International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 3
ISSN: 2088-5334

An Extreme Gradient Boosting for Cancer Feature Extraction and Classification

Teo Voon Chuan [a], Kohbalan Moorthy [a,1], Nasarudin Ismail [b,2], Mohd. Murtadha Mohamad [c], Chan Weng Howe [c]

[a] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia*
[b] *Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia*
[c] *Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia*

Corresponding author: [1]kohbalan@umpsa.edu.my; [2]nasar@uthm.edu.my

***Abstract*— Cancer remains a leading cause of death worldwide; the World Health Organization (WHO) reports that there have been nearly 10 million cancer-related deaths in recent years, with breast cancer affecting over 2.1 million women annually on a global scale, posing significant challenges for early detection and diagnosis. Gene selection, using DNA microarray data, is crucial for reducing the presence of less informative genes and ensuring the selection of genes relevant to disease diagnosis. Cancer classification involves identifying the type of cancer and determining the extent of tumor growth and spread. This research focuses on improving gene selection for cancer classification using the XGBoost classifier, an efficient open-source implementation of the gradient-boosted trees algorithm. The primary goal is to enhance the performance of gene selection, enabling timely and appropriate treatments for cancer patients, as early detection is vital for ensuring a full recovery. Additionally, this research aims to reduce the time and expense associated with gene selection for cancer classification while increasing classification accuracy. The proposed method achieved an accuracy of approximately 93%, with precision, recall, and F1-score values of 93%, 87%, and 90%, respectively. The study highlights the potential of the XGBoost classifier in optimizing gene selection and improving diagnostic processes. Future work will focus on enhancing the accuracy of gene selection for cancer classification and reducing the number of irrelevant genes before proceeding to subsequent processes. This approach holds promises for streamlining the diagnostic process, improving patient outcomes, and offering significant benefits in timely cancer treatment.**

***Keywords*— Machine learning; classification; gene selection; cancer prediction; XGBOOST.**

Manuscript received 25 Sep. 2024; revised 14 Dec. 2024; accepted 12 Mar. 2025. Date of publication 30 Jun. 2025.

I. INTRODUCTION

According to this globalized epoch, cancer remains the highest cause of mortality worldwide, irrespective of the country. According to the 2024 World Health Organization [1], there have been almost 10 million cancer-related deaths in recent years, affecting over 2.1 million women annually with breast cancer (BC) on a global scale. The most prevalent cancers in recent years [2], [3], as reported by the WHO, include breast, lung, and prostate cancers. Although there are numerous types of cancer, detecting and starting fast initiatives can significantly impact a cure. Gene selection plays a crucial role in aiding researchers in cancer classification. Gene selection involves identifying uninformative genes and removing extraneous genes from the gene expression dataset [4]. This technique is crucial for identifying a set of genes relevant to a specific disease. Simply put, gene selection is used to identify informative and significant genes relevant to clinical diagnoses, such as cancer. Cancer classification involves identifying the type of cancer and determining the extent of the tumor's growth, which may spread. Generally, different kinds of tissue cancer are classified according to where they originate, known as histological type. According to the different types of tissue and histology, several types of cancers can be categorized into several groups, including Carcinoma, Leukemia, Lymphoma, Mixed Types, Myeloma, and Sarcoma [5].

For many years, cancer nomenclature has been primarily dictated by the anatomical origin of the tumor, such as "lung cancer," which denotes a malignancy that arises within lung tissue. However, this traditional approach often leads to late-stage detection, as the focus on organ-specific classification can compromise diagnostic precision. By the time a cancer is

identified, the affected organ systems may already be significantly impaired, rendering effective treatment and the prospect of a cure more challenging. In this context, gene selection serves as a critical tool, aiming to isolate the most pertinent genes that contribute to accurate and systematic tumor diagnosis. Despite these advancements, cancer classification continues to pose significant challenges, particularly due to the high-dimensional noise inherent in gene articulation profiles and the common issue of limited sample sizes [6]. Typically, a dataset may contain thousands of genes but only a handful of samples, with the vast majority of these genes being extraneous to the classification task. The inclusion of irrelevant genes can dilute the impact of the genuinely informative ones, thereby hindering classification performance. Hence, the significance of productive gene selection cannot be exaggerated, as it is necessary for enhancing the precision and reliability of cancer classification.

Researchers encounter challenges in early cancer detection due to the large variety of cancer types. Selecting the irrelevant genes for cancer classification is a frequent issue arising from microarray data or a high number of genes. Additionally, researchers face difficulties in handling misleading and irrelevant genes, which is a reason for the complication of the cancer classification process, resulting in increased costs and time associated with identifying the most pertinent genes. The research objective is to identify and eliminate effective selection methods that include uninformative and irrelevant genes. This approach aims to enhance the classification accuracy of cancer while minimizing the associated time and costs. The study focuses on cancer classification based on gene selection techniques, employing the XGBoost Classifier for the gene selection process. Furthermore, it aims to assess the accuracy by using the XGBoost Classifier for cancer classification through gene selection. The research targets explicitly the breast cancer dataset and aims to enhance the performance of related studies. The scope of this research is confined to cancer classification with gene selection using the XGBoost Classifier.

II. Materials and Methods

This area reviews the existing case studies that are based on XGBoost Classification for Gene Selection to classify Cancer. The study was divided into two sections: Cancer Classification and Gene Selection.

A. Gene Selection

Gene expression data is essential for uncovering hidden information for illness diagnosis, especially in cancer treatment, based on gene expression levels [7]. DNA microarrays effectively classify and predict specific types of cancer. As processing power has improved, deep learning (DL) has become common in the healthcare industry. Gene expression datasets, which often have limited samples and a large number of features, require data augmentation to overcome dimensionality issues. This paper reviews DL techniques, including Feed Forward Neural Network (FFN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Autoencoder (AE)for classifying and predicting cancer types using gene expression data analysis. Tapak et al. [8] mentioned that oral cancer (OC) significantly impacts patients' quality of life, especially those with premalignant oral lesions who are at high risk. This study aimed to use machine learning (ML) and deep learning (DL) to find predictive biomarkers for stratifying patient survival and forecasting the time-to-development of OC. An autoencoder retrieved features from 86 patients' gene expression profiles, and a Cox regression model identified the most important ones. Hierarchical clustering identified the group of low and high risk. A random forest (RF) classifier achieved 91.6% accuracy, identifying 21 top genes related to OC development, from the initial 29,096 probes.

Over 500,000 women are affected annually by Cervical cancer [9], but widespread screening is hindered by its tedious detection process. The classification of cervical precancerous cells using computer-aided diagnosis techniques is an issue that this work attempts to address. It selects features using a genetic algorithm and deep learning. The Genetic Algorithm is used to maximize the features that are extracted from restricted data using pre-trained Convolutional Neural Networks, such as GoogLeNet and ResNet-18. A Support Vector Machine (SVM) classifier achieves promising results on two public datasets, validated by 5-fold cross-validation. Akhavan and Hasheminejad [10] note that cancer diagnosis through gene analysis is a key research area in machine learning (ML) and bioinformatics. Microarray technology assesses thousands of genes simultaneously, but the challenge lies in the high gene count versus the few samples, necessitating gene selection. A two-phase gene selection procedure for microarray data is proposed in this research. Initially, genes are treated as training samples, with the gene count reduced through anomaly detection. Subsequently, a guided genetic algorithm identifies the final effective genes. Experimental results demonstrate a 99% reduction in gene expression across datasets, significantly enhancing classification accuracy.

Osama et al. [11] mentions that advancements in biotechnology have significantly improved disease diagnosis and prediction. Raw gene expression analysis, which is crucial for diagnosing conditions such as cancer, often employs small sample sizes and high-dimensional microarray data. To overcome overfitting through dimensionality reduction, this review examines recent machine learning (ML) algorithms for data reduction and classification of microarray gene expression data in tumor diagnosis. It comprehensively covers data preprocessing, feature selection, and extraction techniques, and reviews supervised, unsupervised, and semi-supervised ML algorithms. The study also addresses the difficulties and unanswered questions associated with using gene expression data to accurately classify cancer.

Kayikci and Khoshgoftaar [12] mention that, with a one-in-eight-lifetime risk, breast cancer is one of the significant causes of mortality for women. An early diagnosis is essential for successful therapy. To improve breast cancer prediction, this work presents an attention-based multimodal deep learning model that combines clinical data, copy number changes, and gene expression. The model employs a two-phase approach, utilizing dense and dropout procedures for bimodal attention, followed by a sigmoid-gated attention convolutional neural network for feature extraction. The

findings demonstrate that this approach significantly enhances the identification and detection of breast cancer, potentially improving patient outcomes.

According to Yaqoob et al. [13], gene expression datasets include a wealth of biological information, but it can be difficult to identify significant genes in high-dimensional data because of redundant and irrelevant features. This work combines SVM classifiers with the Sine Cosine and Cuckoo Search Algorithm (SCACSA) for gene selection. The feature collection is first filtered using minimum Redundancy Maximum Relevance (mRMR), and then SCACSA is used to maximize gene selection. Applied to a breast cancer dataset, SCACSA enhances classification accuracy, enabling medical practitioners to make informed decisions about cancer diagnoses by effectively navigating complex gene expression data. Pati et al. [14] investigated the impracticality of using all genes for disease classification in genomics due to time and resource constraints, as not all genes are disease related. To tackle this difficulty, this paper introduces a unique gene subset selection method called Heatmap Analysis and Graph Neural Network (HAGNN). After identifying Regions of Interest (ROIs) from microarray data using heatmap analysis, the technique reduces nodes and edges in a Graph Neural Network (GNN). The resulting gene subset, validated with base classifiers, shows that HAGNN outperforms existing methods, significantly advancing GNN-based gene selection.

B. Cancer Classification

Accurately identifying and diagnosing a patient's specific cancer or tumor type is the primary objective of cancer classification, ensuring prompt and effective treatment. The likelihood of survival and effective therapy is greatly increased by early identification. Consequently, the procedure of classifying cancer needs to be both effective and efficient. Determining the best course of treatment for the patient based on the diagnostic findings requires an accurate classification of cancer. According to the WHO, cancer is the second greatest cause of mortality worldwide. It is a category of diseases characterized by aberrant cell development and spread. Because gene expression reflects both hereditary characteristics and physiological processes, it is essential for early cancer identification.

Recent developments in cancer categorization utilize machine learning (ML), intense learning models, for their capacity to detect gene patterns [15]. Data collection, important datasets, and preprocessing methods for high-dimensional gene expression data are covered, and future research objectives in this area are discussed. Significant progress has been made in machine learning, which is now used in domains including autonomous systems, image identification, and computational linguistics. This research also highlights the practical application of machine learning [16] in cancer classification [17], emphasizing its use in medical data to classify cancer types and predict outcomes. The studies discussed the benefits and drawbacks of supervised, unsupervised, and reinforcement learning. The paper highlights how machine learning can improve cancer diagnosis and therapy, providing researchers and practitioners with information on current and emerging clinical applications.

ML has transformed breast cancer classification by enhancing the accuracy of recurrence and metastasis prediction. ML models have been widely used to classify breast cancer based on histopathological images, clinical biomarkers, and genetic data, leading to improved diagnosis and treatment strategies. Deep learning frameworks, such as the model developed by Jiang et al. [18], have demonstrated exceptional performance in detecting metastatic cancer cells in lymph nodes, achieving an AUC of 0.99, significantly enhancing TNM staging accuracy. Similarly, classification models such as Random Forest (RF) have been highly effective in predicting distant metastases, achieving an accuracy of 93.6% and an AUC of 91.3% [19]. A cross-institutional study in breast cancer recurrence classification found that AdaBoost outperformed other ML algorithms, successfully identifying key biomarkers such as CA125, CEA, fibrinogen (Fbg), and tumor diameter as the most influential predictors [20]. Furthermore, Sukhadia et al. [21] developed a classification model based on tumor size, lymph node status, and hormone receptor status, achieving 85% accuracy in predicting distant recurrence. ML has also been applied to stratify patients based on mortality risk in bone metastatic breast cancer, where a gradient-boosting tree model effectively classified high-risk patients with an AUC of 0.829 [22]. These advancements underscore the pivotal role of machine learning (ML) in breast cancer classification, facilitating more accurate diagnosis, prognosis, and personalized treatment strategies [23].

One of the leading causes of cancer-related mortality among women is metastatic breast cancer (MBC). This project [24] aims to utilize machine learning (ML) models based on blood profile data to develop a non-invasive method for identifying cancer metastases. MBC patients were identified by text mining from Electronic Medical Records (EMR), which showed notable variations in monocyte counts. A Decision Tree (DT) classifier obtained 83% accuracy and an AUC of 0.87 after eliminating outliers. To help doctors improve patient survival outcomes, the DT model was implemented through a web application for reliable MBC diagnosis. The most deadly type of skin cancer, melanoma, is on the rise. A deep learning method for melanoma lesion identification with a GPU-equipped server is presented in this paper [25]. Images of malignant or nonmalignant melanoma are classified using a convolutional neural network (CNN) that has been pre-trained on sizable datasets. The method has shown promise in laboratory settings and is intended to aid dermatologists in early detection. The suggested technique's potential for clinical applications is demonstrated by experimental results that show it outperforms state-of-the-art procedures in terms of diagnostic accuracy. Improving patient survival rates for breast cancer (BC) requires early detection and precise diagnosis.

The study by [26] has proposed a deep learning model (BCCNN) to classify breast cancer MRI images into eight categories, including benign and malignant. The model, alongside five fine-tuned pre-trained models (Xception, InceptionV3, VGG16, MobileNet, and ResNet50), was evaluated using a Kaggle dataset enhanced by GAN techniques. The models were tested across different magnifications, yielding F1-score accuracies of 97.54%, 95.33%, 98.14%, 97.67%, 93.98%, and 98.28% for each

respective model. Breast cancer, a significant public health issue, requires early diagnosis for effective treatment. The research by [27] investigates the application of deep learning (DL) and machine learning (ML) methods in five distinct medical imaging modalities—thermography, histology, MRI, ultrasound, and mammography—for the classification and detection of breast cancer. The utilization of CNNs, ANN, DT, Naive Bayesian Network, SVM, Nearest Neighbor, and deep learning architectures is highlighted in the paper. Results indicate that these methods have a high accuracy rate and can enhance patient outcomes and inform clinical decision-making.

Cancer types such as breast, lung, skin, and blood malignancies (e.g., leukemia and lymphoma) exhibit uncontrolled cell proliferation. Acute lymphoblastic leukemia (ALL) is a significant malignancy that can be challenging to diagnose. Using machine learning (ML) and deep learning (DL), the study by [28] presents a novel method for classifying leukemia. Dataset construction, feature extraction with CNN models that have already been trained, and classification with traditional classifiers are all part of the technique. Four classes make up the dataset: Pro-B, Pre-B, Early Pre-B, and Benign. The study employed the ResNet50 CNN architecture with LR classifiers to achieve a maximum accuracy of 99.84% by combining nature-inspired algorithms, such as PSO and CSO, which could lead to advancements in the real-world categorization of blood cancer. One of the leading causes of death for women globally is breast cancer (BC), and lowering death rates requires early detection.

The work by [29] presents a Deep Reinforcement Learning (DRL)-based BC classification algorithm utilizing large datasets. The model utilizes LIME to describe the results after processing and normalizing the data, selects features using the Gorilla Troops Optimization (GTO) algorithm, and classifies the data using Deep Q Learning (DQL). The GTO-DQL model outperforms conventional techniques with accuracy rates of 98.90%, 99.02%, and 98.88%, respectively, when tested on three UCI datasets (WBCD, WDBC, and WPBC). With about 8% of cases detected, breast cancer is still the second most common cause of mortality for women, after lung cancer. A timely and precise diagnosis is essential because it can show up as discomfort, skin abnormalities, and genetic mutations. The Extreme Gradient Boosting (XGBoost) machine learning technique is employed in this work [30] to enhance the rapid and accurate identification of breast cancer. When used on the Wisconsin breast cancer (diagnostic) dataset, XGBoost demonstrated its efficacy in early diagnosis with an accuracy of 94.74% and a recall of 95.24%.

Every year, more than 2.1 million women worldwide are afflicted with breast cancer (BC). Increasing survival rates requires an early and precise diagnosis. With an emphasis on mammography and histopathologic images, the state of Deep Neural Network (DNN) methods for breast cancer (BC) detection, classification, and segmentation using medical imaging is investigated [31]. It highlights the advantages and drawbacks of various imaging modalities, as well as pre-processing techniques such as scaling, normalization, and data augmentation. The study finds that Convolutional Neural Networks (CNNs) are widely used, with both pre-trained and custom models. Additionally, it identifies 13 significant challenges for future research in BC diagnosis.

Using several CNNs and meta-learning, the study by [32] aims to develop an effective breast cancer classification model on the Breast Ultrasound Images (BUSI) dataset, which encompasses a range of breast abnormalities. Traditional approaches often struggle with the dataset's complexity. The proposed model integrates meta-learning for optimized learning adaptation and transfer learning with Inception, ResNet50, and DenseNet121 for enhanced feature extraction and data augmentation to diversify the dataset. Meta ensemble learning further improves classification accuracy by combining CNN outputs. The study involves dataset preprocessing, training CNNs, applying meta-learning for optimization, and evaluating performance metrics, such as F1 score, recall, precision, and accuracy, against existing systems.

C. Methodology

This research investigates the application of cancer classification using gene selection methods based on the XGBoost classifier, capitalizing on the potential of gene expression profiles for disease diagnostics. A significant challenge in this domain is the disparity between the vast number of genes and the limited size of available datasets. Small sample sizes can compromise classification accuracy due to the inclusion of redundant and uninformative genes, thereby increasing false-positive rates. Employing the XGBoost Classifier addresses this issue by effectively identifying genes that significantly contribute to cancer classification. The study isolates and prioritizes the most informative genes through rigorous preprocessing, which is essential for accurate cancer classification. Subsequently, a search approach is applied to refine this selection, aiming to identify a concise yet highly informative gene subset that optimizes cancer classification accuracy. This methodology not only enhances the precision of cancer diagnostics also streamlines the gene selection process, providing a robust framework for future research and clinical applications.

This study utilized the XGBoost Classifier to identify genes critical for cancer classification. The process began with the input of the original dataset or initial feature subset, followed by the initialization of the classifier. Data processing was conducted to minimize training errors and enhance classification accuracy by ensuring each variable received equal weight. Following this, the stages of population initialization, crossover, mutation, and fitness function setup were performed to prepare for subsequent analyses. Fitness calculations assessed each individual's gene combination based on their specific fitness values, with higher-fitness individuals progressing to the next generation. Selecting the best individuals involves a prefiltering step to narrow down informative genes. The XGBoost Classifier was then employed to optimize the gene subset, retaining only those with scores greater than zero and excluding those with irrelevant scores. For evaluation, Support Vector Machines (SVM) were utilized to assess accuracy, which is particularly suited for small-sample and high-dimensional data classification tasks. The study's results included accuracy rates and graphical representations of cancer classification

outcomes, highlighting the effectiveness of the proposed methodology.

1) Dataset:

In this study, the primary dataset is the Breast Cancer Dataset, which consists of 151 columns representing samples and 54,676 rows representing genes. The dataset comprises six distinct classes: normal, cell line, luminal A, luminal B, HER2, and basal. This chapter provides a comprehensive discussion on the research methodology, elucidating the experimental setup and operational mechanisms of the algorithm employed. Specifically, it focuses on the methodological approach centered around the XGBoost Classifier, detailing the processes of data collection and outlining the anticipated evaluation metrics for assessing the chosen methodology's effectiveness. This rigorous approach ensures clarity in the experimental design and robustness in the results obtained. The overall flow for the method is given in Figure 1.

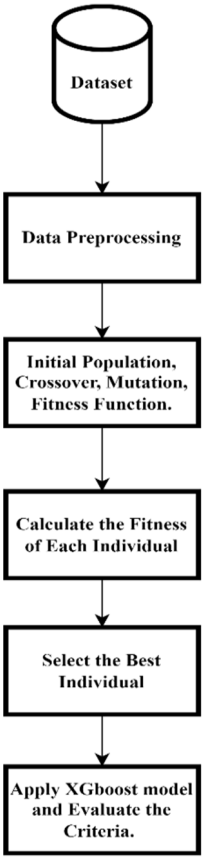

Fig. 1 Methodology of this research

This section presents a comprehensive analysis of the findings from the implementation and testing of the XGBoost classifier method. It includes detailed explanations of the employed methodology and the outcomes derived from the conducted experiments. XGBoost is extensively utilized in cancer prediction and other domains due to its capacity to produce highly accurate models and its versatility in handling diverse types of data. Its ensemble learning strategy, which combines several weak learners to create a strong and accurate prediction model, is responsible for its efficacy. Extreme Gradient Boosting, also known as XGBoost, is a powerful and effective algorithm recognized for its outstanding results in various machine learning applications, including cancer prediction. It is a member of the gradient boosting algorithm family, which builds an ensemble of weak learners—usually decision trees—sequentially to improve prediction accuracy.

2) Feature Extraction:

To classify the cancer, this research introduced an XGBoost classifier to utilize the dataset for testing and training, thereby selecting the gene [33],[34]. Several essential modules are imported for this process, including Pipeline, MultiOutputClassifier, KFold, XGBClassifier, roc_auc score, make multilabel classification, classification report, confusion matrix, and train-test split. These modules facilitate various functions such as dataset division, classifier definition, accuracy calculation, and the generation of classification reports. Following the module importation, the dataset is split into train and test datasets at a ratio of 70% and 30%, respectively.

3) Evaluation

The proposed XGBoost classifier is evaluated using key metrics such as F1 Score, Recall, Precision, and Accuracy. These metrics provide a comprehensive assessment of the model's performance, with equations provided for each metric. The results are compared with previous methodologies, highlighting the improvements achieved by integrating recursive feature elimination with cross-validation (RFECV). The equations for F1 Score [35], Recall [36], Precision [37] and Accuracy [38] are provided in Equations (1)-(4), respectively.

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

Precision assesses the proportion of predictions in the Positive class that align with the ground truth, effectively measuring a classifier's ability to avoid misclassifying negative samples as positive. It provides crucial insights into the model's accuracy specifically within the Positive class, offering a vital perspective on its classification performance.

$$\text{Precision} = \frac{TruePostive}{TruePostive+FalsePostive} \quad (2)$$

The positive class which correctly matches the truth of ground among all actual samples of positive, recall helps to measure the prediction proportion. It evaluates a classifier's ability to correctly identify positive instances, providing insights into its sensitivity to detecting relevant samples within the dataset.

$$Recall = \frac{TruePostive}{TruePostive+FalseNegative}) \quad (3)$$

To evaluate F1 score, by assessing the balance between recall and precision that provides a single metric. It quantifies the accuracy of positive predictions, where a value of 1.0 indicates optimal performance and 0.0 indicates the lowest.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

III. Results and Discussion

This section discusses the testing and results, encompassing the tested datasets, measurement methods employed, and the

findings of the research. Results are compared between previous methodologies and the proposed approach, which integrates RFECV. The RFECV method enhances the robustness of feature selection and model performance evaluation. The discussion highlights the efficacy of the new method in enhancing predictive accuracy and underscores its potential impact on cancer prediction models.

The results from previous research are compared with the findings of this study in Table 1. The dataset used is divided into six categories, labeled from 0 to 5. The classifier's recall, also known as sensitivity, measures how well it can recognize every positive case. For each category, it is computed as the ratio of true positives to the sum of true positives and false negatives. Recall essentially indicates the proportion of real positives that the classifier correctly detects. A comprehensive indicator of model performance is provided by the F1 Score, which is the harmonic mean of recall and precision. An F1 Score approaching 1.0 signifies excellent performance, with 1.0 being ideal and 0.0 being the least desirable. Furthermore, the macro average of the F1 Score provides an overall performance metric across all categories, with a higher average indicating better performance. Support refers to the count of instances within each category in the dataset and is used in evaluating model performance, though it does not influence the comparison of models. For example, a support value of 9 for category 0 means that there are nine instances where category 0 is present in the dataset.

TABLE I
PROPOSED RESEARCH AND PREVIOUS RESEARCH RESULT ANALYSIS

Research Duration	Area	Precision	Recall	F1-score	Support
Previous Research	0	0.50	0.67	0.57	9
	1	0.93	0.72	0.81	18
	2	0.83	1.00	0.91	5
	3	1.00	0.85	0.92	13
	4	1.00	0.43	0.60	14
	5	1.00	1.00	1.00	2
	micro avg	0.84	0.70	0.77	61
	macro avg	0.88	0.78	0.80	61
	weighted avg	0.89	0.70	0.76	61
	samples avg	0.68	0.70	0.69	61
	Accuracy		0.875		
Proposed Research	0	0.86	0.67	0.75	9
	1	0.94	0.83	0.88	18
	2	0.80	0.80	0.80	5
	3	1.00	1.00	1.00	13
	4	0.93	0.93	0.93	14
	5	1.00	1.00	1.00	2
	micro avg	0.93	0.87	0.90	61
	macro avg	0.92	0.87	0.89	61
	weighted avg	0.93	0.87	0.90	61
	samples avg	0.84	0.87	0.85	61
	Accuracy		0.9289		

A. Previous Work with XGBoost Classifier

Following the classification report, various graphs were created utilizing Principal Component Analysis (PCA) to visualize the results. PCA is predominantly used to reduce the dimensionality or number of features within a dataset. In this study, 90 components were initially analyzed in earlier work. For the discussion of results, the focus is placed on the top three PCA components. The graph below compares PCA 1 and PCA 2 from the previous analysis, highlighting the impact of dimensionality reduction on data interpretation and visualization. Figures 2 through 5, which are presented below, were also generated from prior research. Specifically, Figure 2 compares PCA 1 with PCA 2, Figure 3 contrasts PCA 2 with PCA 3, and Figure 4 illustrates the comparison between PCA 1 and PCA 3.

Figure 5 offers a three-dimensional visualization using PCA, comparing the first three principal components: PCA 1, PCA 2, and PCA 3. These graphical representations adeptly illustrate the dimensionality reduction achieved through PCA, providing a clearer understanding of the data structure across these key components. The distinction between two-dimensional and three-dimensional PCA visualizations lies in the number of principal components employed. Principal Component Analysis (PCA) identifies and constructs these components to capture the maximum variance within the dataset, with PC1 capturing the highest variance, followed by PC2, and so forth. Typically, the first two or three components are sufficient to explain the majority of the variance, making it feasible to disregard additional components without a substantial loss of information. Although PCA is not designed as a clustering tool, its ability to reduce dimensionality aids in the visualization of patterns, potentially revealing clusters of gene expression profiles that share similar characteristics. Patterns that might be subtle or indistinct in a 2D PCA plot often become more discernible in a 3D context. However, an analysis of the 3D PCA results in this study reveals that the gene expression profiles do not form well-defined clusters, which complicates the identification of distinct classes or groups. This indicates that the 3D PCA visualization portrays a more complex and dispersed distribution of gene expressions, making it more challenging to pinpoint outliers that could merit further exploration.

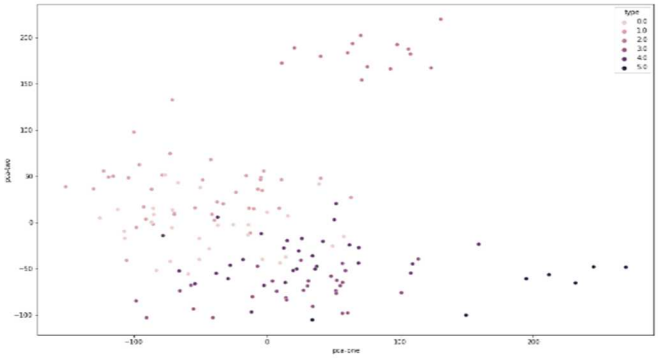

Fig. 2 Comparison from earlier work of PCA 1 and PCA 2 with PGA 2D

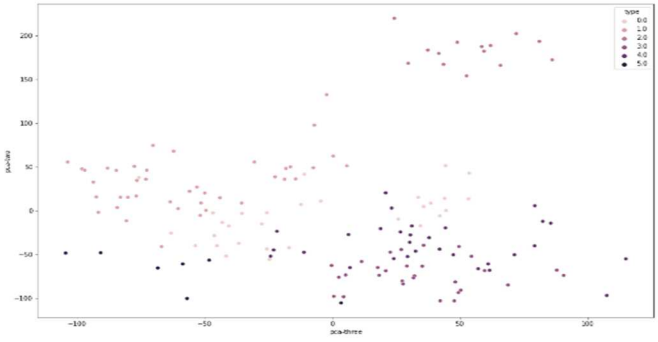

Fig. 3 Comparison from earlier work of PCA 3 and PCA 2 with PGA 2D

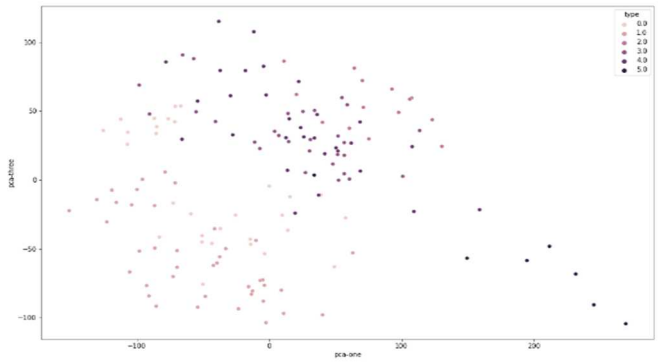

Fig. 4 Comparison from earlier work of PCA 1 and PCA 3 with PGA 2D

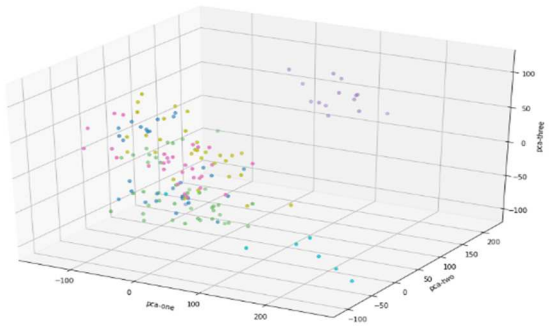

Fig. 5 Comparison from earlier work of PCA 1, 2, and 3 with PGA 3D

B. Recursive Feature Elimination and XGBoost Classifier Proposed Method with Cross Validation.

After the classification report, several PCA graphs were generated. In the proposed method, 63 principal components were used for the PCA analysis. Figure 6 displays the comparison between PCA 1 and PCA 2, while Figure 7 compares PCA 3 and PCA 2. Figure 8 illustrates the comparison between PCA 1 and PCA 3. A close examination of the PCA plots generated by the proposed method reveals that the clusters corresponding to each class demonstrate a stronger association based on gene expression profiles. Analyzing the spatial separation between these clusters proves to be a more effective approach for detecting outliers than focusing on individual variables. The visualizations produced with the updated code clearly show more distinct and cohesive gene clusters compared to those generated by the original code. In this improved method, genes with similar expression profiles are more consistently grouped, while those that do not conform to these clusters are distinctly identified as outliers.

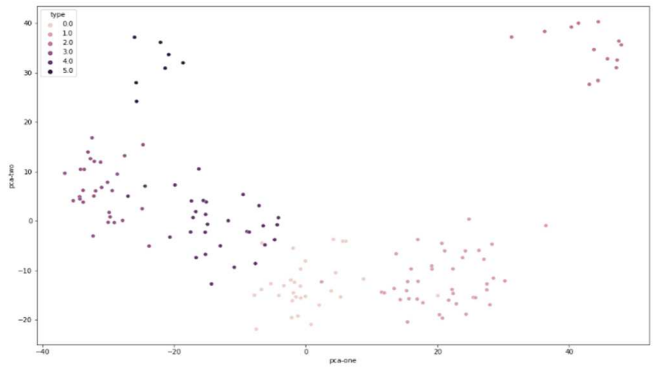

Fig. 6 Comparison of the PCA 1 and PCA 2 work from this study with the PGA 2D.

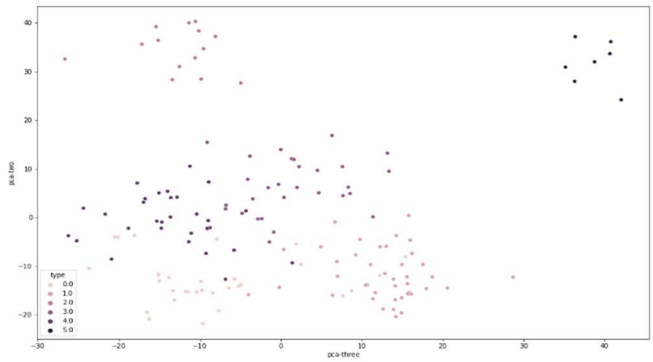

Fig. 7 Comparison from this study of PCA 3 and PCA 2 with PGA 2D

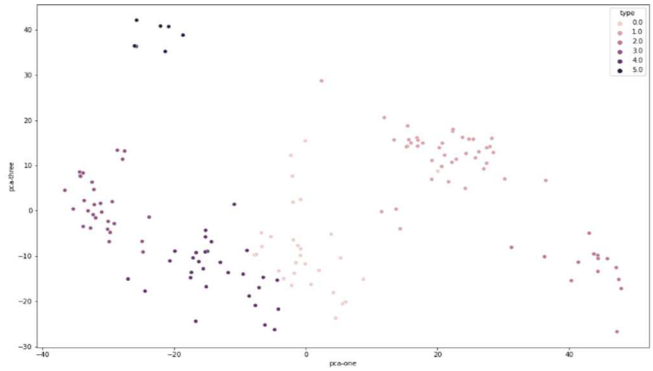

Fig. 8 Comparison from this study of PCA 1 and PCA 3 with the PGA 2D

The comparison between PCA 1 and PCA 2, for example, clearly illustrates that the clusters generated by the proposed method are more distinct and easily identifiable. This enhancement suggests that the proposed method more effectively grouping genes according to their respective classes, resulting in a reduced number of misclassifications across different courses. Figure 9 further exemplifies this improvement through a 3D PCA plot, which vividly captures the refined clustering achieved by the new approach. Upon observation, more identifiable gene clusters are clearly shown based on the different classes in the 3D PCA plot. According to the plot, it is evident that the proposed method's accuracy is higher than that of the previous one, as indicated by the distinct and well-defined clusters of genes. This enhanced clustering suggests a more precise classification of genes, reflecting the improved performance of the proposed method.

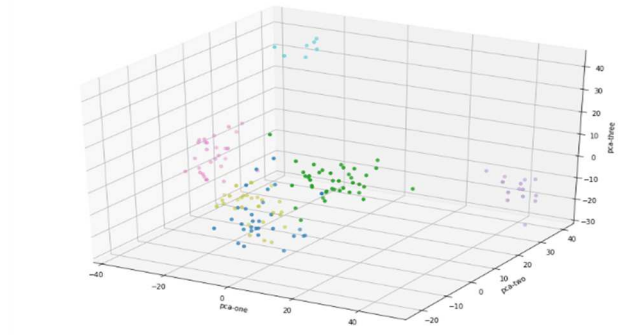

Fig. 9 Comparison from this study of PCA 1, 2, and 3 with PGA 3D

In conclusion, this section presents the results and findings derived from implementing the code. The results consistently indicate that the proposed method, which incorporates contributions from recursive feature elimination with cross-

validation (RFECV), achieves higher accuracy compared to previous work. This enhancement demonstrates that RFECV significantly improves the performance for cancer classification using XGBoost classifier.

IV. Conclusion

This research paper introduces an innovative method for gene selection in cancer classification by employing the XGBoost Classifier. The use of microarray technology enables the development of extensive databases of cancerous tissues based on gene expression data. However, a common challenge in cancer classification lies in the fact that training datasets typically contain a limited number of samples and span multiple categories, which is disproportionate to the vast number of genes involved. In this study, a breast cancer dataset was employed. The most significant genes were identified using RFECV, which was then followed by cancer classification using the XGBoost classifier. The findings indicate that the integration of RFECV with XGBoost significantly enhances both the accuracy of gene selection and the overall performance in cancer classification, compared to using the XGBoost classifier in isolation.

Although the proposed methodology has proven to be effective, further improvements can be made by enhancing the feature elimination method or the search approach. For instance, integrating feature selection methods such as Ant Colony Optimization could potentially increase the effectiveness of the gene selection process. In conclusion, the primary objective is to enhance the accuracy of gene selection for cancer classification and to minimize the inclusion of irrelevant genes before proceeding to subsequent analysis stages. Future work will focus on refining these methodologies to achieve even greater accuracy and efficiency in cancer classification.

Acknowledgment

This research was communicated through monetary assistance from Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216. The authors thank the Ministry of Higher Education Malaysia for providing financial support under Universiti Malaysia Pahang Al-Sultan Abdullah for laboratory facilities and additional financial support under the Internal Research grant RDU2303103.

References

[1] J. Ferlay et al., "Cancer statistics for the year 2020: An overview," *Int. J. Cancer*, vol. 149, no. 4, pp. 778-789, Aug. 2021, doi:10.1002/ijc.33588.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA: Cancer J. Clin.*, vol. 71, no. 1, pp. 7-33, Jan. 2021, doi: 10.3322/caac.21654.

[3] M. J. Kang et al., "Cancer statistics in Korea: Incidence, mortality, survival, and prevalence in 2019," *Cancer Res. Treat.*, vol. 54, no. 2, pp. 330-344, Mar. 2022, doi: 10.4143/crt.2022.128.

[4] N. Mahendran, P. M. D. R. Vincent, K. Srinivasan, and C.-Y. Chang, "Machine learning based computational gene selection models: A survey, performance evaluation, open issues, and future research directions," *Front. Genet.*, vol. 11, Dec. 2020, doi:10.3389/fgene.2020.603808.

[5] M. Lamba, G. Munjal, Y. Gigras, and M. Kumar, "Breast cancer prediction and categorization in the molecular era of histologic grade," *Multimed. Tools Appl.*, vol. 82, no. 19, pp. 29629-29648, Aug. 2023, doi: 10.1007/s11042-023-14918-9.

[6] M. Lee, "Recent advances in generative adversarial networks for gene expression data: A comprehensive review," *Mathematics*, vol. 11, no. 14, Jul. 2023, doi: 10.3390/math11143055.

[7] U. Ravindran and C. Gunavathi, "A survey on gene expression data analysis using deep learning methods for cancer diagnosis," *Prog. Biophys. Mol. Biol.*, vol. 177, pp. 1-13, Jan. 2023, doi:10.1016/j.pbiomolbio.2022.08.004.

[8] L. Tapak et al., "Identification of gene profiles related to the development of oral cancer using a deep learning technique," *BMC Med. Genomics*, vol. 16, no. 1, Feb. 2023, doi: 10.1186/s12920-023-01462-6.

[9] R. Kundu and S. Chattopadhyay, "Deep features selection through genetic algorithm for cervical pre-cancerous cell classification," *Multimed. Tools Appl.*, vol. 82, no. 9, pp. 13431-13452, Apr. 2023, doi: 10.1007/s11042-022-13736-9.

[10] M. Akhavan and S. M. H. Hasheminejad, "A two-phase gene selection method using anomaly detection and genetic algorithm for microarray data," *Knowl.-Based Syst.*, vol. 262, Feb. 2023, doi:10.1016/j.knosys.2022.110249.

[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, Mar. 2023, doi: 10.1016/j.eswa.2022.118946.

[12] S. Kayikci and T. M. Khoshgoftaar, "Breast cancer prediction using gated attentive multimodal deep learning," *J. Big Data*, vol. 10, no. 1, May 2023, doi: 10.1186/s40537-023-00749-w.

[13] A. Yaqoob, N. K. Verma, and R. M. Aziz, "Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm," *J. Med. Syst.*, vol. 48, no. 1, Jan. 2024, doi:10.1007/s10916-023-02031-1.

[14] S. K. Pati, A. Banerjee, and S. Manna, "Gene selection of microarray data using heatmap analysis and graph neural network," *Appl. Soft Comput.*, vol. 135, Mar. 2023, doi: 10.1016/j.asoc.2023.110034.

[15] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, Jan. 2023, doi: 10.3390/bioengineering10020173.

[16] N. F. Idris et al., "Stacking with recursive feature elimination-isolation forest for classification of diabetes mellitus," *PLoS ONE*, vol. 19, no. 5, May 2024, doi: 10.1371/journal.pone.0302595.

[17] A. Yaqoob, R. M. Aziz, and N. K. Verma, "Applications and techniques of machine learning in cancer classification: A systematic review," *Hum.-Centric Intell. Syst.*, vol. 3, no. 4, pp. 588-615, Sep. 2023, doi: 10.1007/s44230-023-00041-3.

[18] B. Jiang et al., "Deep learning applications in breast cancer histopathological imaging: Diagnosis, treatment, and prognosis," *Breast Cancer Res.*, vol. 26, no. 1, Sep. 2024, doi:10.1186/s13058-024-01895-6.

[19] H. Duan et al., "Machine learning-based prediction model for distant metastasis of breast cancer," *Comput. Biol. Med.*, vol. 169, Feb. 2024, doi: 10.1016/j.compbiomed.2024.107943.

[20] D. Zuo et al., "Machine learning-based models for the prediction of breast cancer recurrence risk," *BMC Med. Inform. Decis. Mak.*, vol. 23, 2023, doi: 10.1186/s12911-023-02377-z.

[21] S. S. Sukhadia et al., "Machine learning-based prediction of distant recurrence in invasive breast carcinoma using clinicopathological data: A cross-institutional study," *Cancers*, vol. 15, no. 15, 2023, doi:10.3390/cancers15153960.

[22] F. Xiong et al., "A machine learning-based model to predict early death among bone metastatic breast cancer patients: A large cohort of 16,189 patients," *Front. Cell Dev. Biol.*, vol. 10, Dec. 2022, doi:10.3389/fcell.2022.1059597.

[23] M. A. Ismail, A. O. Ibrahim, and S. Jebaraj, "Machine learning in healthcare: Transformative applications, challenges, and future directions," *Front. Health Inform.*, vol. 13, no. 2, 2024.

[24] M. Botlagunta et al., "Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms," *Sci. Rep.*, vol. 13, no. 1, Jan. 2023, doi: 10.1038/s41598-023-27548-w.

[25] S. R. Waheed et al., "Melanoma skin cancer classification based on CNN deep learning algorithms," *Malays. J. Fundam. Appl. Sci.*, vol. 19, no. 3, pp. 299-305, May 2023, doi: 10.11113/mjfas.v19n3.2900.

[26] B. Abunasser et al., "Convolution neural network for breast cancer detection and classification using deep learning," *Asian Pac. J. Cancer Prev.*, vol. 24, no. 2, pp. 531-544, Feb. 2023, doi:10.31557/APJCP.2023.24.2.531.

[27] M. Radak, H. Y. Lafta, and H. Fallahi, "Machine learning and deep learning techniques for breast cancer diagnosis and classification: A

comprehensive review of medical imaging studies," *J. Cancer Res. Clin. Oncol.*, vol. 149, no. 12, pp. 10473-10491, Sep. 2023, doi:10.1007/s00432-023-04956-z.

[28] W. Rahman et al., "Multiclass blood cancer classification using deep CNN with optimized features," *Array*, vol. 18, Jul. 2023, doi:10.1016/j.array.2023.100292.

[29] S. Almutairi et al., "Breast cancer classification using deep Q learning (DQL) and gorilla troops optimization (GTO)," *Appl. Soft Comput.*, vol. 142, Jul. 2023, doi: 10.1016/j.asoc.2023.110292.

[30] R. Hoque, S. Das, and M. Hoque, "Breast cancer classification using XGBoost," *World J. Adv. Res. Rev.*, vol. 21, no. 2, pp. 1985-1994, Feb. 2024, doi: 10.30574/wjarr.2024.21.2.0625.

[31] B. Abhisheka, S. K. Biswas, and B. Purkayastha, "A comprehensive review on breast cancer detection, classification and segmentation using deep learning," *Arch. Comput. Methods Eng.*, vol. 30, no. 8, pp. 5023-5052, Nov. 2023, doi: 10.1007/s11831-023-09968-z.

[32] M. D. Ali et al., "Breast cancer classification through meta-learning ensemble technique using convolution neural networks," *Diagnostics*, vol. 13, no. 13, Jun. 2023, doi: 10.3390/diagnostics13132242.

[33] T. Bansal and N. Jindal, "An improved hybrid classification of brain tumor MRI images based on conglomeration feature extraction techniques," *Neural Comput. Appl.*, vol. 34, no. 11, pp. 9069-9086, Jun. 2022, doi: 10.1007/s00521-022-06929-8.

[34] J. Hariharan, Y. Ampatzidis, J. Abdulridha, and O. Batuman, "Useful feature extraction and machine learning techniques for identifying unique pattern signatures present in hyperspectral image data," in *Hyperspectral Imaging-A Perspective on Recent Advances and Applications*, IntechOpen, Dec. 2022, doi:10.5772/intechopen.107436.

[35] S. Srivastava et al., "Comparative analysis of deep learning image detection algorithms," *J. Big Data*, vol. 8, no. 1, May 2021, doi:10.1186/s40537-021-00434-w.

[36] T. Pham and L. M. Archibald, "The role of working memory loads on immediate and long-term sentence recall," *Memory*, vol. 31, no. 1, pp. 61-76, Jan. 2023, doi: 10.1080/09658211.2022.2122999.

[37] S. A. Bhat and N. F. Huang, "Big data and AI revolution in precision agriculture: Survey and challenges," *IEEE Access*, vol. 9, pp. 110209-110222, Aug. 2021, doi: 10.1109/access.2021.3102227.

[38] L. C. Ngugi, M. Abelwahab, and M. Abo-Zahhad, "Recent advances in image processing techniques for automated leaf pest and disease recognition-A review," *Inf. Process. Agric.*, vol. 8, no. 1, pp. 27-51, Mar. 2021, doi: 10.1016/j.inpa.2020.04.004.