# SMOTE-ENN-LR: LEVERAGING MACHINE LEARNING FOR BREAST CANCER CLASSIFICATION IN MICROARRAY GENE EXPRESSION WITH EXPLAINABLE AI

*Md Faisal Bin Abdul Aziz*[1,2], *Azree Nazri*[1]*, *Fatematuz Zuhura Evamoni*[3], *Razali Yaakob*[1], *Teh Noranis Mohd Aris*[1], *Zamberi Sekawi*[4], *Tanjim Mahmud*[5,6], *Olalekan Agbolade*[1], *Wajid Syed*[7], *Mohamed N Al Arifi*[7]

[1]Department of Computer Science, Universiti Putra Malaysia, Serdang, Malaysia

[2]Department of Computer Science and Engineering, Comilla University, Bangladesh

[3]Dept. of Biotechnology and Genetic Engineering, Noakhali Science and Technology University, Bangladesh

[4]Department of Medical Microbiology, Universiti Putra Malaysia, Serdang, Malaysia

[5]Department of Computer Science and Engineering, Rangamati Science and Technology University, Bangladesh

[6]Graduate School of Engineering, Kitami Institute of Technology, Kitami, Hokkaido, Japan

[7]Department of Clinical Pharmacy, College of Pharmacy, King Saud University, Saudi Arabia

Emails: faisal@cou.ac.bd[1,2], azree@upm.edu.my[1*], fatematuz@nstu.edu.bd[3], razaliy@upm.edu.my[1], nuranis@upm.edu.my[1], zamberi@upm.edu.my[4], tanjim_cse@yahoo.com[5,6], lokaprof@yahoo.com[1], wali@ksu.edu.sa[7], malarifi@ksu.edu.sa[7]

## *ABSTRACT*

*Breast cancer continues to be a major public health issue worldwide, ranking as the second leading cause of cancer-related deaths among women. Effective early detection and classification are crucial for improving survival rates, yet they are complicated by the challenges posed by imbalanced datasets in microarray gene expression analysis. These imbalances can significantly affect the predictive power and reliability of traditional classification models, underscoring the need for more sophisticated analytical techniques. This study introduces an approach, the SMOTE-ENN-LR method, which combines the Synthetic Minority Over-sampling Technique (SMOTE) with Edited Nearest Neighbors (ENN) for noise removal and Logistic Regression (LR) to accurately classify breast cancer based on microarray data. The SMOTE technique is utilized to over-sample the minority cases in the dataset, thereby addressing the issue of underrepresentation. Simultaneously, the ENN method is employed to clean the data by removing mislabeled instances and noise, which are often prevalent in over-sampled datasets. The cleaned and stable dataset is used to train a LR model, optimizing its ability to discern between cancerous (Abnormal) and non-cancerous (Normal) gene expression profiles effectively. Our comprehensive evaluation shows that the SMOTE-ENN-LR method attained a remarkable classification accuracy of 97.14%, outperforming contemporary state-of-the-art methods. This significant enhancement in accuracy highlights the potential of combining advanced data preprocessing techniques with robust statistical learning models to tackle the inherent challenges of microarray data analysis. Further, we employ Local Interpretable Model-agnostic Explanations (LIME) and SHAP (SHapley Additive exPlanations) to offer an understandings into our model's decision-making process, enhancing the predictions' transparency and interpretability. Moreover, the success of the SMOTE-ENN-LR method in this study paves the way for its application in other areas of medical diagnostics where similar data imbalances may impact the accuracy and effectiveness of disease classification. These results substantiate the effectiveness of the SMOTE-ENN-LR approach in managing the complexities of imbalanced microarray gene expression data, proposing a promising path for upcoming research in medical bioinformatics and precision medicine.*

***Keywords: Breast cancer; Gene expression; Machine learning; Logistic Regression; Classification; Explainable AI***

## 1.0 INTRODUCTION

Breast cancer is a significant global health challenge and the most common cancer among women worldwide [1]. It is caused by the uncontrolled growth of breast cells that can spread to other organs of the body. This disease exhibits complex and diverse behavior at both the cellular and molecular levels, resulting in varied prognostic and clinical outcomes [2-4]. When the regulatory mechanisms fail, breast cells divide uncontrollably, forming masses or lumps [5].

The risk factors for breast cancer are multifaceted, involving both inherent and extrinsic elements. Inherent factors such as age, sex, ethnicity, and genetic predispositions can predispose individuals to tumor formation. In contrast, extrinsic factors influenced by lifestyle, diet, and long-term medical treatments like hormone replacement therapy significantly impact neoplastic processes and can be somewhat controlled to mitigate risk [6]. The risk components of breast cancer are displayed in Fig. 1.

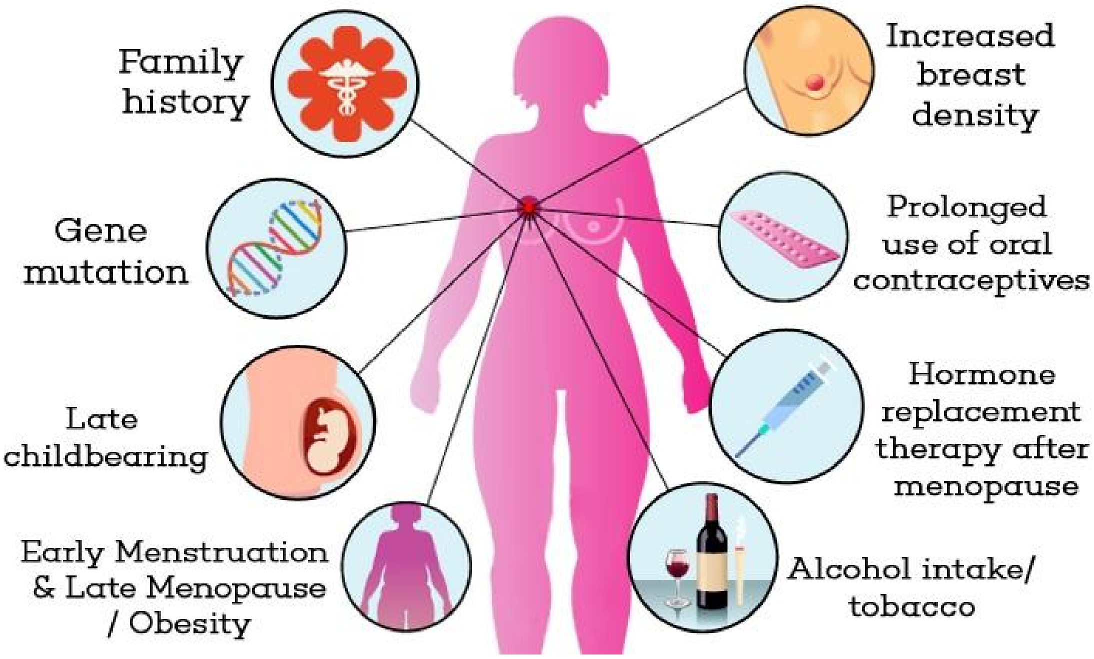
Family history
Gene mutation
Late childbearing
Early Menstruation & Late Menopause / Obesity
Increased breast density
Prolonged use of oral contraceptives
Hormone replacement therapy after menopause
Alcohol intake/ tobacco

Fig. 1: The causes and risk factors of breast cancer [7]

Table 1: A typical Microarray Gene Expression Data for n Samples Across m Genes [8]

| Sample | Gene 1 | Gene 2 | . . . | Gene m | Class Label |
|---|---|---|---|---|---|
| Sample 1 | 8.66 | 7.05 | . . . | 8.44 | C1 |
| Sample 2 | 10.25 | 6.79 | . . . | 8.55 | C2 |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| Sample n | 9.22 | 7.00 | … | 8.39 | Cn |

Microarray technology, a staple in bioinformatics and molecular biology, facilitates the measurement of expression levels across thousands of genes simultaneously. This capability enables it to generate high-dimensional data from minimal samples, which is advantageous for genomic studies but also introduces significant computational demands and the risk of overfitting, known as the "curse of dimensionality" [9-11]. Microarray gene expression (MGE), therefore, plays a pivotal role in supporting medical decision-making, particularly in cancer classification [12-13]. MGE data shows potential for preliminary detection of breast cancer, offering advantages and considerations compared to mammography, X-rays, and MRI. Table 1 provides a symbolic representation of MGE data in the n × (m+1) matrix. The matrix has n rows denoting samples, m columns denoting genes, and the last column indicating the class label, which represents the group to which each sample belongs. The use of machine learning (ML) methods in breast cancer research has become increasingly common due to their ability to train models rapidly and develop predictive systems that aid in effective decision-making. ML algorithms leverage statistical methods to autonomously learn and adapt from data, facilitating classification without human intervention [14-15]. Despite advances in computational power [16], the challenge of imbalanced datasets persists, where the minority class is underrepresented, which can severely skew the model’s performance [17].

This study introduces the SMOTE-ENN-LR approach, integrating SMOTE with ENN and LR to counteract the imbalance in the dataset and refine input features for improved classification accuracy. Furthermore, this paper discusses the importance of model interpretability in clinical settings, utilizing advanced methods like LIME and SHAP to enhance the transparency and reliability of the models, thereby fostering greater trust and acceptance in clinical applications. The main objectives of this study are as follows:

1. To integrate SMOTE with LR to address the imbalance in microarray gene expression data, which improves the representation of minority classes.
2. To employ ENN for data cleaning to enhance model accuracy by reducing noise and removing misleading samples.
3. To develop a reliable breast cancer classification model using SMOTE-ENN-LR with high accuracy.
4. To utilize LIME and SHAP to explain model predictions, thereby increasing the transparency and understandability of the model's outputs to clinicians.
5. To assess the suggested method against benchmarks, emphasizing higher accuracy and interpretability.

## 2.0 LITERATURE REVIEW

Breast cancer is the most common types of cancer affecting humanity, and their early detection is vital for effective treatment. However, breast cancer dataset often suffer from class imbalance, wherein the minority class (positive cases) is considerably smaller than the majority label (negative cases). This imbalance can lead to poor classification performance and biased results. To address this issue, various techniques have been proposed. This literature review aimed to critically evaluate the application of SMOTE-ENN, application of ML and integration of microarray gene expression for the classification of breast cancer. This review seeks to understand the effectiveness, limitations, and their practical implications, particularly in the domain of breast cancer classification.

### 2.1 SMOTE-ENN on Breast Cancer

The SMOTE-ENN has emerged as a broadly employed strategy to handle class imbalance by producing synthetic samples for the minority class, thereby improving the performance of model trained on imbalanced data [18]. SMOTE is an oversampling procedure that creates synthetic minority class samples by interpolating between existing samples [19]. The ENN technique aimed at eliminating noisy samples from both classes to improve the quality of the dataset [18]. However, it is essential to carefully evaluate and analyze the effectiveness of SMOTEENN in addressing the issues posed by imbalanced datasets [20]. Several studies have discovered the use of SMOTE-ENN for breast cancer classification [21-22]. One study compared the performance of SMOTE-ENN with other resampling techniques, such as undersampling, oversampling, SMOTE, and SMOTETOMEK, on medical datasets, including the breast cancer data. The results showed that SMOTE-ENN with LR and ANN achieved the highest accuracy [19]. Introduced by Tomek in 1976, the technique concentrates on identifying and removing samples that form ambiguous boundaries between classes. Therefore, by eliminating such instances, the classification model becomes less susceptible to misclassification and better handles the imbalanced class problem. One-Sided Selection method by Kubat, aimed to balance classes by removing the majority of cases that are close to the minority class. This technique assists in mitigating the challenges posed by imbalanced training sets and enhances the performance of classifiers [19]. In their comprehensive review, He and Garcia discussed various strategies for learning from imbalanced datasets. Their paper covered techniques such as resampling and ensemble methods to handle imbalanced class distributions [23]. A hybrid method to tackle the class imbalance problem in medical datasets was proposed in [24]. Combining techniques such as data-level sampling and ensemble learning, the method aims to improve the classification performance on imbalanced medical data. Another study used SMOTE-ENN to address the class imbalance issue of breast cancer and compared its performance with other methods, such as SMOTE, BLSMOTE (Borderline-SMOTE), and RUS (Random Undersampling) [25]. This study found that SMOTE-ENN + XGBoost achieved the highest accuracy among the fusion models. In addition, a study examined the efficacy of mix-up at the individual level on balanced data subjected to SMOTE-ENN to ensure consistency between the mixed label and the original label while also addressing imbalances within each class [23]. The author found that SMOTE-ENN improved the performance of Mix-up in handling class imbalance. Table 2 provides a concise overview of the key information for each referenced method, including the authors, the method itself, and its associated weaknesses or limitations. Researchers can use this summary to quickly grasp the main aspects of each approach and consider their applicability in their specific context.

In general, SMOTE-ENN is a promising technique for addressing the imbalance issue in the breast cancer dataset. It combines the strengths of SMOTE and ENN to generate synthetic minority class samples and eliminate noisy samples from both classes. Several studies have shown that SMOTEENN can improve the performance of several classifiers and achieve high accuracy in breast cancer classification. However, more study is needed to further explore the potential of SMOTE-ENN and its limitations.

Table 2: Summary Table for Breast Cancer Classification on SMOTE-ENN

| Ref. | Method | Weakness/Limitation |
|---|---|---|
| [18] | SMOTE | May introduce noise in synthetic samples. Sensitive to the choice of k in the algorithm |
| [19] | One-Sided Selection | May lead to information loss by removing potentially relevant majority class samples close to the minority class |
| [23] | Learning from Imbalanced Data | Focused on general strategies; lacks specific details about certain algorithms. Lacks depth in discussing newer approaches and advancements in the field |
| [20] | Ensemble Deep Learning | General review; not specific to addressing imbalanced datasets in healthcare. Limited focus on challenges and limitations in healthcare applications |

## 2.2 The Impact of Machine Learning in Breast Cancer Classification

Machine learning (ML) has emerged as a transformative tool in medical research, offering the potential to revolutionize cancer diagnostics and classification [26-28]. In the context of breast cancer, ML algorithms analyze complex datasets to extract patterns, predict outcomes, and aid clinicians in decision-making [29]. In recent years, the integration of ML techniques with microarray expression data has emerged as a promising avenue for enhancing the accuracy and precision of breast cancer classification. This section explores the applications and advancements of ML in addressing the issues posed by this cancer. The authors emphasized the pivotal role of gene selection in cancer classification, particularly utilizing support vector machines (SVM). SVMs are renowned for their capability to grip high-dimensional data and discern complex relationships within gene expression patterns. This study emphasizes the significance of choosing appropriate features for accurate breast cancer binary classification, offering insights into the integration of SVMs to achieve the goal [30]. ML approaches such as random forest, extra tree, and SVM are used to categorize breast cancer gene expression data into normal and relapse. Grid search cross-validation (CV) is used to optimize the hyperparameters of the methods. The tuned SVM outperforms the others with 97.78% of accuracy [31]. The comparative analysis by Kotsiantis et al. shed light on the effectiveness of different ML methods, presenting a nuanced understanding of their strengths and boundaries in the context of breast cancer classification [32]. Alrefai et al. utilized a convolutional neural network (CNN) to categorize breast cancer using MGE data. The authors reported an accuracy of 95.45% [33]. Table 3 displayed the summary of the breast cancer classification using ML techniques.

Table 3: Summary of breast cancer classification using machine learning

| Ref. | Method Used | Weakness/Limitation |
|---|---|---|
| [30] | Support Vector Machines (SVM) | Limited discussion on the interpretability of selected genes, potentially hindering biological insights. |
| [31] | SVM | Imbalanced data, missing feature selection method. |
| [34] | Various ML techniques | Challenge of handling high-dimensional data and potential sensitivity to noise in microarray expression datasets. |
| [32] | Comparative analysis of multiple methods | Limited exploration of ensemble methods; may not capture the combined strengths of various algorithms effectively. |
| [35] | Gene expression profiling for molecular classification | Dependency on the availability of comprehensive and representative datasets for accurate subtype identification. |
| [17] | Ensemble with Genetic Algorithm | Imbalanced data, computationally intensive, the performance heavily relies on the diversity and performance of the individual feature selection algorithms. |

The application of ML in DNA microarray analysis was expanded by Cho and Won [34], addressing challenges posed by noise and dimensionality. This comprehensive overview laid the foundation for utilizing ML techniques to unravel complex gene expression patterns linked with breast cancer subtypes. Sorlie and Tibshirani studied diagnostic markers and categorizing breast cancer into molecular subtypes, this research paved the way for more personalized and targeted approaches to treatment [35]. In [17], the authors proposed an ensemble filter feature selection approach called EnSNR for the classification of breast cancer. The EnSNR feature subset is generated automatically and a genetic algorithm (GA) is used to generate the model to classify breast cancer using the EnSNR feature subset. Ensemble methods may sacrifice interpretability, making it challenging to extract meaningful biological insights from the selected features. The integration of molecular information provided a more subtle understanding of the heterogeneity within breast cancer.

Therefore, the integration of ML methods with microarray expression data deals a great method to improve the accuracy of breast cancer classification. The insights gained from the reviewed literature lay the foundation for further study and advancements in this critical space of cancer diagnosis.

## 3.0 MATERIALS AND METHODS

The SMOTE-ENN-LR method combines 1) SMOTE for generating synthetic data points, 2) ENN for noise removal, and 3) LR for classification. SMOTE tackles the class imbalance issue by creating synthetic samples of the minority cases, thereby levelling the playing field for ML models. ENN further enhances the quality of the data by eliminating noisy majority class instances that may negatively impact model performance (See algorithm 1). Finally, LR was employed to build a predictive model on the cleaned and augmented data. Fig. 2 shows the system architecture of the study.

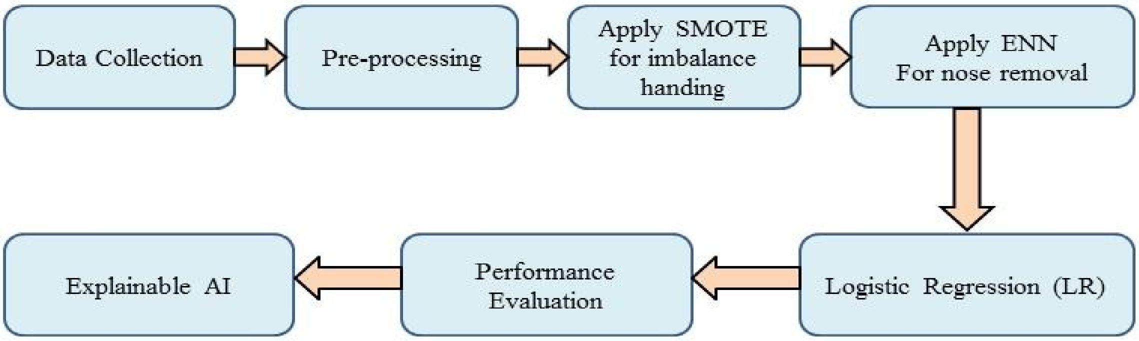

Fig. 2: Workflow of the Proposed Method

### 3.1 Dataset

The dataset was obtained from National Center for Biotechnology Information (NCBI) that hold the accession number GSE57297 [36]. It contains 50,739 genes or features from 25 breast cancer (abnormal) samples and 7 normal samples. Identified 10054 genes that are statistically significant out of a total of 50739 genes, with a P-value < 0.05. The dataset information is presented in Table 4. Table 5 shows some data samples from dataset.

Table 4: Dataset Descriptions

| Cancer | No. of Genes | Total Sample | Sample Size | Sample Types |
|---|---|---|---|---|
| Breast | 10054 | 32 | 25 | Abnormal (1) |
| | | | 7 | Normal (0) |

### 3.2 Data Pre-processing

The quality of data preprocessing directly influences the performance of ML models [37]. In this study, we employed a comprehensive preprocessing strategy tailored to handle the complexities of microarray gene expression data, aimed at optimizing the input for subsequent analysis and model training.

### 3.3 Dataset Loading and Inspection

The dataset was loaded from a specified path to ensure accessibility and reproducibility of results. Initial data inspection was conducted to understand the structure and distribution of the data. This included visualizing the first few rows to verify the integrity and format of the loaded data, which is crucial for identifying any discrepancies early in the preprocessing stage.

**Algorithm 1** Data Pre-processing, Balancing, and Model Evaluation

```
1: Load Dataset
2: Load the dataset from ’path/to/csv/file’
3: Display the first few rows to inspect the dataset structure
4: Separate Features and Target
5: Extract target variable $Y$ from the column ’Class’
```

```
Extract features $X$ by removing the target column 'Class'
Pre-process Data

    a. Filter Genes:
        i. Select genes/features with a p-value less than 0.05.
        ii. Ensure that the dataset includes 10,054 genes/features.
    b. Apply PCA:
        i. Apply Principal Component Analysis (PCA) to reduce dimensionality.
        ii. Retain components that explain 95% of the variance.
Initialize Model
Set up a LR model
Set Up Resampling Technique
Define SMOTE-ENN as the resampling technique to balance the dataset
Specify Edited Nearest Neighbours as part of the SMOTE-ENN process
Create Pipeline
Create a pipeline that combines the resampling tech- nique with the LR model
Define Cross-Validation
Set up a repeated stratified k-fold cross-validation method
Specify the number of splits and repetitions for cross- validation
Evaluate the Model
Use cross-validation to evaluate the model
Calculate metrics: accuracy, precision, recall, and F1- score
Output Metrics
Print the Performance Metrics
Explainable AI
Explanation using LIME and SHAP
```

Table 5: Data Samples from Dataset

| PC1 | PC2 | PC3 | PC4 | . . . | PC24 | PC25 | Class |
|---|---|---|---|---|---|---|---|
| -58.53 | 9.30 | -6.29 | 2.27 | . . . | -2.53 | -6.99 | 0 |
| -91.00 | -7.65 | 12.62 | -8.44 | . . . | 0.88 | 9.49 | 0 |
| -85.76 | -3.39 | 2.34 | -10.54 | . . . | -0.36 | -0.71 | 0 |
| -93.33 | 4.71 | 0.57 | -4.07 | . . . | 2.18 | 0.83 | 0 |
| -74.27 | -11.09 | 7.49 | 10.23 | . . . | -6.58 | 12.48 | 0 |
| 14.56 | 11.91 | 1.85 | -2.46 | . . . | 3.80 | 0.07 | 1 |
| 8.87 | 7.02 | 16.22 | 14.53 | . . . | -7.65 | -13.28 | 1 |
| 10.82 | -35.27 | -9.65 | 39.32 | . . . | -7.16 | 6.06 | 1 |
| 31.76 | 28.86 | -11.45 | 2.85 | . . . | 2.80 | 4.22 | 1 |
| 24.00 | -56.66 | 0.03 | -18.39 | . . . | 4.58 | 2.48 | 1 |
| 40.27 | 4.32 | 25.07 | 11.72 | . . . | 22.32 | 0.52 | 1 |
| 28.76 | -57.43 | -18.07 | -20.55 | . . . | 0.05 | -2.00 | 1 |
| 23.24 | -3.99 | 46.55 | 11.72 | . . . | -2.46 | 0.93 | 1 |
| 9.63 | 9.61 | 17.71 | -14.71 | . . . | -10.87 | 2.78 | 1 |

### 3.3.1 Feature and Target Separation

We segregated the target variable, Y, identified by the column labeled 'Class', from the input features, X. The feature matrix X was derived by excluding the target column, thereby isolating the predictors which are essential for the model training process.

### 3.3.2 Gene Filtering

Given the high dimensionality typical of microarray data, gene filtering was imperative to enhance model performance and reduce computational load:

### 3.3.3 P-value Filtering

We filtered genes based on statistical significance, retaining only those features with a p-value less than 0.05. This step ensured that the features included in the model were statistically relevant to the class outcomes.

### 3.3.4 Feature Consistency

The dataset was pruned to maintain a consistent set of 10,054 genes across all samples, aligning with typical gene expression panels used in clinical settings.

Dimensionality Reduction via PCA: To address the curse of dimensionality and improve the efficiency of our model, Principal Component Analysis (PCA) [38] was applied. PCA served two primary functions:

i. By transforming the data into a set of linearly un- correlated components, PCA reduces the dimensionality of the data while retaining those components that explain a substantial amount of variance, specifically 95% in this study.

ii. The reduced feature set facilitated a more interpretable model by diminishing the noise and less informative variables.

## 3.4 Data Balancing with SMOTE-ENN

The class imbalance prevalent in our dataset was addressed using SMOTE combined with ENN. This hybrid approach not only augmented the minority class through synthetic sample generation but also refined the dataset by removing misleading majority class samples. This preprocessing step was crucial for preventing model bias towards the majority class.

### 3.4.1 Synthetic Minority Over-sampling Technique

SMOTE is a method used to handle class imbalance in ML by generating synthetic examples for the minority class [18]. This technique helps to create a balanced dataset, which enhances the performance of the classifier by improving the representative- ness of the minority class.

SMOTE augments the data by performing the following steps:

i. **Identification of the k-nearest neighbors**: For each sample in the minority class, identify the $k$ nearest neighbors within the same class.

ii. **Synthetic Sample Generation**:
- For each sample $x_i$ in the minority class, select one of its $k$ nearest neighbors $x_{ni}$.
- Generate a synthetic sample $s$ using the formula:

$$s = x_i + \lambda \times (x_{ni} - x_i)$$

where $\lambda$ is a random number between 0 and 1.

iii. **Repetition**: Repeat the process until the dataset is sufficiently balanced.

### 3.4.2 Edited Nearest Neighbors (ENN)

The ENN algorithm is primarily utilized for reducing misclassification noise and improving the quality of datasets, especially those that exhibit class imbalance [39]. It works by removing instances that are likely mislabeled or are considered noise.

ENN operates through the following steps:

i. Neighbor Identification: Identify the $k$ nearest neighbors for each instance in the dataset.

ii. Majority Voting System: For each instance, examine the class labels of its nearest neighbors.

iii. Instance Deletion: If the majority of an instance's neighbors belong to a different class than the instance itself, the instance is removed from the dataset.

Let $X$ be a dataset with instances $x_i$, and $Y$ the corresponding class labels. For an instance $x_i$, denote $N(x_i)$ as the set of $k$ nearest neighbors, and $C(x)$ the class of instance $x$. The ENN rule can be mathematically formulated as:

$$\text{Remove } x_i \texttt{ if } \sum_{x_j \in N_k(x_i)} I(C(x_j) \neq C(x_i)) > k/2 \tag{1}$$

where $I$ is the indicator function, returning 1 if the condition is true and 0 otherwise.

Often, ENN is combined with SMOTE to both augment the minority class and ensure the quality of the augmented dataset.

This hybrid approach, known as SMOTE-ENN, provides a balanced and cleaner dataset for training robust models.

### 3.5 LR model

LR is extensively used for classification tasks [40]. It measures the probability of the target variable belonging to one class based on independent predictors.

The model uses a logistic function defined as:

$$\sigma(z) = 1 / (1 + e^{-z}) \tag{2}$$

where $z$ is a linear combination of input features. Given model parameters $\beta$ and inputs $X$, the probability that an observation belongs to class 1 is:

$$P(y = 1|X) = \sigma(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n) \tag{3}$$

### 3.6 Fitting the Model

The parameters of the LR model are estimated using maximum likelihood estimation, seeking to maximize the likelihood of observing the given data.

### 3.7 Use in Classification

To classify a new observation, the model calculates the predicted probability. If $P(y = 1|X) > 0.5$, the outcome is classified as class 1, otherwise as class 0.

### 3.8 Explainable AI (XAI) Approach

#### 3.8.1 Local Interpretable Model-agnostic Explanations

LIME provides insights into complex ML model predictions by creating simple, local surrogate models that approximate the predictions of the original model within a small neighborhood around the input being explained [41]. The core principle of LIME is to perturb the input data and observe how the predictions change, thus gaining insight into which features significantly influence the output [42]. By applying weights to these perturbations based on their proximity to the original input, LIME ensures that the local surrogate model—typically a linear model or decision tree—faithfully represents the original model's behavior in the vicinity of the input. This approach allows users to understand which features contribute to the decision made by a model, providing transparency and enhancing trust, especially in critical applications like healthcare and finance, where interpretability is essential. Mathematical Formulation:

Let $f$ be the complex model's prediction function, and $\xi(x)$ be the explanation model for an instance $x$. LIME solves the following optimization problem:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{4}$$

where:

- $G$ is the class of interpretable models, such as linear models or decision trees.
- $L$ is a loss function measuring how unfaithful $g$ is at approximating $f$ in the locality defined by $\pi x$, the proximity measure centered at $x$.
- $\Omega(g)$ is a complexity measure of the model $g$ (regularization).
- $\pi_x$ often takes the form of an exponential kernel based on the Euclidean distance, weighted by some hyperparameter.

#### 3.8.2 SHapley Additive exPlanations

SHAP is a process to elucidate the output of any ML model by measuring the influence of each feature to the prediction [43]. SHAP standards are presented on the concept of Shapley values from cooperative game theory [41]. These values offer a consistent and locally accurate attribution of feature importances across different types of data and models.

Mathematical Formulation:
The SHAP value $\phi_i$ for a feature $i$ is calculated as follows:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)] \tag{5}$$

where:

- $F$ is the set of all features.
- $S$ is a subset of features excluding $i$.
- $|S|$ is the cardinality of subset $S$.
- $f(S)$ is the model prediction when only the features in $S$ are used.

## 4.0 RESULTS

### 4.1 Model Parameter Settings

In this study, we meticulously configured the SMOTEENN-LR model to ensure robustness and reliability in the classification of breast cancer from MGE. Given the complexity of handling imbalanced datasets and the high dimensionality of microarray data, specific parameter settings are critical for optimizing model performance. Below, we outline the key parameters and their settings:

### 4.2 Cross-validation Strategy

To evaluate the model's performance comprehensively, we employed a Repeated Stratified K-Fold CV strategy. This approach confirms that each fold of the data is utilized for both training and testing, and that the data is split in a way that preserves the percentage of samples for each class.

### 4.3 Number of Splits

The cross-validation process was configured to use 7 splits (see table 6). This setting was chosen to provide a balance between computational efficiency and model evaluation thoroughness, allowing for detailed performance metrics across multiple subsets of data.

Table 6: Model Parameter Settings

| Parameter | Value |
|---|---|
| Cross-validation Strategy | Repeated Stratified K-Fold |
| Number of Splits | 7 |
| Number of Repeats | 3 |
| Random State | 1 |
| Scoring Metrics | Accuracy, Precision , Recall , F1-score |
| Parallel Processing | Enabled (n_jobs=-1) |

### 4.4 Number of Repeats

Each cross-validation cycle was repeated 3 times to ensure stability and reliability in the performance estimates. This repetition helps mitigate the variance that might arise from the random partitioning of data in each fold.

### 4.5 Random State

A constant random state of 1 was used to initialize the random number generator. This setting ensures the reproducibility of the model training and validation process, allowing other researchers to replicate our results under the same experimental conditions.

### 4.6 Parallel Processing

To expedite the computational process, parallel processing was enabled with n_jobs=-1. This setting allows the model to utilize all available CPU cores, significantly reducing the time required for cross-validation and model training. These parameter settings were carefully selected to optimize the model for the specific challenges posed by microarray data analysis, particularly in dealing with imbalanced datasets. By enhancing the oversampling of minority classes through SMOTE and reducing noise via ENN before applying LR, our approach not only addresses the imbalance in the data but also improves the overall predictive accuracy, as evidenced by our experimental results.

### 4.7 Evaluation Metrics

Performance metrics were employed to assess the performance and effectiveness of the proposed methods. The confusion matrix summarizes the performance of the proposed methods on a set of test data. In this study, the following performance metrics [44] were utilized to critically assess the performance of the proposed cancer classification approach. A confusion matrix is illustrated in the Table 7, which can be employed to observe the performance of a classifier.

Table 7: Confusion Matrix

| **Actual Class** | **Predicted Class** | |
|---|---|---|
| | **Positive (P)** | **Negative (N)** |
| **Positive (P)** | True Positive (TP) | False Negative (FN) |
| **Negative (N)** | False Positive (FP) | True Negative (TN) |

The evaluation metrics for classification models are defined as follows:

- **Accuracy**: The ratio of accurate outcomes to the total cases analysed.

$$\text{Accuracy} = (TP + TN) \,/\, (TP + FN + FP + TN) \tag{6}$$

- **Precision**: The ratio of accurate positive identifications.

$$\text{Precision} = TP \,/\, (TP + FP) \tag{7}$$

- **Recall** (Sensitivity): The proportion of actual positives that were correctly identified.

$$\text{Recall} = T\ P \,/\, T\ P + F\ N \tag{8}$$

- **F-measure**: The harmonic mean of precision and recall.

$$\text{F-measure} = 2 \times \text{Precision} \times \text{Recall} \,/\, (\text{Precision} + \text{Recall}) \tag{9}$$

- **ROC Curve**: A graphical representation demonstrating the diagnostic efficacy of a binary classifier system as its discrimination threshold is adjusted.

$$\text{ROC curve} = \text{Sensitivity vs } (1 - \text{Specificity}) \tag{10}$$

This section briefly explains the experimental findings across three stages: feature selection, comparison analysis, and evaluation of the classification phase. Table 8 presents the performance metrics of the SMOTE-ENN-LR approach applied to microarray expression data for breast cancer.

Table 8: Performance Analysis Results of Breast Cancer Classification

| **Cancer** | **Precision (%)** | **Recall (%)** | **F1-score (%)** | **Accuracy (%)** |
|---|---|---|---|---|
| Breast | 96.43 | 98.21 | 97.96 | 97.14 |

The method was assessed using a 7-fold CV with three repeats on the LR model. The metrics, such as Precision, Recall, F1-score, and Accuracy, all expressed in percentages. The method achieved a notable precision of 96.43%, suggesting that the vast majority of positive predictions were indeed true positives. The recall for breast cancer was 98.21%, suggesting that the model successfully identified most of the actual positive cases. The F1-score, which balances precision and recall, was exceptionally high at 97.96%. This indicates a well-balanced model in terms of both false positives and false negatives. The model demonstrated an overall accuracy of 97.14%, affirming its capability to correctly classify both positive and negative cases. The outcomes exhibit the effectiveness of the proposed approach in addressing the imbalance problem in MGE data for breast cancer classification. The higher precision and recall outcomes indicate that the model is reliable in identifying true positives while minimizing false positives, which is crucial in medical diagnoses. The F1-scores are particularly notable, reflecting a robust balance between precision and recall, a key aspect in imbalanced datasets. Fig. 3 represents the receiver operating characteristic (ROC) curve generated for breast cancer.

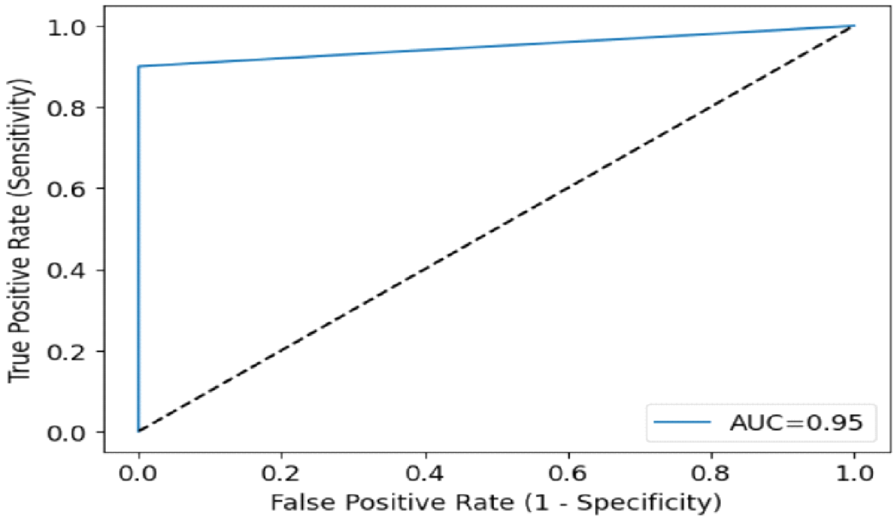

Fig. 3: ROC curve

The SMOTE-ENN-LR method shows high precision (96.43%), recall (98.21%), and F1-score (97.96%). These metrics indicate the method's robustness in handling imbalanced datasets, particularly in ensuring a minimal number of false positives (high precision) and a high detection rate of true positives (high recall). The SMOTE-ENN-LR's accuracy of 97.14% significantly outperforms that of other existing methods. This suggests that the combined approach of balancing the dataset and then applying LR is more effective in this context. The superior performance of SMOTE-ENNLR in terms of accuracy and the available metrics suggests its suitability for breast cancer microarray expression data. Using SMOTE-ENN-LR effectively addresses the imbalance problem, enhancing the model's predictive power.

### 4.8 Explaining Predictions with LIME

The LIME visualization Fig. 4 provided outlines the contribution of various principal components (PCs) towards a model's prediction that a particular breast cancer case is abnormal based on microarray gene expression data.

Prediction Probabilities:

Class 0 (Normal): 0.000 (No probability or very low probability of the sample being normal)

Class 1 (Abnormal): 1.00 (Extremely high or absolute certainty probability of the sample being abnormal)

The Fig. lists several principal components (PCs) with their respective contribution scores to the prediction decision. Each feature contribution is expressed as a positive or negative value indicating the direction and strength of the influence on the prediction outcome (in this case, toward abnormal).

a. PC1 (31.76) Influence: Strong positive contribution (15.75)
   Interpretation: Higher values of PC1 significantly increase the probability of the sample being classified as abnormal.
b. PC22 (16.54)
   Influence: Positive contribution (0.33)
   Interpretation: High values of PC22 moderately increase the likelihood of an abnormal classification.
c. PC4 (2.85)
   Influence: Positive contribution (0.02)
   Interpretation: Slightly positive influence on the prediction of abnormal.
d. PC23 (-1.07)
   Influence: Positive contribution (0.02)
   Interpretation: The value range of PC23 slightly pushes the decision toward abnormal.
e. PC24 (2.80)
   Influence: Positive contribution (0.02)
   Interpretation: Similar to PC4, contributes slightly toward an abnormal prediction.
f. PC2 (28.86)
   Influence: Positive contribution (0.02)
   Interpretation: Higher values of PC2 slightly favor an abnormal classification.
g. PC10 (-21.47)
   Influence: Positive contribution (0.02)
   Interpretation: The lower value of PC10 slightly supports the abnormal classification.

h. PC5 (5.89)
Influence: Positive contribution (0.02)
Interpretation: Mid-range values of PC5 have a minor positive impact on the abnormal prediction.

i. PC12 (12.81)
Influence: Positive contribution (0.02)
Interpretation: Higher values of PC12 slightly influence the sample toward being classified as abnormal.

j. PC13 (3.68)
Influence: Positive contribution (0.02)
Interpretation: Indicates that a mid-range value of PC13 subtly pushes toward an abnormal classification.

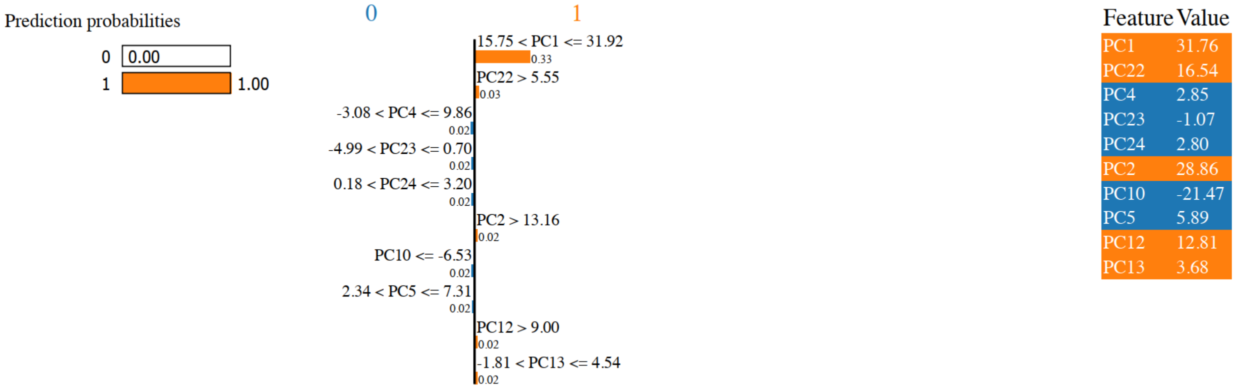

Fig. 4: Local interpretability with LIME-1[0-Normal, 1-Abnormal]

The second LIME visualization Fig. 5 provided illustrates how different principal components (PCs) contribute to a specific prediction for a microarray gene expression data sample in breast cancer classification. This explanation focuses on the classification probabilities for two classes (likely 0 for ’normal’ and 1 for ’abnormal’), with the provided sample having probabilities of 0.790 for class 0 and 0.211 for class 1, suggesting a leaning towards the ’normal’ classification but with considerable uncertainty.

- PC1 (<= -4.22): This feature heavily influences the prediction towards class 0 (normal) with a high contribution score of 0.98, suggesting that lower values of PC1 are strongly indicative of a normal classification.
- PC18 (0.59 < PC18 <= 3.34): Sits within a range that mildly supports the normal class with a contribution of 0.00, indicating neutrality in this context.
- PC9 (1.54 < PC9 <= 9.86): Also shows a neutral contribution (0.00), despite the value falling within a significant range, suggesting that its impact is less decisive under the given model.
- PC13 (> 4.54): With a value of 6.72, exceeding 4.54, it again shows a neutral influence on the prediction outcome.
- PC20 (<= -5.58): A very low value of -13.14 contributes neutrally, which is interesting as it suggests that extreme values of PC20 might not be influential for this model.
- PC11 (-6.93 < PC11 <= -0.25): Shows a small range and a neutral effect on the decision, highlighting less impact on the classification.
- PC8 (> 8.93): A high value of 17.27 shows a neutral contribution, indicating that this feature in high ranges does not sway the prediction considerably.
- PC24 (> 3.20): With a value of 9.76, much higher than 3.20, it still contributes neutrally, suggesting that for this model, higher values of PC24 are not determining factors.
- PC16 (0.35 < PC16 <= 7.73): Falls within a contributing range but shows a neutral impact with a score of 0.00.
- PC22 (-0.69 < PC22 <= 5.55): Also demonstrates a neutral impact with its value of 0.99.

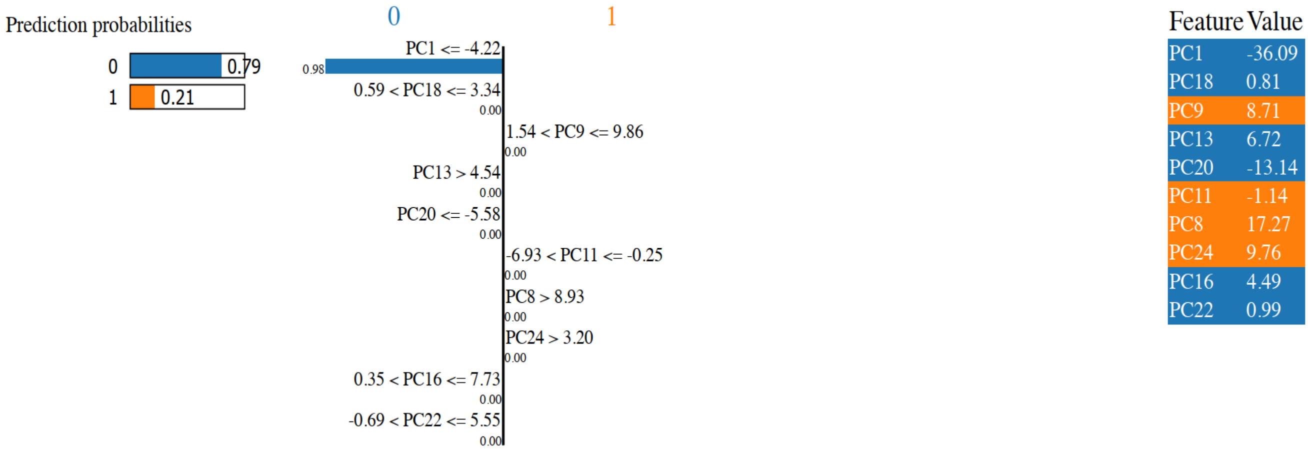

Fig. 5: Local interpretability with LIME-2[0-Normal, 1-Abnormal]

### 4.9 Understanding Model Predictions: A SHAP Analysis

The provided SHAP visualization (refer to Fig. 6) offers an in-depth look at how different features, denoted as principal components (PCs), influence a specific model prediction. Here we focus on a single prediction instance, examining how each component nudges the model's output away from or towards the expected average outcome.

- **Base Value and Prediction**: The base value, denoted by [$f(X)$] = 0.75, represents the expected model output when no specific feature information is included. The actual prediction for the instance is significantly lower, ($x$) = 0, indicating a shift towards class '0'.
- **Negative Contributions**: Features such as PC1, with a value of -58.53, show a substantial negative impact, reducing the probability of the outcome being class '1'. Similarly, other PCs like PC23, PC22, and others collectively push the prediction towards class '0'.
- **Positive Contributions**: Conversely, PCs like PC19 and PC17, though minor, try to counterbalance by pushing the prediction slightly towards class '1'.

The aggregation of feature effects culminates in a final prediction that deviates markedly from the base expectation, underscoring the importance of principal components like PC1 in this decision-making process.

### 5.0 DISCUSSION

The SMOTE-ENN-LR approach for classifying breast cancer using MGE data demonstrated highly promising results, significantly improving the reliability and accuracy of diagnostic predictions. The precision rate achieved was 96.43%, indicating a high likelihood that the positive predictions made by the model are true positives. This is vital in medical settings, where the cost of false positives—such as unnecessary treatments and associated stress for patients—can be substantial. Additionally, the model exhibited a recall of 98.21%, suggesting it successfully identified the majority of real positive samples. The higher recall rate is critical for effective screening programs where missing a true case can delay crucial treatment, adversely affecting patient outcomes.

The balanced F1-score of 97.96% highlights the model's robustness, effectively managing the trade-off between precision and recall—a key aspect in dealing with imbalanced datasets. Moreover, the accuracy of 97.14% affirms the model's competence in accurately classifying both positive and negative samples, confirming its utility in a clinical environment. These metrics not only illustrate the effectiveness of the SMOTE-ENN-LR method in handling class imbalance and noise in the data but also showcase its superiority over traditional methods, which often struggle under similar conditions. The integration of SMOTE and ENN pre-processing ensures that the LR model is trained on balanced and cleaned data, enhancing the predictive performance and reliability necessary for medical diagnostics. One potential limitation of this study is the relatively small number of normal breast cancer samples in the dataset.

### 6.0 COMPARISON WITH STATE-OF-ART METHOD

In the domain of breast cancer classification using MGE data, our proposed SMOTEENN-LR method demonstrates significant advancements over several state-of-the-art techniques, including Genetic Algorithms (GA) [17], Support Vector Machines (SVM) [45], and kNearest Neighbors (kNN) [46]. These traditional methods achieved accuracies of 92.39%, 90%, and 91.24%, respectively. However, they do not inherently address the serious matter of class imbalance, which can lead to biased predictions favoring the majority class, a significant drawback in medical diagnostics.

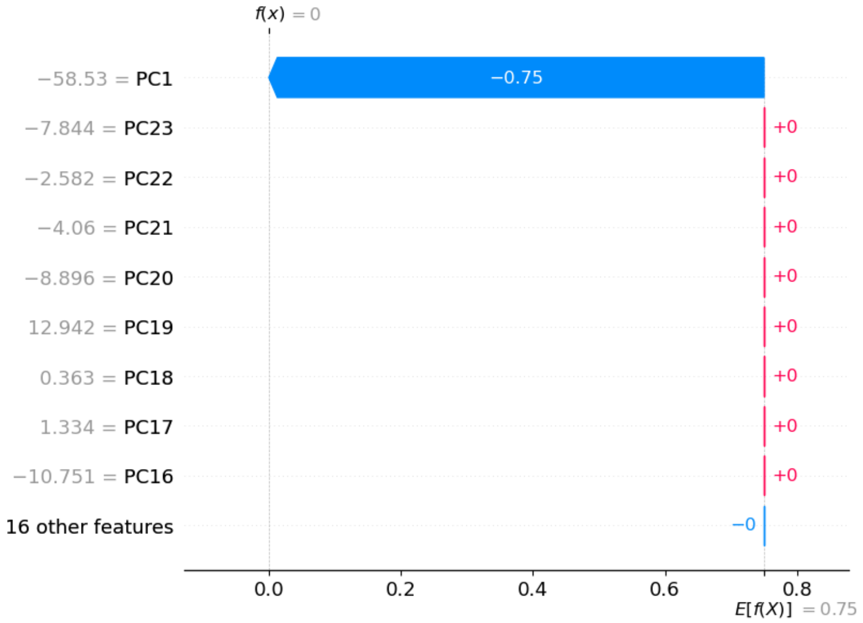

Fig. 6: SHAP visualization showing the impact of principal components on the model's prediction.

Our approach leverages the SMOTE to generate synthetic samples for the minority class, effectively balancing the dataset. This is complemented by the ENN technique, which further improves the dataset by eliminating noisy instances before applying LR. This dual approach not only addresses class imbalance but also ensures data integrity, resulting in enhanced predictive performance.

As shown in Table 9, the SMOTE-ENN-LR method achieved an accuracy of 97.14%, precision of 96.43%, recall of 98.21%, and an F1-score of 97.96%. These metrics clearly surpass the performance of the other methods listed, demonstrating the efficacy of our approach in producing a more balanced and accurate predictive model. Moreover, our study uniquely incorporates Explainable AI (XAI) techniques, specifically LIME and SHAP, to provide transparency into the model's decision-making process. Compared to other methods listed in Table 9 (such as Catboost [26], GRU-RNN [27], and DNN post-SMOTE [28] that did not utilize XAI frameworks like LIME and SHAP, our comprehensive approach of integrating SMOTE and ENN, followed by LR and supported by XAI, sets a new benchmark in terms of both performance and interpretability.

In summary, the SMOTE-ENN-LR method not only achieves superior predictive accuracy but also addresses the class imbalance problem and enhances model interpretability through XAI. This makes it exceptionally suitable for critical healthcare applications, where accurate and explainable diagnosis is paramount.

Table 9: Comparison with State-of-art Method

| **Ref.** | **Method** | **Precision (%)** | **Recall (%)** | **F1-score (%)** | **Accuracy (%)** | XAI |
|---|---|---|---|---|---|---|
| [26] | Catboost | - | - | - | 92 | X |
| [27] | GRU-RNN | - | - | - | 97.8 | X |
| [28] | DNN post SMOTE | - | - | - | 81 | X |
| [17] | GA | - | - | - | 92.39 | X |
| [45] | SVM | - | 87 | - | 90 | X |
| [46] | kNN | - | - | - | 91.24 | X |
| Our Study | SMOTE-ENN-LR | 96.43 | 98.21 | 97.96 | 97.14 | YES |

## 7.0 CONCLUSION

In this paper, a comprehensive strategy is introduced to handle class imbalance issues in ML. It begins by using SMOTE to create synthetic instances of the minority class, effectively increasing its representation in the dataset. ENN comes into play next, ensuring the dataset is devoid of noisy instances from the majority class by comparing class labels of neighbors. After this preprocessing, LR is trained on the refined data, enabling the model to predict the probability of fitting to a specific class. This method has proven successful in clinical diagnosis, where accurate predictions are critical. Therefore, addressing class imbalance and noise, SMOTE-ENNLR contributes to more robust and reliable ML models. Finally, the experimental findings reveal that the suggested SMOTE-ENN-LR method achieved a test accuracy of 97.14%. The SMOTE-ENN-LR approach yields superior classification accuracy compared to existing methods.

## Contributions of this Study

The main contributions of this study are as follows:

1. Introduced the SMOTE-ENN-LR method, which combines SMOTE, ENN, and LR to address the class imbalance in breast cancer gene expression data.
2. Utilized SMOTE to enhance the representation of minority classes through oversampling, and applied ENN to remove noise from the dataset, ensuring cleaner and more relevant data for training the model.
3. Achieved a high classification accuracy of 97.14%, demonstrating the effectiveness of the SMOTE-ENNLR approach in predicting breast cancer from MGE data.
4. Demonstrated superior performance over state-of-the-art methods for breast cancer prediction, providing evidence of the robustness and reliability of the SMOTE-ENN-LR model.
5. Implementation of LIME to provide local interpretative insights into the model predictions and SHAP to decipher the role of each parameters to the prediction across the model.
6. Tackled the prevalent issues of imbalanced data and noise in medical datasets, which are crucial for enlightening diagnostic accuracy and reducing the risk of misdiagnosis.

## Future Work

As we look to the future, our research will aim to further refine the SMOTE-ENN-LR method by increasing the dataset size to include a more diverse range of gene expressions and clinical scenarios. This expansion will enable us to validate our findings more comprehensively and explore the potential reduction of false-positive rates, thereby enhancing the diagnostic precision of breast cancer classification. We will collect large dataset for future research.

## Author Contributions

Conceptualization, M.F.B.A.A., F.Z.E, and A.N.; methodology, M.F.B.A.A., F.Z.E. and A.N.; software, M.F.B.A.A.; validation, M.F.B.A.A., and A.N.; formal analysis, M.F.B.A.A., F.Z.E, and A.N.; data curation, M.F.B.A.A., F.Z.E; writing—original draft preparation, M.F.B.A.A.,; writing—review and editing, M.F.B.A.A., F.Z.E. O.A, A.N., and T.M.; visualization, M.F.B.A.A.,T.M.; supervision, A.N., R.Y., T.N.M.A. and Z.S.; funding acquisition, W.S. and M.N.A.A. All authors have read and agreed to the published version of the manuscript.

**Funding**

This study was supported by the Ongoing Research Funding Program, King Saud University, Riyadh, Saudi Arabia, (ORF-2025-81).

**Acknowledgments**

The authors of this study extend their appreciation to Ongoing Research Funding Program (ORF-2025-81), King Saud University, Riyadh, Saudi Arabia. Additionally, we utilized Quillbot for grammar refinement, which greatly enhanced the quality of the manuscript.

## Competing Interests

The authors have stated that there are no competing interests.

## REFERENCES

[1] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi, “Prediction of Breast Cancer using Machine Learning Approaches,” *J. Biomed. Phys. Eng.*, vol. 12, no. 3, pp. 297–308, 2022, doi: 10.31661/jbpe.v0i0.2109-1403.

[2] P. A. Jones and S. B. Baylin, “The Epigenomics of Cancer,” *Cell*, vol. 128, no. 4, pp. 683–692, 2007, doi: 10.1016/j.cell.2007.01.029.

[3] M. Mokoatle, V. Marivate, D. Mapiye, R. Bornman, and V. M. Hayes, “A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application,” *BMC Bioinformatics*, vol. 24, no. 1, p. 112, 2023, doi: 10.1186/s12859-023-05235-x.

[4] P. Muhammed Jamshed Alam, S. Akter, and T. Mahmud, “An Expert System to Detect Uterine Cancer under Uncertainty,” *IOSR J. Comput. Eng.*, vol. 16, no. 5, pp. 36–47, 2014, doi: 10.9790/0661-16513647.

[5] M. Mohaiminul Islam, S. Huang, R. Ajwad, C. Chi, Y. Wang, and P. Hu, “An integrative deep learning framework for classifying molecular subtypes of breast cancer,” *Comput. Struct. Biotechnol. J.*, vol. 18,

pp. 2185–2199, 2020, doi: 10.1016/j.csbj.2020.08.005.

[6] M. Kamińska, T. Ciszewski, K. Łopacka-Szatan, P. Miotła, and E. Starosławska, "Breast cancer risk factors," *Prz. Menopauzalny*, vol. 14, no. 3, pp. 196–202, 2015, doi: 10.5114/pm.2015.54346.

[7] Medindia, "Breast Cancer." Accessed: Jan. 10, 2023. [Online]. Available: https://www.medindia.net/patientinfo/breast-cancer.htm

[8] M. F. B. Abdul Aziz, A. Nazri, R. Yaakob, F. Z. Evamoni, T. N. Mohd Aris, and Z. Sekawi, "A Comprehensive Survey on Different Machine Learning Approaches for Breast Cancer Prediction based on Medical Imaging Modalities and Microarray Gene Expression.," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 9, pp. 2615–2625, 2023.

[9] A. Garg and V. Mago, "Role of machine learning in medical research: A survey," *Comput. Sci. Rev.*, vol. 40, 2021, doi: 10.1016/j.cosrev.2021.100370.

[10] J. Kim, Y. Yoon, H. J. Park, and Y. H. Kim, "Comparative Study of Classification Algorithms for Various DNA Microarray Data," *Genes (Basel).*, vol. 13, no. 3, 2022, doi: 10.3390/genes13030494.

[11] S. Osama, H. Shaban, and A. A. Ali, "Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review," *Expert Syst. Appl.*, vol. 213, p. 118946, 2023, doi: 10.1016/j.eswa.2022.118946.

[12] F. Alharbi and A. Vakanski, "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review," *Bioengineering*, vol. 10, no. 2, 2023, doi: 10.3390/bioengineering10020173.

[13] N. Pochet, F. De Smet, J. A. K. Suykens, and B. L. R. De Moor, "Systematic benchmarking of microarray data classification: Assessing the role of non-linearity and dimensionality reduction," *Bioinformatics*, vol. 20, no. 17, pp. 3185–3195, 2004, doi: 10.1093/bioinformatics/bth383.

[14] Y. Amethiya, P. Pipariya, S. Patel, and M. Shah, "Comparative analysis of breast cancer detection using machine learning and biosensors," *Intell. Med.*, vol. 2, no. 2, pp. 69–81, May 2022, doi: 10.1016/J.IMED.2021.08.004.

[15] T. Mahmud, P. Michal, and M. Fumito, "Exhaustive Study into Machine Learning and Deep Learning Methods for Multilingual Cyberbullying Detection in Bangla and Chittagonian Texts," *Electronics*, vol. 13, no. 9, p. 1677, 2024.

[16] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review," *Expert Syst. Appl.*, vol. 167, no. October 2020, pp. 1–20, 2021, doi: 10.1016/j.eswa.2020.114161.

[17] S. Hengpraprohm and S. Jungjit, "Ensemble feature selection for breast cancer classification using microarray data," *Intel. Artif.*, vol. 23, no. 65, pp. 100–114, 2020, doi: 10.4114/intartif.vol23iss65pp100-114.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[19] M. Kubat, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," in *International Conference on Machine Learning*, 1997, p. 179.

[20] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 757–774, 2023, doi: 10.1016/j.jksuci.2023.01.014.

[21] R. Gupta, R. Bhargava, and M. Jayabalan, "Diagnosis of Breast Cancer on Imbalanced Dataset Using Various Sampling Techniques and Machine Learning Models," in *International Conference on Developments in eSystems Engineering, DeSE*, 2021, pp. 162–167. doi: 10.1109/DESE54285.2021.9719398.

[22] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *J. Biomed. Inform.*, vol. 107, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[24] B. Han and E. Eskin, "Interpreting meta-analyses of genome-wide association studies," *PLoS Genet.*, vol. 8, no. 3, 2012, doi: 10.1371/journal.pgen.1002555.

[25] D. Devi, S. kr Biswas, and B. Purkayastha, "Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance," *Pattern Recognit. Lett.*, vol. 93, pp. 1339–1351, 2017, doi: 10.1016/j.patrec.2016.10.006.

[26] G. Sultan and S. Zubair, "An ensemble of bioinformatics and machine learning approaches to identify shared breast cancer biomarkers among diverse populations," *Comput. Biol. Chem.*, vol. 108, p. 107999, 2024, doi: 10.1016/j.compbiolchem.2023.107999.

[27] S. Babichev, I. Liakh, and I. Kalinina, "Applying the Deep Learning Techniques to Solve Classification Tasks Using Gene Expression Data," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3368070.

[28] A. Kishore, L. Venkataramana, D. V. V. Prasad, A. Mohan, and B. Jha, "Enhancing the prediction of IDC breast cancer staging from gene expression profiles using hybrid feature selection methods and deep learning architecture," *Med. Biol. Eng. Comput.*, 2023, doi: 10.1007/s11517-023-02892-1.

[29] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[30] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Mach. Learn.*, 2002, doi: 10.1023/A:1012487302797.

[31] N. M. Ali, R. Besar, and N. A. A. Aziz, “A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation,” *Bull. Electr. Eng. Informatics*, vol. 12, no. 2, pp. 1047–1054, 2023, doi: 10.11591/eei.v12i2.4838.

[32] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “A comparative analysis of machine learning methods in classification of breast cancer.,” *Artif. Intell. Med.*, vol. 41, no. 1, pp. 27–41, 2007.

[33] N. Alrefai *et al.*, “An integrated framework based deep learning for cancer classification using microarray datasets,” *J. Ambient Intell. Humaniz. Comput.*, 2023, doi: 10.1007/s12652-022-04482-9.

[34] S.-B. Cho and H.-H. Won, “Machine learning in DNA microarray analysis for cancer classification,” in *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, 2003, pp. 189–198.

[35] T. Sorlie and R. Tibshirani, “Gene expression profiling for molecular classification of breast cancer and identification of diagnostic markers.,” *Genome Biol.*, vol. 4, no. 7, p. 58, 2003.

[36] J. Fu, W. Allen, A. Xia, Z. Ma, and X. Qi, “Identification of biomarkers in breast cancer by gene expression profiling using human tissues,” *Genomics Data*, vol. 2, pp. 299–301, 2014, doi: 10.1016/j.gdata.2014.09.004.

[37] A. Q. Md, S. Kulkarni, C. J. Joshua, T. Vaichole, S. Mohan, and C. Iwendi, “Enhanced Preprocessing Approach Using Ensemble Machine Learning Algorithms for Detecting Liver Disease,” *Biomedicines*, vol. 11, no. 2, p. 518, 2023, doi: 10.3390/biomedicines11020581.

[38] S. Laghmati, S. Hamida, K. Hicham, B. Cherradi, and A. Tmiri, “An improved breast cancer disease prediction system using ML and PCA,” *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 33785–33821, 2024, doi: 10.1007/s11042-023-16874-w.

[39] M. F. Kabir and S. Ludwig, “Classification of Breast Cancer Risk Factors Using Several Resampling Approaches,” in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 2018, pp. 1243–1248. doi: 10.1109/ICMLA.2018.00202.

[40] S. Prusty, S. Patnaik, S. K. Dash, and S. G. Priyadarsini Prusty, “SEMeL-LR: An improvised modeling approach using a meta-learning algorithm to classify breast cancer,” *Eng. Appl. Artif. Intell.*, vol. 129, p. 107630, 2024, doi: 10.1016/j.engappai.2023.107630.

[41] V. Vimbi, N. Shaffi, and M. Mahmud, “Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer’s disease detection,” *Brain Informatics*, vol. 11, no. 1, p. 10, 2024.

[42] B. U. Maheswari, A. A, A. Avvaru, A. Tandon, and R. P. de Prado, “Interpretable Machine Learning Model for Breast Cancer Prediction Using LIME and SHAP,” in *IEEE 9th International Conference for Convergence in Technology*, 2024, pp. 1–6. doi: 10.1109/i2ct61223.2024.10543965.

[43] S. Bai, S. Nasir, R. A. Khan, S. Arif, A. Meyer, and H. Konik, “Breast Cancer Diagnosis: A Comprehensive Exploration of Explainable Artificial Intelligence (XAI) Techniques,” *arXiv Prepr. arXiv2406.00532*, 2024, [Online]. Available: http://arxiv.org/abs/2406.00532

[44] T. Mahmud, K. Barua, S. U. Habiba, N. Sharmen, M. S. Hossain, and K. Andersson, “An Explainable AI Paradigm for Alzheimer’s Diagnosis Using Deep Transfer Learning,” *Diagnostics*, vol. 14, no. 3, p. 345, 2024, doi: 10.3390/diagnostics14030345.

[45] J. Wu and C. Hicks, “Breast cancer type classification using machine learning,” *J. Pers. Med.*, vol. 11, no. 2, pp. 1–12, 2021, doi: 10.3390/jpm11020061.

[46] K. Adem, “Diagnosis of breast cancer with Stacked autoencoder and Subspace kNN,” *Phys. A Stat. Mech. its Appl.*, vol. 551, 2020, doi: 10.1016/j.physa.2020.124591.