a greater challenge. This further supports the conclusion that the classification performance is influenced by the specific properties of the datasets. In the context of algorithmic performance, the study revealed that linear discriminant analysis and nearest neighbour classifiers exhibit somewhat superior performance compared to decision tree classifiers.

The study conducted in Kaminski et al., (2021) considered an evaluation of the performance of five distinct classification algorithms. The evaluation was carried out on three separate DNA microarrays, each associated with cancer outcomes. The study revealed that ensemble approaches, which integrate many classifiers into a unified classifier, have superior performance in terms of classification accuracy.

In their study, Kim et al. (2022) conducted a comparative analysis selected distinct machine learning methods across seven diverse gene expression datasets. The study's overall finding suggests that the choice of gene selection method significantly impacts classification performance. Specifically, it was observed that classical techniques, such as linear discriminant analysis, exhibit strong performance when applied to datasets that have undergone gene selection. However, when considering overall performance, it can be shown that support vector machines exhibit the highest level of performance, regardless of whether gene selection is employed or not. The random forest method closely follows in terms of performance.

Reshan et al., (2023) proposed a system to detect and divide breast cancer into two categories using the Wisconsin Diagnostic Breast Cancer (WDBC). The work aim at using the fewest features to attain the highest accuracy of prediction using multi-model features and ensemble machine learning (EML) techniques. The work incorporates voting, bagging, stacking, and boosting as combination techniques for the classifier. Recursive feature elimination was used to find the most important features. The work was able to distinguish effectively benign breast tumors from malignant cancers.

However, the findings of Kim et al., (2022) were in opposition to those of Prajapati et al. (2022), who concluded that gene selection methods are not significant. Instead, their conclusions align with the earlier studies conducted by Ben-Dor et al. (2020) and Dudoit et al. (2022), which emphasized that the characteristics of the datasets had the most substantial influence on performance. The selection of the machine learning algorithm was deemed to be the primary determinant in attaining a substantial classification accuracy. The majority of methods provide satisfactory performance, with the support vector machine being widely acknowledged as the optimal algorithm.

In their study, Pirooznia et al. (2018) conducted an analysis on a total of eight distinct public datasets derived from microarray investigations. The researchers employed a diverse range of eight machine learning algorithms for their analysis. Various feature selection methods were incorporated, and the models' performance was evaluated both before and after the feature selection process. The study's findings indicate that the accuracy of algorithms is influenced by the nature of the dataset, and the presence of noise in the data is a significant challenge. Additionally, it was shown that the utilization of feature selection had a positive impact on the accuracy of predictions. All models exhibited improved performance when employing the selected feature selection methods; nevertheless, no algorithm emerged as the definitive superior choice. Both support vector machines and random forests consistently demonstrate high performance across all datasets.

Varshali et al., (2023) applied supervised and unsupervised machine learning algorithms with Spark Big data for microarray data classification. The algorithms include K-Nearest Neighbor, Naive Bayes, Decision Trees, Support Vector Machine (SVM), Logistic Regression, and

ensemble models such as AdaBoost, XGBoost, and Random Forest algorithms. The Wisconsin breast cancer diagnostic data was used for testing the proposed method. Performance comparison was carried out based on the accuracy of prediction. The work yield positive findings useful in the clinical context.

In the study carried out by Önskog et al., (2021), the work examined the impact of normalization and gene selection techniques on microarray datasets. Subsequently, they proceeded to evaluate the performance of these techniques across eight distinct datasets. The employed gene selection approaches were exclusively filter-based, and the findings indicated a positive correlation between gene selection utilizing t-statistics and the efficacy of machine learning methodologies. This study additionally corroborated the notion that the efficacy of machine learning algorithms varies among datasets; nonetheless, it revealed that support vector machines continuously exhibit high performance.

In a recent study conducted by Raza and Hasan (2022), a comparative analysis was performed on ten distinct machine learning algorithms. The objective of the study was to identify the most effective algorithm by evaluating its performance on a singular dataset pertaining to prostate cancer. Additionally, the researchers employed t-statistics as the selected approach for feature selection. The researchers discovered that the performance of Bayes Net was superior compared to other widely-used methods such as support vector machine and random forests, which exhibited comparatively lower performance. The primary finding suggests that the efficacy of machine learning algorithms is significantly impacted by the inherent properties or attributes of the dataset upon which the algorithm is employed. This demonstrates that it is recommended to assess machine learning algorithms by conducting tests on multiple datasets, as it is challenging to extrapolate the performance based just on a single dataset.

In the study conducted by Wu et al. (2020), a novel methodology is introduced for the automated detection and diagnosis of brain cancers. The dataset included in this study underwent preprocessing, involving the conversion of MRI images into matrix representations. Subsequently, the MRI images are classified utilizing the probabilistic neural network (PNN) classification technique. The results indicate that the proposed approach achieves a diagnosis accuracy of over 73%. Alur et al., (2023) proposed a methodology for the classification of microarray data using neural networks, specifically focusing on the identification and categorization of brain tumors. The utilization of a neural network-based classifier in this approach yielded an accuracy rate of 83%. In Manikandan, et al., (2023), five machine learning algorithms were used to predict the survival of breast cancer. The algorithms include Naïve Bayes, Decision tree classifier, Ada Boost, XGBoost, and Gradient Boosting classifier. The research classifies the alive and death status of breast cancer patients using the Surveillance, Epidemiology, and End Results (SEER) dataset. It was observed that the Decision Tree algorithm outperforms other machine learning algorithms in terms of classification accuracy.

## 3. Materials and methods

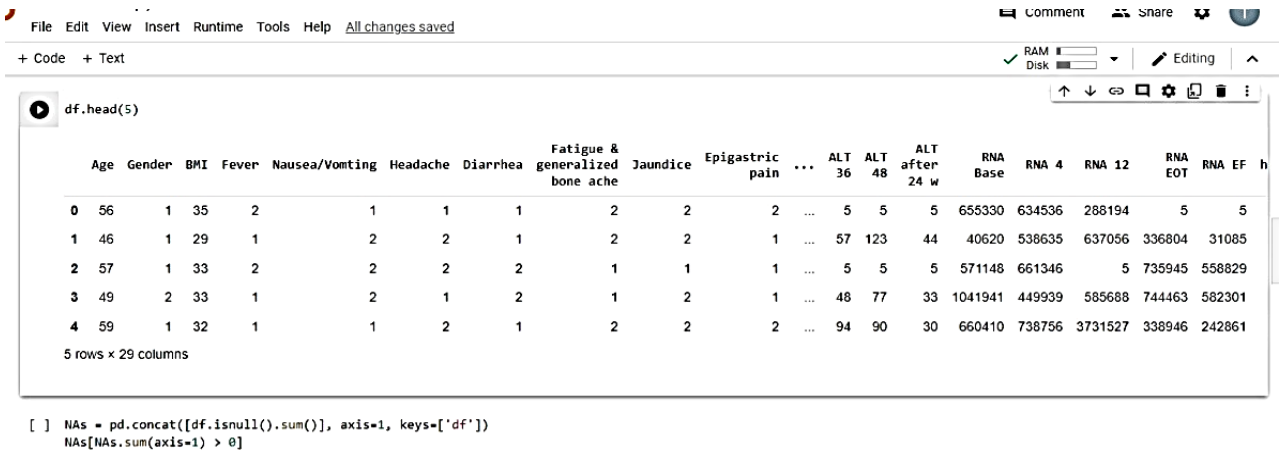
### 3.1 Dataset Acquisition

In this study, the dataset was acquired from a publicly available repository located at https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida, specifically chosen for research purposes. This repository functions as a significant resource within the domain of breast cancer gene expression analysis. The dataset consists of 151 distinct samples, each of which is extensively characterized by a comprehensive set of 54,676 genes. Comprehensive study has categorized the genes into 6 unique classes, each of which serves as a representative of various subtypes of breast cancer. Moreover, the dataset has a distinct category that represents breast tissue in a healthy state, hence facilitating a comparative analysis of gene expressions between healthy and malignant breast tissue. The label distribution of the dataset is shown in table 1 for the 6 classes.

The classification of the dataset and the ample representation of genes being studied present possibilities for employing advanced analytical methods and machine learning models to acquire knowledge regarding breast cancer diagnosis and subtyping. The dataset serves as the fundamental basis for this research, enabling the exploration of data preprocessing, feature selection and classification techniques employed. A sample of the first few rows of the dataset is shown in figure. Additionally, A description of some of the features/variables of the dataset is shown in table 2.

**Table 1: Gene label distribution**

| Gene class label | Number present |
|---|---|
| basal | 41 |
| luminal_B | 30 |
| HER | 30 |
| luminal_A | 29 |
| cell_line | 14 |
| normal | 7 |




| | Age | Gender | BMI | Fever | Nausea/Vomting | Headache | Diarrhea | Fatigue & generalized bone ache | Jaundice | Epigastric pain | ... | ALT 36 | ALT 48 | ALT after 24 w | RNA Base | RNA 4 | RNA 12 | RNA EOT | RNA EF | h |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 56 | 1 | 35 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | ... | 5 | 5 | 5 | 655330 | 634536 | 288194 | 5 | 5 | |
| 1 | 46 | 1 | 29 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | ... | 57 | 123 | 44 | 40620 | 538635 | 637056 | 336804 | 31085 | |
| 2 | 57 | 1 | 33 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | ... | 5 | 5 | 5 | 571148 | 661346 | 5 | 735945 | 558829 | |
| 3 | 49 | 2 | 33 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | ... | 48 | 77 | 33 | 1041941 | 449939 | 585688 | 744463 | 582301 | |
| 4 | 59 | 1 | 32 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | ... | 94 | 90 | 30 | 660410 | 738756 | 3731527 | 338946 | 242861 | |

**Figure 1: Sample first five rows of the dataset**

**Table 2: Dataset feature description**

| Variable | Label |
|---|---|
| Age | Age Sexual orientation at birth (M=male, F=Female) |
| Gender | Sexual orientation at birth (M=male, F=Female) |
| Body mass index | BMI |
| Presence of fever | fever |
| Nausative feeling/vomiting | Nausative/ vomiting |
| Presence of headache | Headache |
| Diarrhoea | Diarrhoea |
| Fatigue and general body ache | Fatigue and general body ache |
| Jaundice | Jaundice |
| Epigastric pain | Epigastric pain |

### 3.2. Data Preprocessing

In this research, data cleaning and normalization was performed. Using Pandas' "isnull()" function, the first phase in the data cleaning process involved checking for any missing values in the dataset. It was confirmed that the dataset has no missing values. The next stage was to convert the categorical data into an integer representation after the data had been cleaned. To do this, the data was fed via the `OrdinalEncoder` from the `sklearn` package. The `OrdinalEncoder` is a preprocessing technique that assigns a unique integer value to each distinct category in the categorical features.

### 3.3. Data Splitting

The dataset was partitioned into two separate sets in order to enable the training and evaluation of the machine learning models in this study. A significant portion, specifically 80%, of the data has been allotted for the purpose of training. This allocation is intended to facilitate adequate training of the machine learning models. During the training phase, the models acquire knowledge about the associations between the characteristics and the class labels, thereby developing the ability to make predictions. The testing dataset is the remaining 20% of the entire dataset. The purpose of the testing set is to evaluate the generalization ability of the models on previously encountered data. Hence the train test split used is arbitrarily chosen as 80/20. The metric offers an assessment of the model's performance in terms of accuracy, precision, sensitivity, and F1 score during the prediction process. The justification for partitioning the dataset into distinct training and testing sets is to assess the model's efficacy in replicating real-world circumstances.

### 3.4 Feature Selection

This section focuses on the fundamental process of feature selection and outlines the utilization of the Chi-square test to maximize this critical part of our research. The Chi-square test is especially appropriate for analyzing categorical variables, hence its adoption for feature selection in this research. In order to conduct the Chi-square test, contingency tables (also known as cross-tabulations) were created for each feature in relation to the

class labels. A summary of the frequency of occurrences for each attribute in relation to the different classes, is shown in table 3.

**Table 3: Contingency Table for "Age" Feature:**

| Age Group | Breast Cancer (Class 1) | Healthy Tissue (Class 2) | Total |
|---|---|---|---|
| Under 30 | 10 | 20 | 30 |
| 30-50 | 25 | 15 | 40 |
| 50 and above | 40 | 5 | 45 |
| Total | 75 | 40 | 115 |

The process of determining the value of the Chi-square statistics involves the comparison of the observed frequencies in the contingency table with the expected frequencies that would be observed if there was no relationship between the feature and the class labels. The specified procedure is as follows:

The expected counts (E) for each cell in the contingency table were calculated assuming independence, suggesting that there is no association between the variable "Age" and the class labels. The value of E is obtained by the process of spreading the total counts in a proportional manner across the various class and age groups.

The formula used to calculate the value of E for the intersection of the categories "Under 30" and "Breast Cancer" is as follows: E = (Total count for "Under 30" * Total count for "Breast Cancer") / Grand Total count. In this case, the calculation is (30 * 75) / 115, resulting in an approximate value of 19.57. The same methodology was employed to compute the value of E for each cell inside the table.

The Chi-square statistic, denoted as $\chi^2$, is computed by applying the following formula: $\chi^2 = \Sigma((O - E)^2 / E)$, where O represents the observed frequency and E represents the expected frequency for each cell in the table.

The $\chi^2$ statistic was calculated for the "Age" variable using observed counts (O) from the contingency table. An example of such an observed count is 10 for the combination of "Under 30" and "Breast Cancer". The aforementioned computation was executed for every individual cell within the table. The p-value was calculated to establish the statistical significance of $\chi^2$. In practical applications, researchers commonly employ a Chi-square distribution table or statistical software to determine the p-value corresponding to the computed $\chi^2$ value. A decrease in p-values signifies a heightened level of correlation between the variable "Age" and the categorical variables "Breast Cancer" and "Healthy Tissue". After the Chi-square test, the characteristics were sorted according to their corresponding p-values. Features that have lower p-values were deemed to have stronger correlations with the classes, making them more pertinent for categorization. The dataset used for training and classification consisted of features that exhibited the strongest statistically significant relationships with the class labels, as determined by the lowest p-values.

### 3.5. Classification Algorithms

The Classification of the microarray data was executed within a Python Integrated Development Environment (IDE). Three distinct boosting algorithms were employed for classification: AdaBoost, Gradient Boost, and X-Gradient Boosting.

## 4. Result and Discussion

The performance of each classifier was assessed using various machine learning model evaluation metrics, which include sensitivity, precision, accuracy and F1 Score.

### 4.1 Result of Algorithm performance

This section, we present the results of our analysis and engage in a detailed discussion of the findings. The performance of each classifier is summarized in table 4 for each of the metrics experimented on.

**Table 4: Performance Comparison of the Classifiers**

| PERFORMANCE METRICS | Adaboost | Gradient Boost | XGBoost |
|---|---|---|---|
| Sensitivity | 90.01 | 90.20 | 96.08 |

| | | | |
|---|---|---|---|
| Precision | 89.41 | 91.10 | 97.25 |
| Accuracy | 91.06 | 93.11 | 98.18 |
| F1 score | 90.03 | 91.62 | 96.82 |
| Training Time (sec) | 42 secs | 38 secs | 11 secs |

The results clearly indicate that XGBoost outperforms both Gradient Boost and Adaboost across all considered metrics. It exhibits superior generalization capabilities and notably faster training times, making it an exceptional choice for the classification of microarray data.

### 4.2. Comparison with existing Algorithms

To assess the competitiveness of the boosting approach, the result obtained was compared with several state-of-the-art machine learning algorithms employed in the classification of the same microarray dataset. The results are presented in table 5.

**Table 5: Comparison of results with Other Algorithms Based on Classification Accuracy**

| Authors | Algorithms/ Methods Used | Accuracy | Training Time |
|---|---|---|---|
| Ben-Dor et al, 2020 | Linear Discrimination + KNN + Decision Tree Classifier | 90.0% | N/A |
| Sung Bae & Hong-Hee, 2019 | Linear Discriminant + Support Vector Machine | 90.93% | N/A |
| Dudoit et al. 2022 | Support Vector Machine | 96.8% | N/A |
| Pirooznia et al. 2018 | Support Vector Machine + Random Forest Algorithm | 93.48% | N/A |
| Onskog et al. 2021 | SVM + KNN + Random Forest + Logistic Regression | 94.7% | N/A |
| Proposed Technique | XGBoost | 98.18% | 11 secs |

The findings of this study clearly indicates that the XGBoost algorithm outperforms other contemporary boosting techniques in terms of accuracy. Specifically, it achieves a noteworthy accuracy rate of 98.18% and in addition, a notably rapid training duration of under 11 seconds.

The notable accomplishment of XGBoost is its remarkable performance, which is shown in the significant value of the precision. In the domain of micro array data analysis, the ability of machine leaning algorithms to continuously generate correct predictions across a wide range of heterogeneous datasets is of paramount importance. Hence, the outcome of this work align with the potential of XGBoost as a robust machine learning method that can significantly impact breast cancer research and detection.

## 5. Conclusion and further work

The analysis and classification of microarray data have become essential tools in the field of early illness diagnosis, specifically in relation to breast cancer. The significant results of this study highlight the effectiveness of the methodology. The XGBoost classifier has exhibited promising capabilities in the timely identification of breast cancer subtypes, as evidenced by its high accuracy of 98.18%, precision of 97.25%, and sensitivity of 96.08%. The aforementioned findings have the potential to enhance the accuracy of diagnoses and, as a result, facilitate the prompt and suitable implementation of medical measures.

The major limitation of this study is the requirement to benchmark on more sophisticated datasets as well as inability to carry out a combined boosting approach.

The result of this study demonstrates the effectiveness of XGBoost. However, it is important to state that the domain of boosting techniques has a diverse range of variations and potential avenues for exploration. Subsequent research endeavors will undertake a more comprehensive investigation of these variations, with the aim of further optimizing the categorization model.

## References

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, N. (2020). Tissue classification with gene expression profiles. Journal of Computational Biology, 7:559-584.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Jr, M. & Haussler, D. (2020). Knowledge-based analysis of microarray gene expression data by using suport vector machines, in ‘Proc. Natl. Acad. Sci.’, Vol. 97, pp. 262–267.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2022). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 97:457, 77–87.

Freund, Y., & Schapire, R. (2019). A short introduction to boosting. J. Japan. Soc. for Artif. Intel., 14:5, 771– 780.

Tan, A. C. & Gibert, D. (2022). 'Ensemble machine learning on gene expression data for cancer classification', Applied Bioinformatics 20(3), s75–s83.

Yeang, C., Ramaswamy, S., Tamayo, P. & *et al.* (2021). 'Molecular classification of multiple tumor types', Bioinformatics 17(Suppl 1), 316– 322.

Risky, F. W. P., Santi, W. P. & Santi, P. R. (2018). Boosting Support Vector Machines for Imbalanced Microarray Data. *INNS Conference on Big Data and Deep Learning. Procedia Computer Science 144. pp. 174–183.*

Prajapati, S., Das, H. & Gourisaria, M. K. (2022). "Microarray Data Classification using Machine Learning Algorithms," OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON), Raigarh, Chhattisgarh, India, 2023, pp. 1-6, doi: 10.1109/OTCON56053.2023.10113990.

Sharmila, V., Ruthra, S., Sathish, V., & Vijay, M. (2019). Booster in High Dimensional Data Classification, International Journal of Engineering Research & Technology (IJERT) RTICCT – 2019 (Volume 7 – Issue 01)

Góngora Alonso, S., Marques, G., Agarwal, D., de la Torre Díez, I., & Franco-Martín, M. (2022). Comparison of Machine Learning Algorithms in the Prediction of Hospitalized Patients with Schizophrenia. Sensors, 22(7). https://doi.org/10.3390/s22072517

Alur V, Raju V, Vastrad B, Vastrad C, Kavatagimath S, Kottur shetti S. (2023). Bioinformatics Analysis of Next Generation Sequencing Data Identifies Molecular Biomarkers Associated With Type 2 Diabetes Mellitus. Clinical Medicine Insights: Endocrinology and Diabetes; 16. doi:10.1177/11795514231155635

Sung S, Ferlay H., Siegel J., Laversanne R. L, Soerjomataram M., Jemal I. A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.

Varsha Nemade, & Vishal Fegade. (2023). Machine Learning Techniques for Breast Cancer Prediction. *International Conference on Machine Learning and Data Engineering*. Procedia Computer Science 218 pp.1314–1320. doi:10.1016/j.procs.2023.01.110.

Varshali J., Praneet S., Umesh K., Mayank P., Sarita S., Surjeet D. (2023). A breast cancer risk predication and classification model with ensemble learning and big data fusion. *Decision Analytics Journal* 8.100298. pp.1-13. https://doi.org/10.1016/j.dajour.2023.100298

Kurian B. and Jyothi V.L, "Comparative Analysis of Machine. (2022). Learning Methods for Breast Cancer Classification in Genetic Sequences. *Journal of Environmental and Public Health*. Pp.1-6. https://doi.org/10.1155/2022/7199290

Yifan D., Jialin. L, and Boxi F., (2021). Forecast model of breast cancer diagnosis based on RF-AdaBoost. In Proceedings of the International Conference on Communications, Information System And Computer Engineering (CISCE), pp. 716–719, IEEE, Beijing, China.

Manikandan P., Durga U. & Ponnuraja C. (2023). An integrative machine learning framework for classifying SEER breast cancer. *Scientific Reports*. 13:5362. https://doi.org/10.1038/s41598-023-32029-1

Reshan M. S. A, Amin S, Zeb MA, Sulaiman A, Alshahrani H, Azar AT, Shaikh A. (2023). Enhancing Breast Cancer Detection and Classification Using Advanced Multi-Model Features and Ensemble Machine Learning Techniques. *Life*.; 13(10):2093. https://doi.org/10.3390/life13102093

Kumar, M.; Singhal, S.; Shekhar, S.; Sharma, B.; Srivastava, G. (2022). Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning. *Sustainability*, 14, 13998