Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

Predicting the ER Status Using the Gene Expression Profiles

Nadia Khizar[1], Khalid Mahmood Aamir[2], Mohamad Deriche[3] , Abdul Jaleel[4]

1. Department of Computer Science, University of Sargodha, Sargodha, Punjab, Pakistan.
2. Department of Information Technology, University of Sargodha, Sargodha, Punjab, Pakistan.
3. College of Engineering and Information Technology, Ajman University, Ajman University, UAE.
4. University of Engineering and Technology, Lahore, Pakistan.

Nadiagondal786@gmail.com khalid.aamir@uos.edu.pk m.deriche@ajman.ac.ae abduljaleel@uet.edu.pk

Corresponding author: **Nadia khizar,** nadiagondal786@gmail.com

Abstract

Breast cancer is the most frequently diagnosed malignancy among women globally, making it the leading cause of cancer incidence worldwide. This type of cancer is particularly challenging due to its high degree of molecular heterogeneity, with many different subtypes exhibiting distinct genetic and biological characteristics. Estrogen receptors (ERs) play a crucial role in breast cancer by acting as receptors for the hormone estrogen, often referred to as the hormone receptor (HR). The presence or absence of ERs, known as ER status, significantly impacts treatment decisions, making accurate prediction of ER status essential for personalized treatment plans. Various methods have been proposed in the literature to predict ER status based on gene expression profiles using microarray genome-wide gene expression profiling. The issue of class imbalance, indicated by a notable variation in the number of samples per class and the curse of dimensionality, is a significant concern with microarray datasets. This paper introduces a novel technique, Fuzzy XGBoost, to predict ER status in breast cancer cases while addressing the class imbalance problem. The proposed algorithm enhances the traditional XGBoost the method by incorporating fuzzy logic principles, improving its robustness, computational power, and accuracy in handling imbalanced datasets. Our research demonstrates the effectiveness of the proposed Fuzzy XGBoost algorithm in accurately predicting ER status across multiple gene expression profile datasets. This approach not only

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

contributes to advancements in breast cancer treatment but also showcases the power of computational methods in bioinformatics. We evaluate our method using four datasets: GSE2990, GSE3494, GSE6532, and GSE7390, achieving predictive accuracies of 100%, 92%, 100%, and 96%, respectively. The successful application of Fuzzy XGBoost in this context highlights its potential for broader use in bioinformatics, where leveraging computational power to address challenges such as imbalanced data is critical for predictive accuracy and personalized treatment.

Keywords: Gene Expression, Breast Cancer, Computational Power, Machine Learning, Fuzzy XGBoost

1-Introduction

Breast cancer (BC) is a deadly disease. According to all the latest figures from the American Cancer Society (ACS), BC will affect one in eight women at some point over their lifetime. The ACS cancer statistics for 2024 show that the incidence of BC was 32%, and its death rate of 15% is very high by the standards for cancer in females [1]. The advent of microarray-based gene expression profiling has substantially furthered our understanding of BC. It is now recognized, rather than as one kind of disease consisting of different histological forms and clinical behaviors in women at the most basic level, that BC is a group of diseases, each characterized by distinct molecular anomalies. Gene expression profiling has revealed that estrogen receptor (ER)-positive and ER-negative breast cancers are fundamentally different diseases from the transcriptomic point of view. Moreover, there may exist further molecular classification within these categories; such differences can be significant in affecting the prognosis for ER-positive breast cancer patients. Based on this understanding, a molecular classification system and prognostic multi-gene classifiers have been developed with microarray technology. These classifiers are currently under test in clinical trials; they are also being incorporated into actual clinical practice [2]. A number of factors can affect a patient's prognosis and how well they respond to cancer treatment: histological grading, tumor kind, and size, lymph node involvement, and cellular receptors on the tumor [3]. Making treatment options needs careful evaluation of cellular receptors such as ER, PR, and HER2. These receptors are the primary focus of the BC molecular subtyping methodology, a categorization method based on conventional immunohistochemical markers [4].

1.1-Why is knowing hormone receptor status important?

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

Breast cancer cells removed during a biopsy or surgery are tested for the presence of specific proteins known as estrogen or progesterone receptors. If these receptors are present, it indicates that the hormones estrogen and progesterone can attach to them, stimulating the cancer to grow. Based on the presence or absence of these receptors, cancers are classified as hormone receptor-positive or hormone receptor-negative. Determining cancer's hormone receptor status is crucial because it guides treatment options. Suppose the cancer cells have one or both of these hormone receptors; hormone therapy drugs can be used to lower estrogen levels or block estrogen from acting on breast cancer cells. This treatment is effective for hormone receptor-positive breast cancers but not for tumors that are hormone receptor-negative (both ER—and PR-negative).

All invasive breast cancers should be tested for both hormone receptors either on the biopsy sample or when the tumor is surgically removed. Approximately three out of four breast cancers have at least one of these receptors [5].

In examining biopsy or surgery samples of BC cells, it is necessary to understand if you have particular proteins called estrogen and progesterone receptors. Tumour growths are speeded up when the estrogen and progesterone receptors come in contact with either hormone. The presence or absence of these proteins determines whether a tumor is classified as HR+ HR-. When selecting a therapy, HR status must be considered. Receptors are proteins that can join with unique substances in the circulation system or reside within or on cell surfaces. Normal breast tissue and BC cells themselves contain receptors that can combine estrogen and progesterone. Both of these hormones are critical for the development and survival of these cells. BC cells may display none, a single one, or all of the following receptors.

- When a breast cancer (BC) contains estrogen receptor, it will be ER+.
- BCs for which breast tumors express progesterone receptors are called PR+ tumors.
- If it is positive for either or both of these receptors, we may call it hormone receptor-positive (HR+).
- Cancer cells without HR do not have any hormone receptors since the cancer cells are not estrogen or progesterone.

During or after a biopsy, hormone receptors should be checked in all cases of invasive BC. One of these receptors is usually present in around 75% of breast tumors [6]. This research study presents a new methodology called the" Fuzzy XGBoost" to predict the ER status of BC. The ER status is a major determining factor for treatment decisions. We used four independent validation datasets, GSE2990, GSE3495, GSE6532, and GSE7390, to test the efficiency of our strategy through predictive analysis. These four publicly available microarray datasets used for this study suffer from the problem of class imbalance. This set of datasets provides broad perspectives on different BC cases, thereby improving the extent and transferability of our

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

results. This is an important finding, given that the prediction of ER status could be informative about BC patients and personalized treatment strategies in real applications. The curse of dimensionality, which involves a large number of genes with a small number of samples, is the main problem with microarray data. In an effort to overcome this constraint, scientists looked at a number of statistical and machine learning classification approaches as well as many feature selection strategies that can identify the most important genes. Another critical concern related to microarray datasets is the emergence of the class imbalance problem, which is characterized by a significant difference in the number of samples per class [7]. So, our study tries to demonstrate whether Fuzzy XGBoost can make correct predictions when testing each dataset based on these efficient relationships. The name XGBoost, though it is abbreviated for Extreme Gradient Boosting. Gradient Boosting is a well-known and widely-used method in machine learning, as it provides a framework for effective and efficient core algorithms and the best machine learning package for regression, classification, and ranking with parallel tree boosting [8]. Initially, a highly predictive gradient-boosting method [9]. It was created due to the sequential structure of training in which each decision tree builds upon the mistakes of its predecessors. Its implementation presented challenges. Even for small models, this method significantly increased training durations. Gradient boosting was revolutionized with the release of XGBoost [10], which used parallel processing across many CPU cores and optimized data organization for quicker lookups. This method improved predicted Accuracy and significantly decreased training durations for the model. While XGBoost fundamentally adheres to gradient boosting, it handles weak learners differently. In contrast to conventional gradient boosting, XGBoost integrates weak learners more effectively. Improved CPU core use in a multi-threaded context leads to greater computational efficiency and performance advantages.

1.2-Main Contribution

- This research introduces "Fuzzy XGBoost," a novel computational technique tailored for predicting estrogen receptor (ER) status in breast cancer patients.
- A comparison between traditional XGBoost and the innovative Fuzzy XGBoost showcases substantial enhancements and benefits achievable through fuzzy logic methodologies.
- Using microarray genome-wide gene expression profiling, the study underscores the computational power of predicting ER status, highlighting its relevance in real-world biological data applications.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

- Addressing the complexity of molecular heterogeneity in breast cancer, the study applies Fuzzy XGBoost to accurately forecast ER status, which is essential for understanding the diverse genetic and biological characteristics among different cancer subtypes.
- The study demonstrates how Fuzzy XGBoost can inform the development of more tailored treatment strategies by precisely identifying an individual's ER status, paving the way for personalized approaches in breast cancer treatment and beyond.

1.3-Paper Organization

The paper is divided into several sections. In the introduction, breast cancer is introduced from a statistical point of view, and the importance of determining estrogen receptor (ER) status is presented. This section also gives an insight into the proposed methodology. In the Related Work section, we review methods currently in use to predict ER status. We also describe the strengths and weaknesses of these methods and identify areas where our research should fill in the gaps they left. The Methodology section briefly introduces the datasets used in this study (GSE2990, GSE3494, GSE6532, and GSE7390) and the procedures of microarray gene expression profiling. Then, our proposed Fuzzy XGBoost algorithm architecture is introduced in detail. We then describe the experimental protocol. The Results and Discussion section details the performance of the Fuzzy XGBoost algorithm on all four datasets, including accuracy percentages and a statistical cross-validation comparison with standard XGBoost. Then, we interpret our results and consider their significance. In the conclusion section, we summarize the significant contributions of this study, including the development of the Fuzzy XGBoost technique and its accurate predictive capabilities for estrogen receptor status. The findings emphasize the effectiveness of these methods in the context of current breast cancer research and highlight their potential impact on personalized treatment strategies.

2-Literature review

Breast cancer research aims to improve how we diagnose and treat the disease, especially by looking at different types, like those involving estrogen receptors (ER). Traditionally, ER status has been determined using a method called immunohistochemistry (IHC). Recently, scientists have started using gene expression profiling to predict ER status. They use machine learning and statistical models for this purpose, but these methods often need more balanced data and complex datasets. This review discusses the progress made in predicting ER status and the new computational techniques developed to overcome these challenges. A comprehensive summary of existing literature on XGboost is presented in Table 1.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

Table 1: Summary of Existing Studies on XGBoost Algorithm.

Ref	Dataset	Methodology	Objective	Accuracy
[11]	GSE162726	XGBoost	BC biomarker prediction	0.82
[12]	GEO, GTEx	XGBoost	Gene expression prediction	91.5%
[13]	TCGA, GEO	XGBoost	Improved lesion diagnosis	96.38%
[14]	Genes dataset	AdaBoost, XGBoost	BC Metastasis prediction	95%,90%
[15]	SEER database	XGBoost	PBone metastasis prediction	0.98

Gene expression profiling has been extensively employed for characterizing cell status, reflecting bodily health, diagnosing genetic disorders, and more. Despite the gradual reduction in genome-wide expression profiling costs in recent years, collecting expression profiles for numerous genes remains financially demanding. By examining the values of the 943 landmark genes, it became possible to forecast the expression levels of the remaining target genes, given the average strong correlation among gene expressions in humans. As a result, the author created an algorithm to forecast the levels of gene expression in this study. The technique, which combined numerous tree models and offered improved interpretability, was based on XGBoost. By employing the GEO dataset in combination with an RNA-seq dataset, the efficacy of the XGBoost model was assessed and compared against the performance of other contemporary models. Through an evaluation using test data, the XGBoost model demonstrated superiority over the older D-GEX method, LR, and K-nearest neighbors (KNN) techniques. Notably, predictions derived from the XGBoost model exhibited a notably diminished total error, underscoring its enhanced performance [12].

Through an integrated learning approach, a suitable machine-learning model was developed to discern primary lesions in cases of primary metastatic tumors. This innovation enhanced Accuracy and diagnostic efficacy in identifying the source of primary lesions. Subsequently, features with expression levels lower than the defined threshold were eliminated. Two methods were employed for feature selection, followed by utilizing XGBoost for classification. The outcomes of this study indicated that each cancer type exhibited its corresponding classification accuracy through the amalgamation of tumor data and machine learning techniques. This Accuracy might be used to forecast where initial metastatic tumors would expand. This machine-learning-centered approach could serve as an alternative diagnostic method, supplementing the assessment of machine-learning model processing alongside clinical

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

pathological conditions. After the optimal model was chosen through a 10-fold cross-validation process, its validity was confirmed using an independent test set [13].

The study encompassed a cohort of 97 (BC) patients, with 46 individuals (or 47%) exhibiting distant metastases and 51 individuals (or 53%) without such metastases. The research focused on assessing the expression levels of 24,481 genes within these subjects. Through the combined utilization of Boruta and LASSO techniques, researchers successfully pinpointed biomarker genes associated with distant metastasis in BC patients—statistical analysis involved employing the Mann-Whitney U test to evaluate gene expression variations between the two groups. Effect sizes (Cohen's d) and odds ratios were also calculated. Two examples of tree-based algorithms—AdaBoost and XGBoost—were employed to predict the occurrence of distant metastases in BC. A comprehensive evaluation of performance metrics facilitated comparing study findings and assessing prediction model effectiveness [14].

This retrospective research examined a total of 283,373 BC patients using data from the Surveillance, Epidemiology, and End Results (SEER) database. Multivariate logistic regression (LR) was adopted in the analysis to identify factors contributing to bone metastasis risk among BC patients. Regression with Cox proportional hazards was also used to identify predictive factors correlated with bone metastatic BC (BCBM). Six machine-learning algorithms were used to build diagnostic and prognostic models that compared the identified risk and prognosis parameters. Among these models, the best performance for both diagnosis and prognosis was achieved by the XGBoost method, which produced an area under the ROC curve (AUC) of 0.98 for diagnosis and an AUC of 0.88 for prognosis [15].

The research underscored the applicability of Recursive Feature Selection-Random. Forest is a feature selection technique in datasets, such as gene expression profiles, which exhibit a substantial number of features concerning the sample size. Upon assessing the model's performance, it became evident that XGBoost surpassed the other four machine-learning approaches in subsets consisting of 25 and 20 genes, both across the complete dataset and the driver dataset [16].

Machine learning models have been extensively employed in computer-aided prognostic systems to predict the spread of BC. However, these systems still have a number of difficulties. To address these problems, they have suggested a comprehensible method for forecasting the spread of BC from clinicopathological information. Their process involves training a CatBoost classifier with cost sensitivity and using the LIME explainer for patient-level explanations. They tested their approach on a publicly available dataset of 716 BC patients. The results indicate that the cost-sensitive CatBoost model is better than conventional models and boosts

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

algorithms when considering metrics such as accuracy (76.5%) and recall (79.5%), with an F1 score of 77% [17].

3-Materials and methods

In this section, we will present the proposed methodology in detail. We started the method by acquiring publicly available microarray gene expression profile datasets. The dataset was collected from gene expression omnibus (GSE3494, GSE2990, GSE6532, GSE7390). Details of each dataset are presented in this section for more information. After dataset acquisition, we preprocessed the data by deleting metadata from the datasets. After preprocessing, we split the data into training and testing. Then, we performed fuzzy encoding on the datasets. The steps of fuzzy encoding are also discussed in detail in this section. We also presented our proposed methodology in this section. This study introduces a novel approach named Fuzzy XGBoost. We applied this approach to predict BC's estrogen receptor (ER) status using four distinct gene expression profile datasets. The proposed method is visually depicted in Figure 1. Detailed information about each dataset, as well as the methodology, is provided in this section. The summary of all datasets is shown in Table 2. It's worth noting that all the datasets used in this study are publicly accessible.

Machine learning algorithms may need help dealing with unbalanced datasets. Specialized methods must be used when handling such datasets to prevent the model from being biased in favor of the dominant class. In the literature, several strategies, including data level, algorithm level, and a combination of data and algorithm level, have been put out to address the imbalance data issue [18]. There are three categories for data level-based procedures (resampling techniques): oversampling, under-sampling, and hybrid approaches, combining both oversampling and under-sampling [19]. However, the imbalance problem is addressed by algorithm-based solutions for unbalanced datasets, which permit the assignment of distinct class weights to penalize the minority class's misclassification more severely [20]. Another family of computational strategies for addressing the problem of class imbalance is ensemble learning. By integrating many models, they can increase the model's capacity to handle unbalanced data [21]. At the data level, the issue of excessive imbalance in this study is lessened by the creation of synthetic samples for the minority class. The synthetic minority oversampling method, or SMOTE [22], is an algorithm created to address the problem of unbalanced datasets. In this study, we have used ensemble learning to overcome unbalanced issues.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

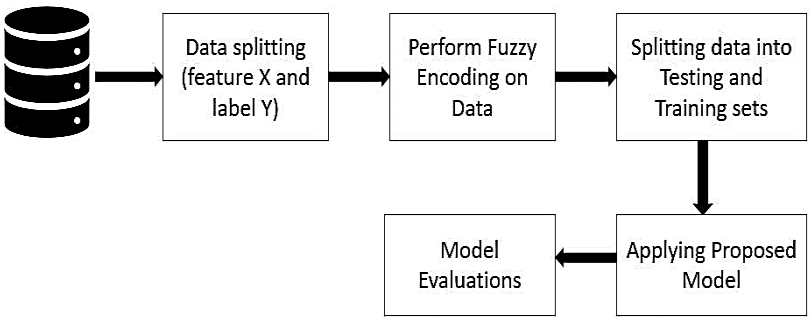

Figure 1. Proposed Methodology flow chart.

3.1-Dataset description

The dataset was collected from gene expression omnibus (GSE3494, GSE2990, GSE6532, GSE7390). A comprehensive description of all datasets is presented in Table 2. Table 3 presents the total number of samples, class distribution of majority and minority classes of ER status, and missing values. The Details of each dataset are presented as follows:

3.1.1-GSE2990

The Gene Expression Omnibus (GEO) database of the National Centre for Biotechnology Information (NCBI) makes the dataset easily accessible. This database serves as a repository for functional genomics data, encompassing a collection of valuable resources [23]. The histologic grade of BC is crucial for prognostication. However, 30% to 60% of tumors are categorized as grade 2, which doesn't reliably indicate the likelihood of recurrence. It was examined whether the histologic grade of breast tumors correlates with gene activity and whether this correlation might improve grading precision [24].Dataset Source link https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2990.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

Table 2. Comprehensive information of all datasets.

Accession No	Name	Platform	Status	Samples	Features	Types	Repository
GSE2990	Understanding the Molecular Basis of Histologic Grade To Improve Prognosis	HG-U133A	Publicly available	189	22283	Numeric	GEO
GSE3494	Molecular characterization of the tumor microenvironment in BC	HGU133A, HU-U133B	Publicly available	502	22283	Numeric	GEO
GSE6532	Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas using genomic grade	HGU133A, HU-U133B, HU-U133 plus 2.0 Array	Publicly available	741	22283	Numeric	GEO
GSE7390	Strong Time Dependence of the 76-Gene Prognostic Signature	HG-U133A	Publicly available	198	22283	Numeric	GEO

3.1.2-GSE3494

You may view the dataset in the GEO database for free at this link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse3494, which is a directory of functional genomics data supported by the (NCBI). The GSE3494 dataset contains gene expression data from 251 breast tumor patients and clinical data like age, lump size & hormone receptor (HR), p53 status, and patient survival [25] [26]. Based on the gene expression patterns, three subtypes of samples were separated: luminal A, luminal B, and basal-like.

3.1.3-GSE6532

The Gene Expression Grade Index improved grading precision and predictive value by grouping patients into subgroups with increased and decreased recurrence chances, notably in Grade 2 tumors [27]. The examination of 64 primary breast tumor microarray tests for gene identification revealed its potential as a critical ally in the fight against BC. With the aid of an

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

additional 129 trials, this was further confirmed. Together, this research showed how important and applicable it is [28] [29]. Genomic grading may successfully distinguish between two separate subgroups of ER+ BC by employing a straightforward and consistently repeatable procedure across several datasets. This study highlights the significance of cell proliferation-related genes in determining the prognosis of ER+ BC. The GEO database, which the NCBI describes as a comprehensive collection of functional genomics data, hosts the dataset and makes it freely available to all users. Dataset Source link https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse6532.

3.1.4-GSE7390

The dataset is freely accessible on the GEO database, which the NCBI describes as a directory of functional genomics data. Independent research has shown that the 76-gene signature can forecast the likelihood of early distant metastases in BC patients who do not have lymph node involvement. The clinical value of gene expression patterns in boosting risk assessment and treatment options is further highlighted by this TRANSBIG study, underscoring their usefulness for improving patient care. Significance of incorporating these signatures into accepted clinical practice [30] [31]. Dataset Source link https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse7390.

Table 3. Dataset distribution details: majority and minority class distributions of the datasets and missing values in each dataset.

Dataset	ER samples (Total)	ER+	ER-	Missing values
GSE2990	146	96	43	7
GSE3494	251	213	34	4
GSE6532	263	200	45	18
GSE7390	198	134	64	Nill

3.2-Proposed Methodology

Our proposed methodology includes several important steps, starting with selecting the right dataset and evaluating how well our algorithm performs. First, we gathered the necessary datasets. After acquiring them, we preprocessed the data by deleting metadata from the data, extracting the features and labels from the datasets, and then splitting the data into training and testing sets. We applied fuzzy encoding, which helps manage any uncertainties in the data. The steps for fuzzy encoding are explained below. Next, we used the XGBoost ensemble learning technique. XGBoost is great at combining predictions from multiple models to boost accuracy.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

By adding fuzzy logic to XGBoost, our method aims to tackle the challenges of imbalanced data, high dimensionality, and how computational power increases. Table 4 provides an overview of our proposed algorithm. It lays out each step and technique we used, showing how the proposed methodology was used to predict ER status accurately.

3.2.1-Fuzzy encoding

- Normalize the features using a scaling function: This initial step involves adjusting the scale of input features to ensure that all features contribute equally to the analysis, preventing any single feature from dominating due to its scale.
- Conduct Principal Component Analysis (PCA) with 10 components on the normalized features: PCA transforms the data into a new coordinate system where the axes align with the directions of maximum variance. This transformation generates a new set of variables, known as principal components, aimed at capturing the most significant information within the data. Here, 10 principal components are retained for subsequent analysis.
- Perform k-means clustering with 2 clusters on the PCA-transformed features: K-means clustering partitions data points into clusters based on their similarity. In this case, the Paraformer features undergo k-means clustering with a predefined number of clusters set to 2, effectively dividing the data into two distinct groups.
- Calculate fuzzy memberships using pairwise distances between data points and cluster centroids: Fuzzy memberships quantify the degree to which data points belong to each cluster. These memberships are computed using the distances between data points and the clusters' centroids. Fuzzy logic allows for partial memberships, indicating varying degrees of belongingness to different clusters.
- Rearrange the cluster indices to derive fuzzy labels: In the final step, the cluster indices obtained from k-means clustering are reshaped to create fuzzy labels. This reshaping incorporates fuzzy memberships, providing a representation of the uncertainty or partial membership of data points within each cluster.

3.2.2 Proposed Algorithm

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

Algorithm

Step 1 Data Preparation:
We separated each dataset's features (X) and labels (Y).
Step 2 Fuzzy Encoding:
We performed Fuzzy Encoding, which involves a series of steps:
Scaling the Data: Adjusting data values to work well together.
Reducing Dimensions using PCA: Finding key features in the data.
Grouping Similar Data with K-means: Creating clusters of similar data.
Measuring Membership using pdist2: Calculating how closely data belongs to clusters.
Assigning Labels by reshaping: Labeling data based on clusters.
Step 3 Data Splitting:
Split the fuzzy-encoded data into training and testing sets:
Set a training ratio of typically 0.7 (70% for training, 30% for testing).
Step 4 Model Training: We trained the Fuzzy XGBoost algorithm using the fit ensemble model (GentleBoost, AdaBoostM2, LogitBoost, etc.).
Step 5 Model Evaluation: We evaluate our proposal using accuracy, precision, recall, and F1 score.

4-Results and Discussion

This section will present the outcomes of our proposed methodology's implementation. We employed four distinct gene expression profile datasets for our analysis. Initially, we divided the datasets into training and testing sets, and then we applied the traditional XGBoost algorithm to assess its performance. The results of the conventional XGBoost classifier are summarized in Table 5. Most of the results are biased due to data imbalanced issues. We present the confusion matrix on traditional XGBoost for each dataset to evaluate the algorithm's performance. The confusion matrix on GSE2990 using XGBoost is presented in Figure 2. The confusion matrix on GSE3494 using XGBoost is presented in Figure 3. The confusion matrix on GSE6532 using XGBoost is presented in Figure 4. The confusion matrix on GSE7390 using XGBoost is presented in Figure 5. Subsequently, we evaluated the efficacy of our proposed Fuzzy XGBoost algorithm. First, we have divided labels (Y) and features (X) for the research dataset. The second is the Fuzzy perfume encoding. We have perfumed numerous phases in undefined encoding. 1) Data scaling,2) PCA Dimension Reduction, 3) Using K-means to Group Similar Data Measurement of Membership using pdist2.5) Label assignment through reshape.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

Third, datasets should be divided into 30% and 70% for testing and training. Fourth, train the Fuzzy XGBoost on the fit-ensemble algorithm. The datasets were partitioned into 70% training and 30% testing subsets, ensuring a stratified split based on the outcome. The confusion matrices for the proposed Fuzzy XGBoost model are shown in Figures 6 through 9. Specifically, Figure 6 presents the results for the GSE2990 dataset, Figure 7 for the GSE3494 dataset, Figure 8 for the GSE6532 dataset, and Figure 9 for the GSE7390 dataset. For the GSE2990 dataset, we trained the classifier using LogitBoost and GentleBoost with 100 trees and a learning rate of 0.01. The classifier for GSE3494 was trained using LogitBoost and GentleBoost, 140 trees, with a learning rate of 0.1. Similarly, for GSE6532, the classifier utilized GentleBoost with 100 number trees and a learning rate of 0.01. Lastly, concerning GSE7390, the classifier underwent training with LogitBoost and GentleBoost, using 100 numbers trees and a learning rate of 0.1. In the case of the GSE2990 dataset, Fig 10 illustrates both Accuracy and average Accuracy using Fuzzy XGBoost. Similarly, for the GSE3494 dataset, Fig 11 portrays the Accuracy and average Accuracy using Fuzzy XGBoost. The utilization of the GSE6432 dataset results in Fig 12, which demonstrates Accuracy and average Accuracy using Fuzzy XGBoost. Then, when considering the GSE7390 dataset, Fig 13 presents the Accuracy and average Accuracy using Fuzzy XGBoost.

Table 4. Performance Analysis of Traditional XGBoost Algorithm on Microarray Datasets.

Datasets	Accuracy	Precision	Recall	F1 score
GSE2990	68.29%	60.14%	54,12%	51.76%
GSE3494	88.00%	65.71%	60.26%	62.10%
GSE6532	90.54%	81.88%	79.44%	80.58%
GSE7390	86.44%	85.28%	78.92%	81.27%

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

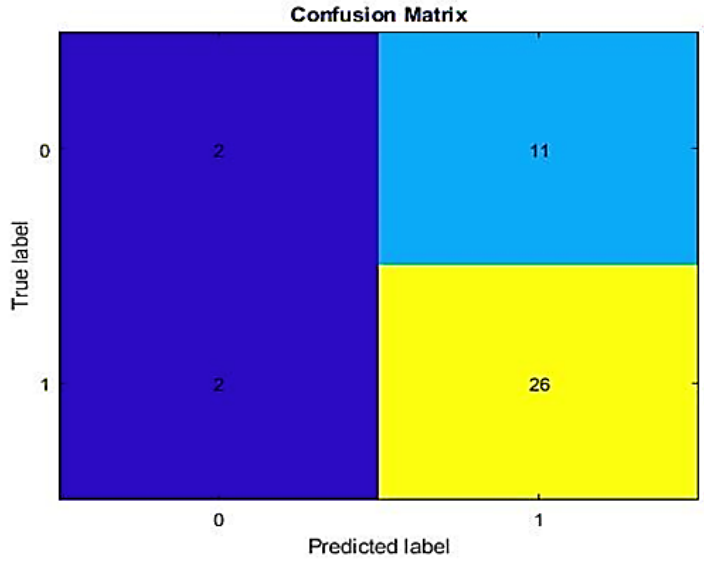

Figure 2. The figure shows the confusion matrix of the GSE2290 dataset on the traditional XGBoost Algorithm.

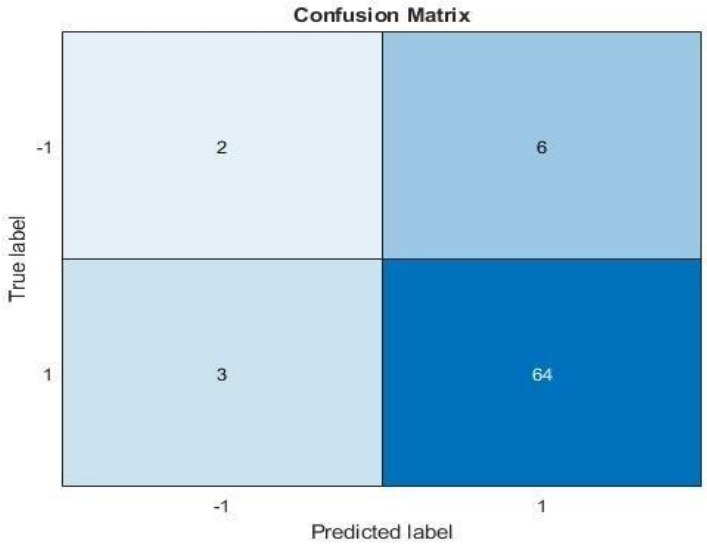

Figure 3. The figure shows the confusion matrix of the GSE3494 dataset on the traditional XGBoost Algorithm.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

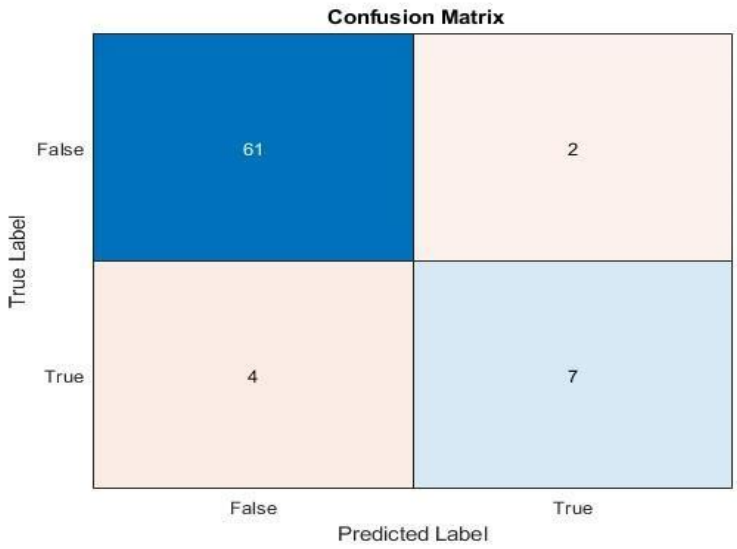

Figure 4. The figure shows the confusion matrix of the GSE6532 dataset on the traditional XGBoost Algorithm.

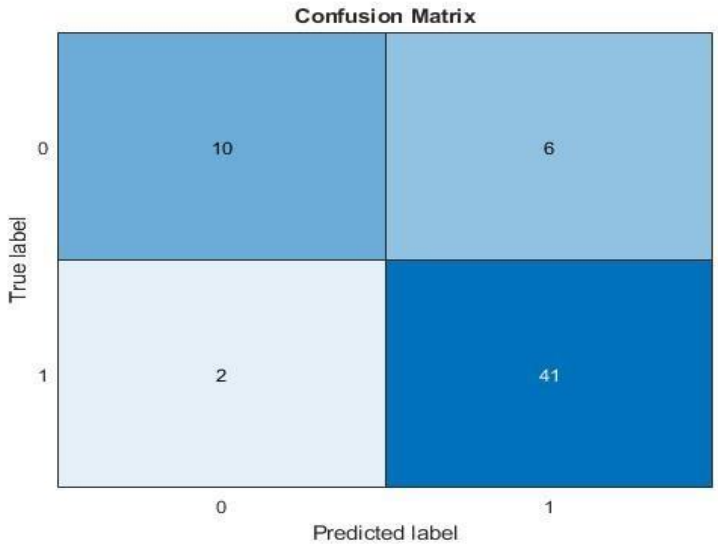

Figure 5. The figure shows the confusion matrix of the GSE7390 dataset on the traditional XGBoost Algorithm.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

Table 5. Performance Analysis of our Proposed Fuzzy XGBoost Algorithm on Microarray Datasets.

Datasets	Accuracy	Precision	Recall	F1 score
GSE2990	100	1	1	1
GSE3494	94.66	1	0.89	0.94
GSE6532	100	1	1	1
GSE7390	96.61	0.94	1	0.97

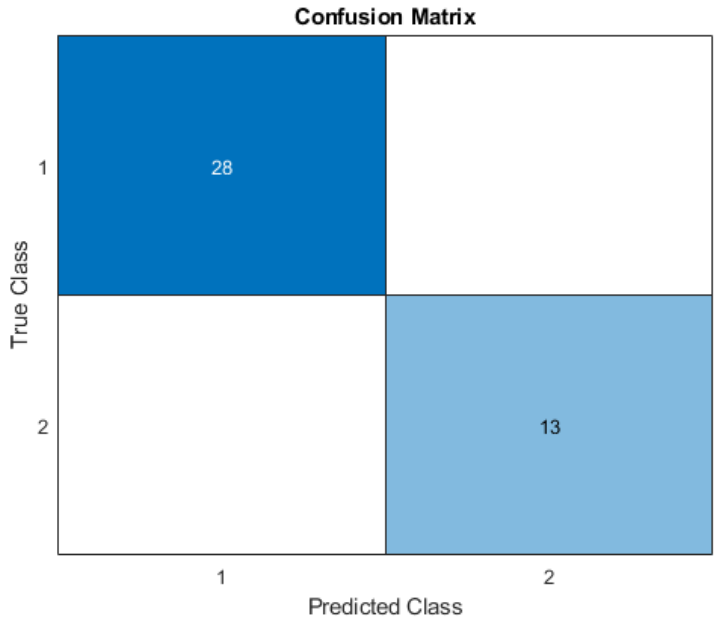

Figure 6. The figure shows the confusion matrix of the GSE2290 dataset on the Proposed Fuzzy XGBoost Algorithm.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

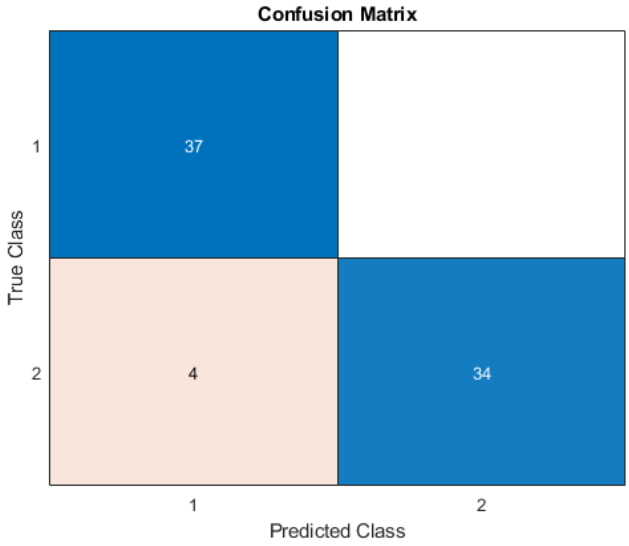

Figure 7. The figure shows the confusion matrix of the GSE3494 dataset on the Proposed Fuzzy XGBoost Algorithm

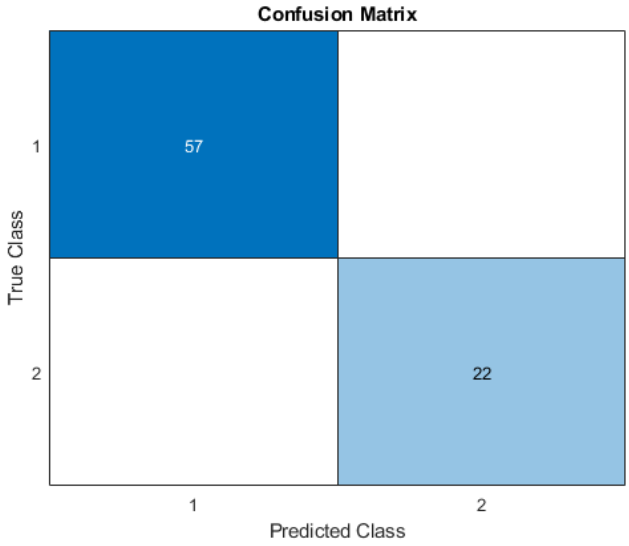

Figure 8. The figure shows the confusion matrix of the GSE6532 dataset on the Proposed Fuzzy XGBoost Algorithm.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

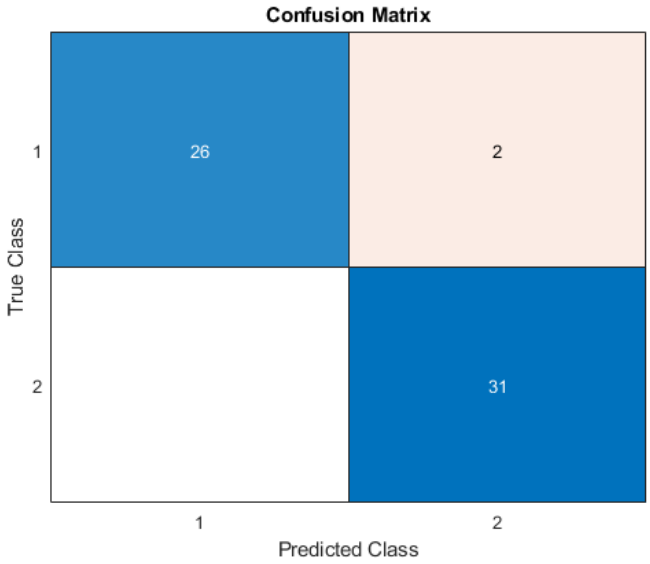

Figure 9. The figure shows the confusion matrix of the GSE7390 dataset on the Proposed Fuzzy XGBoost Algorithm.

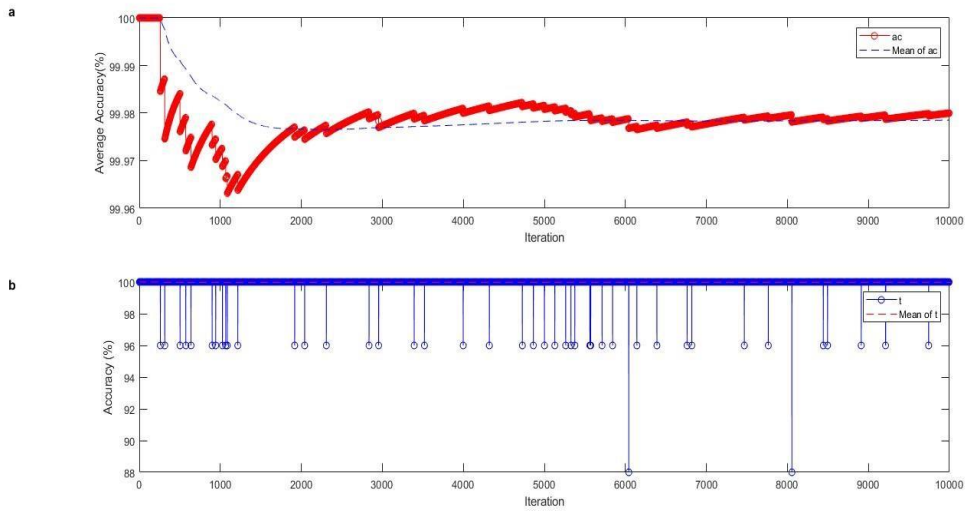

Figure 10. Accuracy on the GSE2290 dataset using the proposed Fuzzy XGBoost algorithm a) represents average accuracy on each iteration, and **b)** represents the accuracy on each iteration. The algorithm has been executed 10,000 times.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

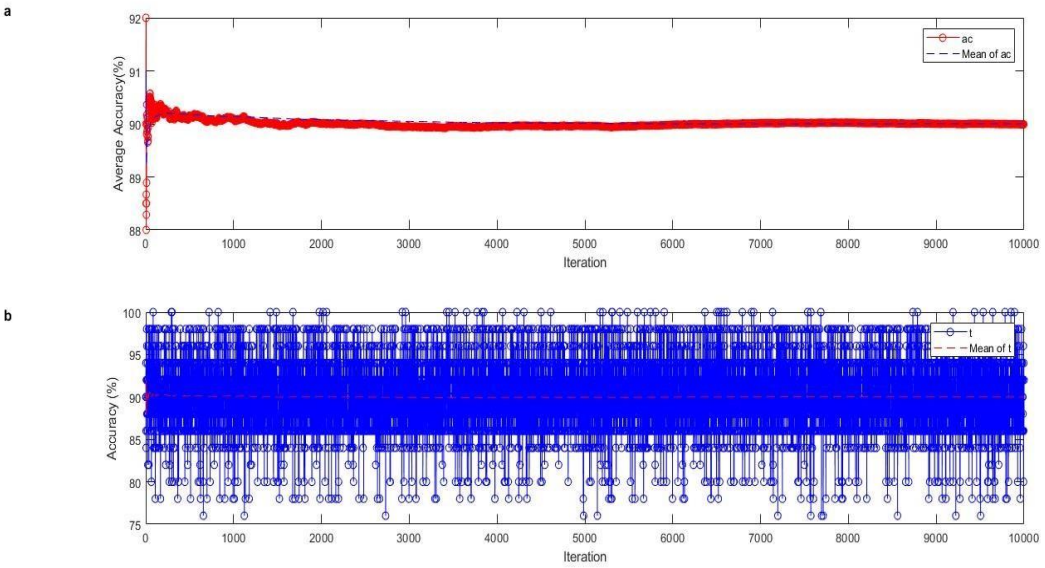

Figure 11. Accuracy on the GSE3494 dataset using the proposed Fuzzy XGBoost algorithm a) represents average accuracy on each iteration, and **b)** represents the accuracy on each iteration. The algorithm has been executed 10,000 times.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

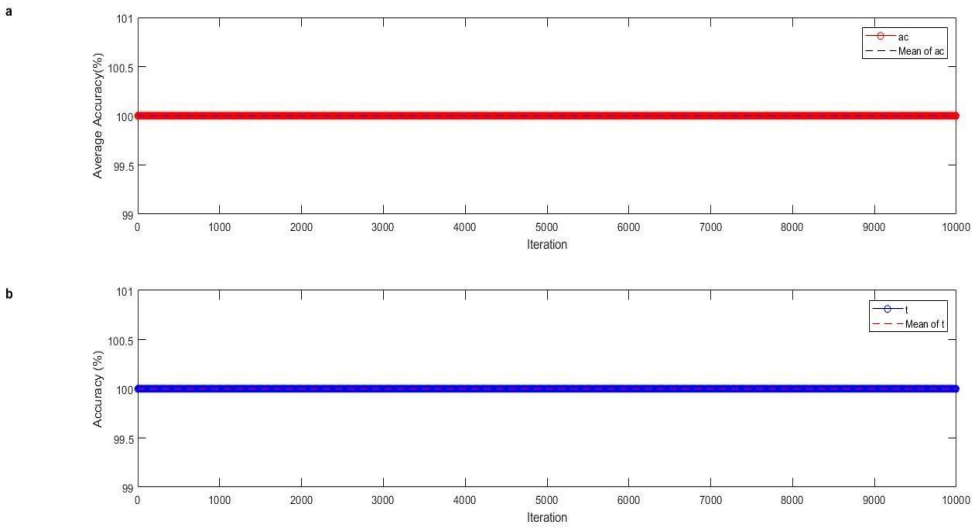

Figure 12. Accuracy on the GSE6532 dataset using the proposed Fuzzy XGBoost algorithm a) represents average accuracy on each iteration, and **b)** represents the accuracy on each iteration. The algorithm has been executed 10,000 times.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

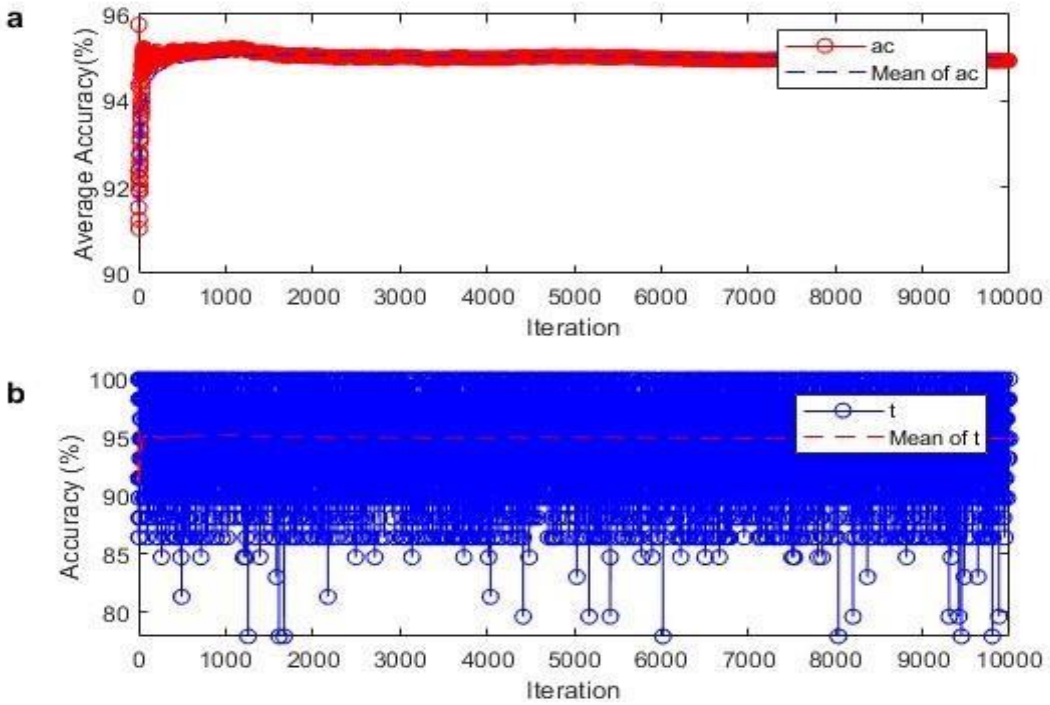

Figure 13. Accuracy on the GSE7390 dataset using the proposed Fuzzy XGBoost algorithm a) represents average accuracy on each iteration, and **b)** represents the accuracy on each iteration. The algorithm has been executed 10,000 times.

5-Conclusion

In conclusion, ERs play a pivotal role in BC; they can be the deciding factor for treatment choice as well as a modifier of prognosis. The proposed Fuzzy XGBoost method reported in this study provided an accurate prediction of ER status for BC cases by incorporating fuzzy logic theories into an XGBoost framework. By doing so, the approach was found to have potential value across different datasets, raising the standard of accuracy for predicted ER status. This method harnesses the power of advanced computational techniques and modern technology to deliver precise and reliable results, demonstrating the importance of personalized BC treatment strategies where a person's ER status can inform individualized therapy.

Our comparison between the proposed Fuzzy XGBoost and traditional XGBoost highlights that integrating fuzzy logic into the processing model can increase accuracy. The proposed Fuzzy

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

XGBoost method was validated across various datasets, including GSE2990, GSE3495, GSE6532, and GSE7390, which are all imbalanced yet consistently demonstrated high prediction accuracy. These results further underscore that Fuzzy XGBoost is a reliable tool for clinical decision-making across different contexts and a robust predictive model in this field.

Accurate determination of ER status is essential for a good prognosis and response to treatment in BC. It significantly impacts the long-term survival of patients suffering from this disease. Several factors, such as histological grade, tumor type and size, lymph node involvement, and receptor expression, all contribute to this. Therefore, the proposed Fuzzy XGBoost can be considered a unique methodology for predicting the ER status of such patients, potentially transforming existing approaches to treating BC with personalization.

Leveraging computational power and cutting-edge technology, this method enhances the quality of patient care, aids in better treatment choices, and deepens the understanding of the complex nature of hormone receptors and BC development. These findings suggest that the proposed method could revolutionize the approach to personalized BC treatment.

References

[1] D. S. Dizon, "Cancer statistics 2024: All hands on deck," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 2, pp. 12-49, 2024.

[2] J. S. Reis-Filho and L. Pusztai, "Gene expression profiling in breast cancer: classification, prognostication, and prediction," *The Lancet*, vol. 378, no. 9805, pp. 1812-1823, 2011.

[3] F. K. Al-Thoubaity, "Molecular classification of breast cancer: A retrospective cohort study," *Annals of Medicine and Surgery*, vol. 49, pp. 44-48, 2020.

[4] A. Goldhirsch, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, and H.-J. Senn, "Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011," *Annals of Oncology*, vol. 22, no. 8, pp. 1736-1747, 2011.

[5] American Cancer Society, "Breast cancer hormone receptor status," *American Cancer Society*, 2023.

[6] American Cancer Society, "Understanding a Breast Cancer Diagnosis: Breast Cancer Hormone Receptor Status," *American Cancer Society*, 2023.

[7] C. J. Lin WJ, "Class-imbalanced classifiers for high-dimensional data," *Briefings in Bioinformatics,* vol. 14, no. 1, p. 13–26, 2013.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

[8] T. Chen *et al*., "Xgboost: extreme gradient boosting," *R package version* 0.4-2, vol. 1, no. 4, pp. 1-4, 2015.

[9] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics,* vol. 29, no. 5, p. 1189–1232, 2001.

[10] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[11] Q. Li, H. Yang, P. Wang, X. Liu, K. Lv, and M. Ye, "XGBoost-based and tumor-immune characterized gene signature for the prediction of metastatic status in breast cancer," *Journal of Translational Medicine*, vol. 20, no. 1, p. 177, 2022.

[12] W. Li, Y. Yin, X. Quan, and H. Zhang, "Gene expression value prediction based on XGBoost algorithm," *Frontiers in Genetics*, vol. 10, p. 1077, 2019.

[13] S. Chen *et al*., "A novel XGBoost method to infer the primary lesion of 20 solid tumor types from gene expression data," *Frontiers in Genetics*, vol. 12, p. 632761, 2021.

[14] S. Akbulut, F. H. Yagin, and C. Colak, "Prediction of breast cancer distant metastasis by artificial intelligence methods from an epidemiological perspective," *Istanbul Medical Journal*, vol. 23, no. 3, pp. 210-215, 2022.

[15] X. Zhong, Y. Lin, W. Zhang, and Q. Bi, "Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning," *Scientific Reports*, vol. 13, no. 1, p. 18301, 2023.

[16] A. Thalor, H. K. Joon, G. Singh, and S. Roy, "Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer," *Computational and Structural Biotechnology Journal*, vol. 20, p. 3, 2022.

[17] I. Maouche, L. S. Terrissa, K. Benmohammed, and N. Zerhouni, "An explainable AI approach for breast cancer metastasis prediction based on clinicopathological data," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 12, pp. 3321-3329, 2023.

[18] N. Bhandari, R. Walambe, K. Kotecha, and S. P. Khare, "A comprehensive survey on computational learning methods for analysis of gene expression data," *Frontiers in Molecular Biosciences*, vol. 9, p. 907150, 2022.

[19] K. Upadhyay, P. Kaur, and D. K. Verma, "Evaluating the performance of data level methods using KEEL tool to address class imbalance problem," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 9741-9754, 2022.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

[20] M. Zheng, F. Wang, X. Hu, Y. Miao, H. Cao, and M. Tang, "A method for analyzing the performance impact of imbalanced binary data on machine learning models," *Axioms*, vol. 11, no. 11, p. 607, 2022.

[21] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *International Journal of Distributed Sensor Networks*, vol. 18, no. 6, p. 15501329221106935, 2022.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[23] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207-210, 2002.

[24] C. Sotiriou *et al*., "Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262-272, 2006.

[25] L. D. Miller *et al*., "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *Proceedings of the National Academy of Sciences*, vol. 102, no. 38, pp. 13550-13555, 2005.

[26] K. D. Sullivan, M. D. Galbraith, Z. Andrysik, and J. M. Espinosa, "Mechanisms of transcriptional regulation by p53," *Cell Death & Differentiation*, vol. 25, no. 1, pp. 133-143, 2018.

[27] S. Loi *et a*l., "Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade," *Journal of Clinical Oncology*, vol. 25, no. 10, pp. 1239-1246, 2007.

[28] S. Loi *et al*., "Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen," *BMC Genomics*, vol. 9, pp. 1-12, 2008.

[29] S. Loi *et al*., "PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor–positive breast cancer," *Proceedings of the National Academy of Sciences*, vol. 107, no. 22, pp. 10208-10213, 2010.

[30] C. Desmedt *et al*., "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clinical Cancer Research*, vol. 13, no. 11, pp. 3207-3214, 2007.

Power System Technology
ISSN:1000-3673

Received: 15-05-2024 Revised: 12-06-2024 Accepted: 25-07-2024

[31] P. Patil, P.-O. Bachant-Winner, B. Haibe-Kains, and J. T. Leek, "Test set bias affects reproducibility of gene signatures," *Bioinformatics*, vol. 31, no. 14, pp. 2318-2323, 2015.