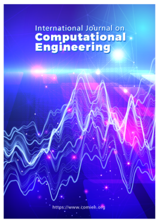
comien

International Journal on
Computational Engineering

Journal Homepage www.comien.org/index.php/comien



Gene Selection for Cancer Classification Based on XGBoost

Teo Voon Chuan [a], Md Raihanul Islam Tomal [a], Kohbalan Moorthy [a], Chan Weng Howe [b]

[a] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia.*
[b] *Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia.*
*Corresponding author: *kohbalan@umpsa.edu.my*

***Abstract*— Cancer remains a leading cause of death worldwide; World Health Organization (WHO) referees 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 further enhancing the accuracy of gene selection for cancer classification and reducing the number of irrelevant genes before proceeding to subsequent processes. This approach holds promise for streamlining the diagnostic process, improving patient outcomes, and offering significant benefits in the timely treatment of cancer.**

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

Manuscript received 6 Jan. 2025; revised 25 Feb. 2025; accepted 24 Mar. 2025. Date of publication 30 Jun. 2025.

I. Introduction

In this era of globalization, cancer remains a leading cause of mortality worldwide, irrespective of the country. According to the World Health Organization (2024), there have been nearly 10 million cancer-related deaths in recent years, with breast cancer (BC) affecting over 2.1 million women annually on a global scale. The most prevalent cancers in recent years, as reported by WHO, include breast, lung, and prostate cancers. Although there are numerous types of cancer, early detection and treatment can significantly improve the chances of a cure. Gene selection plays a crucial role in aiding researchers in cancer classification. Gene selection involves reducing redundant and less informative genes in a gene expression dataset, such as a DNA microarray, to ensure that the selected genes are directly related to disease diagnosis [1]. This technique is essential for identifying a set of relevant genes associated with a specific disease. Simply put, gene selection is employed to identify the informative and significant genes pertinent to clinical diagnoses, such as cancer. Cancer classification involves identifying the type of cancer and determining the extent to which a tumor has grown and spread. Generally, cancers are classified based on the type of tissue from which they originate, known as histological type. Based on histology or tissue type, hundreds of distinct cancers can be categorized into several groups, including Carcinoma, Leukemia, Lymphoma, Mixed Types, Myeloma, and Sarcoma [2].

Traditionally, cancer nomenclature has focused primarily on organ location; for instance, "lung cancer" refers to a tumor that originates in the lung tissues. Consequently, many cancers are detected in their later stages due to accuracy problems, resulting in compromised or malfunctioning organ systems, which makes achieving a cure difficult even after treatment. Gene selection aims to identify the most relevant genes that assist in diagnosing tumors precisely and systematically. However, cancer classification remains challenging due to the high-dimensional noise in gene expression profiles and the issue of small sample sizes. This means that a dataset will typically contain thousands of genes but only a few samples, with most of these genes being irrelevant to the classification task. Including all genes in the analysis can impede classification performance by obscuring the contribution of informative genes. Therefore, effective gene selection is crucial for improving the accuracy of cancer classification.

Based on the problem background, the problem statement of this research is articulated as follows. First, researchers face obstacles in making early cancer detections due to the vast number of cancer types. Additionally, the selection of irrelevant genes for cancer classification is a common issue, arising from the high dimensionality of gene or microarray data. Furthermore, researchers encounter difficulties in dealing with irrelevant and misleading gene expression samples, which complicates the cancer classification process and increases both the time and cost of identifying the most associated genes. The aim of this research is to minimize the selection of irrelevant and less informative genes through effective gene selection methods. By doing so, it seeks to enhance the accuracy of cancer classification while reducing the time and cost involved. This study focuses on cancer classification based on gene selection approaches, implementing the XGBoost Classifier in the gene selection process for cancer classification. Additionally, it aims to verify the performance of gene selection for cancer classification using the XGBoost Classifier. The research specifically focuses on the breast cancer dataset and aims to improve the accuracy of relevant research. The scope of this study is limited to cancer classification with gene selection using the XGBoost Classifier.

A. Literature Review

This area review the existing case studies which is Gene Selection for Cancer Classification Based on XGBoost Classifier. The study divided into two section, Gene Selection and Cancer Classification.

B. Gene Selection

Ravindran U et al. [3] mentioned, gene expression data is crucial for extracting hidden information for disease diagnosis, particularly in cancer treatment, based on gene expression levels. DNA microarray efficiently classifies and predicts specific cancer types. Deep learning (DL) has become prevalent in healthcare due to increased computing power. Gene expression datasets, with limited samples and many features, require data augmentation to overcome dimensionality issues. This paper reviews DL techniques, including Feed Forward Neural Network (FFN), Convolutional Neural Network (CNN), Autoencoder (AE), and Recurrent Neural Network (RNN), for classifying and predicting cancer types using gene expression data analysis. Leili Tapak et al. [4] mentioned, oral cancer (OC) significantly impacts patients' quality of life, especially those with oral premalignant lesions who are at high risk. This study aimed to identify prognostic biomarkers for predicting the time-to-development of OC and stratifying patient survival using machine learning and deep learning. Using gene expression profiles from 86 patients, an autoencoder extracted features, and a Cox regression model selected significant ones. Hierarchical clustering identified high-risk and low-risk groups. A random forest classifier achieved 91.6% accuracy, identifying 21 top genes related to OC development, from the initial 29,096 probes.

Cervical cancer affects over 500,000 women annually [5], but widespread screening is hindered by its tedious detection process. This study addresses the challenge of classifying cervical pre-cancerous cells using computer-aided diagnosis tools. It employs Deep Learning and a Genetic Algorithm for feature selection. Pre-trained Convolutional Neural Networks, GoogLeNet and ResNet-18, extract features from limited data, which are then optimized using the Genetic Algorithm. A Support Vector Machines classifier achieves promising results on two public datasets, validated by 5-fold cross-validation. Motahare Akhavan et al. [6] mention, Cancer diagnosis via gene analysis is a key research area in bioinformatics and machine learning. Microarray technology assesses thousands of genes simultaneously, but the challenge lies in the high gene count versus few samples, necessitating gene selection. This paper proposes a two-phase gene selection method for microarray data. Initially, genes are treated as training samples, reducing gene count through anomaly detection. Subsequently, a guided genetic algorithm identifies the final effective genes. Experimental results show a 99% gene reduction across datasets, significantly enhancing classification accuracy. Sarah Osama et al. [7] mention that advancements in biotechnology have significantly improved disease diagnosis and prediction. Analyzing raw gene expression, crucial for identifying diseases like cancer, often involves high-dimensional microarray data with small sample sizes. This review examines recent machine learning (ML) algorithms for data reduction and classification of microarray gene expression data to diagnose tumors, addressing overfitting through dimensionality reduction. It comprehensively covers data preprocessing, feature selection, and extraction techniques, and reviews supervised, unsupervised, and semi-supervised ML algorithms. Additionally, the paper discusses the challenges and open questions in gene expression data for accurate cancer classification. Safak Kayikci et al. [8] mention as, breast cancer is a leading cause of death among women, with a lifetime risk of one in eight. Early diagnosis is crucial for effective treatment. This study introduces an attention-based multimodal deep learning model that integrates clinical data, copy number alterations, and gene expression to enhance breast cancer prediction. The model employs a two-phase approach: a sigmoid gated attention convolutional neural network for feature generation, followed by dense and dropout processes for bi-modal attention. Results indicate that this methodology significantly improves breast cancer detection and diagnosis, potentially leading to better patient outcomes.

Abrar Yaqoob et al. [9] mention gene expression datasets contain vast biological information, yet identifying crucial genes within high-dimensional data is challenging due to redundant and unimportant features. This study introduces the Sine Cosine and Cuckoo Search Algorithm (SCACSA) for gene selection, paired with Support Vector Machine (SVM) classifiers. Initially, minimum Redundancy Maximum Relevance (mRMR) filters the feature set, followed by SCACSA to optimize gene selection. Applied to a breast cancer dataset, SCACSA enhances classification accuracy, aiding medical practitioners in making informed cancer diagnosis decisions by effectively navigating complex gene expression data. Soumen et al. [10] investigating all genes for disease classification in genomics is impractical due to time and resource constraints, as not all genes are disease-related. This study presents a novel gene subset selection technique, Heatmap Analysis and Graph Neural Network (HAGNN), to address this challenge. The method involves heatmap analysis to identify Regions of Interest (ROIs) from microarray data, followed by node and edge reduction 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

C. Cancer Classification

The primary goal of cancer classification is to accurately identify and diagnose the specific type of cancer or tumor in a patient to ensure timely and appropriate treatment. Early detection significantly increases the chances of survival and successful treatment. Therefore, the cancer classification process must be both efficient and effective. Accurate cancer classification is crucial for determining the most suitable treatment plan for the patient based on the diagnostic results. Cancer, a group of diseases marked by abnormal cell growth and spread, is the second leading cause of death globally, per the WHO. Gene expression analysis is crucial for early cancer detection, reflecting biochemical processes and genetic traits. This study [11] reviews recent advancements in cancer classification using machine learning, particularly deep learning models, for their ability to identify gene patterns. It covers data collection, key datasets, and preprocessing techniques for high-dimensional gene expression data, concluding with future research directions in this field. Machine learning has seen significant advancements, finding applications in fields like computational linguistics, image identification, and autonomous systems. This paper [12] reviews the practical use of machine learning in cancer classification, highlighting its implementation on medical data to categorize cancer types and predict outcomes. It covers supervised, unsupervised, and reinforcement learning, discussing their pros and cons. The review underscores the potential of machine learning to enhance cancer diagnosis and treatment, offering insights for scholars and practitioners on current and future applications in clinical settings.

Metastatic Breast Cancer (MBC) is a leading cause of cancer-related deaths in women. This study [13] aims to develop a non-invasive system for diagnosing cancer metastases using machine learning (ML) models on blood profile data. Text mining from Electronic Medical Records (EMR) helped identify MBC patients, revealing significant differences in monocyte levels. After removing outliers, a Decision Tree (DT) classifier achieved 83% accuracy and an AUC of 0.87. The DT model was deployed via a web application for robust MBC diagnosis, aiding physicians in improving patient survival outcomes. Melanoma, the deadliest form of skin cancer, is increasing. This study [14] presents a deep learning system for melanoma lesion detection using a GPU-equipped server. A convolutional neural network (CNN) pre-trained on large datasets is used to classify images as malignant or nonmalignant melanoma. The system aims to assist dermatologists in early detection, showing promise in laboratory settings. Experimental results indicate that the proposed technique surpasses state-of-the-art methods in diagnostic accuracy, highlighting its potential in clinical applications. Early detection and accurate diagnosis of breast cancer (BC) are critical for improving patient survival rates. This study [15] proposes a deep learning model (BCCNN) to classify breast cancer MRI images into eight categories, including both benign and malignant types. 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. This review [16] examines the application of machine learning and deep learning techniques in breast cancer classification and diagnosis across five medical imaging modalities: mammography, ultrasound, MRI, histology, and thermography. The study highlights the use of Nearest Neighbor, SVM, Naive Bayesian Network, DT, ANN, and deep learning architectures, including CNNs. Findings indicate that these techniques achieve high accuracy rates and have the potential to enhance clinical decision-making and patient outcomes.

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. This research [17] presents a novel approach for leukemia classification using machine learning and deep learning. The methodology includes dataset building, feature extraction using pre-trained CNN models, and classification with conventional classifiers. The dataset comprises four classes: Benign, Early Pre-B, Pre-B, and Pro-B. By incorporating nature-inspired algorithms like PSO and CSO, the study achieved a maximum accuracy of 99.84% using the ResNet50 CNN architecture and LR classifiers, offering potential improvements in real-world blood cancer classification. Breast cancer (BC) is a leading cause of death among women worldwide, with early detection crucial for reducing mortality rates. This study [18] introduces a big data-based BC classification model using Deep Reinforcement Learning (DRL). The model processes and normalizes data, selects features via the gorilla troops optimization (GTO) algorithm, and employs Deep Q learning (DQL) for classification, with LIME explaining outputs. Evaluated on three UCI datasets (WBCD, WDBC, WPBC), the GTO-DQL model surpasses traditional methods, achieving accuracy rates of 98.90%, 99.02%, and 98.88%, respectively. Breast cancer remains a leading cause of death among women, with approximately 8% diagnosed, second only to lung cancer. Manifesting through genetic changes, pain, and skin alterations, early and accurate diagnosis is crucial. This study [19] utilizes the Extreme Gradient Boosting (XGBoost) machine learning technique to enhance the swift and precise detection of breast cancer. Applied to the Wisconsin breast cancer (diagnostic) dataset, XGBoost achieved an accuracy of 94.74% and a recall of 95.24%, demonstrating its effectiveness in early diagnosis.

Breast cancer (BC) affects over 2.1 million women annually worldwide. Early and accurate diagnosis is crucial for improving survival rates. This review [20] examines the current state of Deep Neural Network (DNN) techniques for BC detection, classification, and segmentation using medical imaging, focusing on mammography and histopathologic images. It highlights the benefits and limitations of various imaging modalities and pre-processing methods like data augmentation, scaling, and normalization. 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. This study [21] aims to develop an efficient breast cancer classification model using meta-learning and multiple convolutional neural networks (CNNs) on the Breast Ultrasound Images (BUSI) dataset, which includes various breast lesions. Traditional approaches often struggle with the dataset's complexity. The proposed model integrates meta-learning for optimized learning adaptation, 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 like accuracy, precision, recall, and F1 score against existing systems.

II. Material and Method

This research explores the utility of the XGBoost Classifier in gene selection for cancer classification, leveraging gene expression profiles' potential 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 presence of redundant and unimportant genes, thereby increasing false positive rates. Employing the XGBoost Classifier addresses this issue by effectively identifying genes that significantly contribute to cancer classification. Through rigorous preprocessing, the study isolates and prioritizes the most informative genes, essential for accurate cancer classification. Subsequently, a search approach is applied to refine this selection, aiming to identify a concise yet highly informative subset of genes that optimize cancer classification accuracy. This methodology not only enhances the precision of cancer diagnostics but also streamlines the gene selection process, providing a robust framework for future research and clinical applications..

In this study, the XGBoost Classifier was utilized to identify genes that are 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 subsequent generations. The selection of the best individuals involved a prefiltering step to narrow down informative genes. The XGBoost Classifier was then employed to optimize the gene subset, retaining only those genes with scores greater than zero, thereby excluding irrelevant ones. For evaluation, Support Vector Machines (SVM) were utilized to assess accuracy, which is particularly suited for high-dimensional, small-sample data classification tasks. The results of the study included accuracy rates and graphical representations of cancer classification outcomes, highlighting the effectiveness of the proposed methodology.

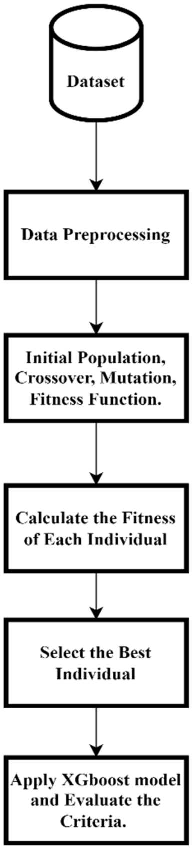

Dataset
Data Preprocessing
Initial Population, Crossover, Mutation, Fitness Function.
Calculate the Fitness of Each Individual
Select the Best Individual
Apply XGboost model and Evaluate the Criteria.

Fig. 1 Methodology of this research

A. Dataset

In this study, the primary dataset utilized is the Breast Cancer Dataset, which consists of 151 columns representing samples and 54,676 rows representing genes. The dataset encompasses six distinct classes: basal, HER, luminal_B, luminal_A, cell_line, and normal. 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.

A. Experiments

This chapter 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 effectiveness is attributed to its ensemble learning approach, which combines multiple weak learners to create a robust and precise predictive model. XGBoost, an abbreviation for Extreme Gradient Boosting, is a powerful and efficient algorithm renowned for its high performance in various machine learning tasks, including cancer prediction. It belongs to the family of gradient boosting algorithms, which sequentially constructs an ensemble of weak learners, typically decision trees, to enhance predictive accuracy.

B. Feature Extraction

The XGBoost classifier is employed for training and testing the dataset and for performing gene selection for cancer classification. Several essential modules are imported for this process, including \`train_test_split\`, \`confusion_matrix\`, \`roc_auc_score\`, \`classification_report\`, \`make_multilabel_classification\`, \`XGBClassifier\`, \`KFold\`, \`MultiOutputClassifier\`, and \`Pipeline\`. 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 training and testing sets, with the training set comprising 70% of the data and the testing set comprising 30%.

C. Evaluation

The evaluation of the proposed XGBoost Classifier in this research employs key metrics such as Accuracy, Precision, Recall, and F1 Score for validation. Accuracy serves as a fundamental metric, quantifying the proportion of correct predictions made by the classifier. It provides a singular measure to assess the overall performance of the model in accurately predicting the classes within the dataset. The equations for Accuracy, Precision, Recall, and F1 Score are provided in Equations (1)-(4), respectively.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{1}$$

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.

$$Precision = \frac{TruePostive}{TruePostive + FalsePostive} \tag{2}$$

Recall measures the proportion of predictions in the Positive class that correctly match the ground truth among all actual Positive samples. It evaluates a classifier's ability to correctly identify positive instances, providing insights into its sensitivity to detecting relevant samples within the dataset.

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

The F1 score represents the harmonic mean of precision and recall, providing a single metric to assess the balance between these two measures. It quantifies the accuracy of positive predictions, where a value of 1.0 indicates optimal performance and 0.0 indicates the lowest.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

III. Result 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 recursive feature elimination with cross-validation (RFECV). The RFECV method contributes to enhancing feature selection robustness and model performance evaluation. The discussion emphasizes the efficacy of the new method in improving predictive accuracy and highlighting its potential impact on cancer prediction models.

TABLE I
PROPOSED RESEARCH AND PREVIOUS RESEARCH RESULT ANALYSIS

RESEARC DURATION	**AREA**	**PRECISION**	**RECALL**	**F1-SCORE**	**SUPPORT**
	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
PREVIOUS RESEARCH	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		
	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
PROPOSED RESEARCH	5	1.00	1.00	1.00	2
	MICRO AVG	0.93	0.87	0.90	61

RESEARC DURATION	AREA	PRECISION	RECALL	F1-SCORE	SUPPORT
	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		

Table 1 presents the outputs from previous research compared with the results introduced in this study. The dataset comprises six classes, labeled from 0 to 5. Recall, also known as sensitivity, measures a classifier's ability to correctly identify all positive instances. It is defined as the proportion of true positives to the sum of true positives and false negatives for each class. In other words, recall represents the fraction of actual positives that were correctly identified by the classifier.

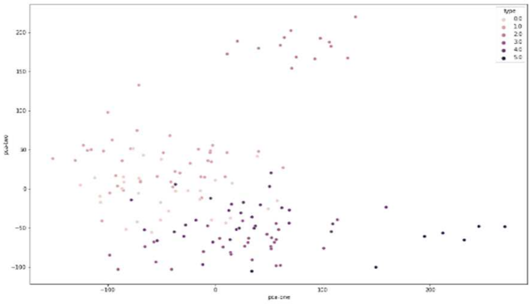


Fig. 2 2D PCA Grpah from Previous Work, Compare between PCA 1 and PCA 2.

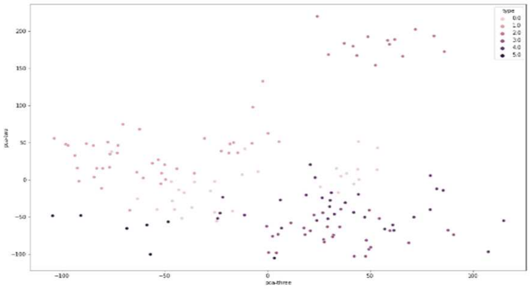


Fig. 3 2D PCA Grpah from Previous Work, Compare between PCA 2 and PCA 3

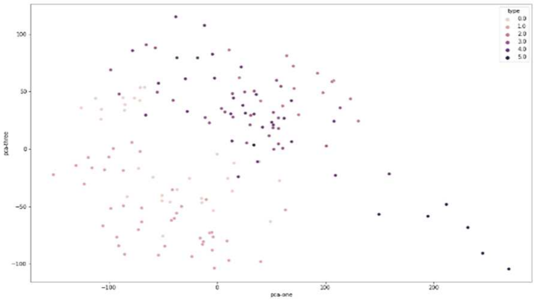


Fig. 4 2D PCA Grpah from Previous Work, Compare between PCA 1 and PCA 3

The F1 Score, a weighted harmonic mean of recall and precision, serves as a comprehensive metric for evaluating model performance. An F1 Score closer to 1.0 indicates superior performance, with 1.0 being the optimal score and 0.0 the worst. Additionally, the macro average of the F1 Score provides an overall performance measure across all classes, where a higher average score is preferable. Support refers to the number of instances of each class in the dataset and is utilized in the performance evaluation process without affecting model comparisons. For instance, a support value of 9 for class 0 indicates that there are 9 observations with actual occurrences of class 0 in the dataset.

A. *Previous Work with XGBoost Classifier*

After the classification report, several graphs are generated based on the results using Principal Component Analysis (PCA). PCA is primarily employed to reduce the number of dimensions or features in a dataset. In this research, a total of 90 components were analyzed in the previous work. For the purpose of result discussion, only the top 3 PCA components are discussed. The graph below illustrates a comparison between PCA 1 and PCA 2 from the previous work, highlighting the dimensionality reduction and its impact on data visualization and interpretation. Below are Figures 2 through 5, which are generated from previous work. Figure 2 illustrates the comparison between PCA 1 and PCA 2. Figure 3 presents the comparison between PCA 2 and PCA 3. Figure 4 shows the comparison between PCA 1 and PCA 3.

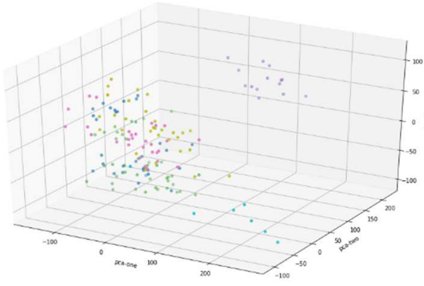


Fig. 5 3D PCA Grpah from Previous Work, Compare between PCA 1, PCA 2 and PCA 3

Finally, Figure 5 depicts a 3D PCA graph, comparing PCA 1, PCA 2, and PCA 3. These visualizations effectively demonstrate the dimensionality reduction achieved through PCA and provide insights into the data structure across different principal components.

The primary distinction between 2D and 3D PCA lies in the number of principal components utilized for visualization. Principal Component Analysis (PCA) constructs principal components to capture the maximum variance within the dataset: PC1 represents the highest variance, followed by PC2, which reflects the second-highest variance, and so forth. Consequently, the top two or three principal components can account for the majority of the variance, allowing for the exclusion of additional components without significant loss of information. While PCA is not inherently a clustering technique, it facilitates the visualization of patterns by reducing dimensionality, which can reveal clusters of expression profiles with similar characteristics. Such patterns may be difficult to discern in a 2D PCA plot but become more apparent in a 3D representation. Upon examining the 3D PCA results, it is evident that the gene expression profiles do not form distinct clusters, making it challenging to identify classes or groups. This suggests that the 3D PCA visualization presents a more complex and scattered arrangement of gene expressions, complicating the identification of outliers that warrant further investigation.

B. *Proposed Method with XGBoost Classifier and Recursive Feature Elimination 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 2 and PCA 3.

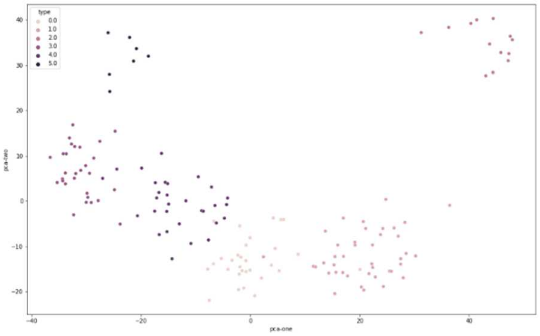


Fig. 6 2D PCA Grpah from Proposed Method, Compare between PCA 1 and PCA 2.

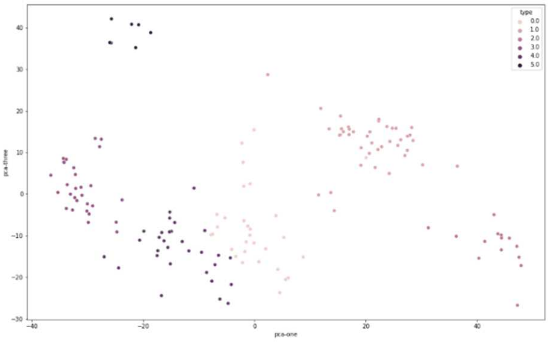


Fig. 7 2D PCA Grpah from Proposed Method, Compare between PCA 2 and PCA 3

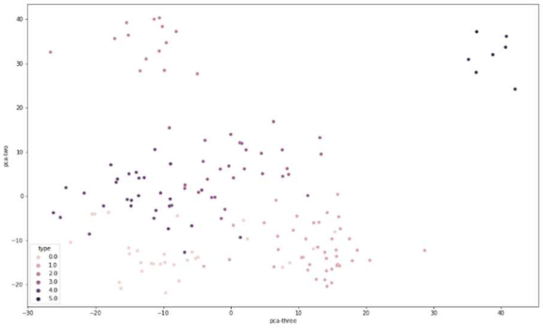


Fig. 8 2D PCA Grpah from Proposed Method, Compare between PCA 1 and PCA 3

Figure 8 presents the comparison between PCA 1 and PCA 3. Observing the PCA plots for the proposed method, it is evident that clusters for each of the classes indicate a higher relationship based on gene expression profiles. Examining the distances between clusters is more effective for identifying outliers than analyzing one variable at a time. It is clearly seen that the graphs produced using the new code exhibit more defined gene clusters compared to the original code. Genes based on classes with similar expression profiles are now clustered together, while those that do not cluster are identified as outliers.

For instance, the plot comparing PCA 1 and PCA 2 clearly demonstrates that the clusters produced by the proposed method are more discernible. This indicates that genes based on classes in the proposed method are grouped together more effectively, with fewer genes being incorrectly classified into other classes. Figure 9 represents the 3D PCA plot produced by the proposed method.

Upon observation, the 3D PCA plot from the proposed method clearly shows more identifiable gene clusters based on the classes. The plot indicates that the proposed method achieves higher accuracy compared to previous work, as evidenced 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.

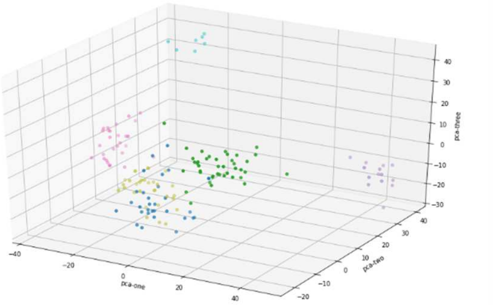


Fig. 9 3D PCA Grpah from Proposed Method, Compare between PCA 1, PCA 2 and PCA 3

In conclusion, this section has focused on the explanation of the results and findings derived from the implementation of 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 of the XGBoost classifier in gene selection for cancer classification.

IV. Conclusion

In this research, a novel approach for gene selection in cancer classification utilizing the XGBoost Classifier has been proposed. Microarray technology facilitates the creation of comprehensive databases of cancerous tissues based on gene expression data. However, training datasets for cancer classification often have a small sample size and consist of multiclass categories compared to the number of genes involved. In this study, a breast cancer dataset was used. The most critical genes were selected using recursive feature elimination with cross-validation (RFECV), followed by cancer classification using the XGBoost classifier. The results demonstrate that combining RFECV with XGBoost yields higher accuracy in gene selection and cancer classification compared to using the XGBoost classifier alone.

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

The authors would like to 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 Internal Research grant RDU2303103.

References

[1] N. Mahendran, P. M. Durai Raj 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.

[2] 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.

[3] 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.

[4] L. Tapak, M. K. Ghasemi, S. Afshar, H. Mahjub, A. Soltanian, and H. Khotanlou, “Identification of gene profiles related to the development of oral cancer using a deep learning technique,” BMC Med. Genomics, vol. 16, no. 1, p. 35, Feb. 2023, doi: 10.1186/s12920-023-01462-6.

[5] 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.

[6] M. Akhavan and S. M. H. Hasheminejad, “A two-phase gene selection method using anomaly detection and genetic algorithm for microarray data,” Knowledge-Based Syst., vol. 262, p. 110249, Feb. 2023, doi:10.1016/j.knosys.2022.110249.

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

[8] S. Kayikci and T. M. Khoshgoftaar, “Breast cancer prediction using gated attentive multimodal deep learning,” J. Big Data, vol. 10, no. 1, p. 62, May 2023, doi: 10.1186/s40537-023-00749-w.

[9] 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, p. 10, Jan. 2024, doi:10.1007/s10916-023-02031-1.

[10] 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, p. 110034, Mar. 2023, doi:10.1016/j.asoc.2023.110034.

[11] F. Alharbi and A. Vakanski, “Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review,” Bioengineering, vol. 10, no. 2, p. 173, Jan. 2023, doi:10.3390/bioengineering10020173.

[12] A. Yaqoob, R. Musheer Aziz, and N. K. Verma, “Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review,” Human-Centric Intell. Syst., vol. 3, no. 4, pp. 588–615, Sep. 2023, doi: 10.1007/s44230-023-00041-3.

[13] 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, p. 485, Jan. 2023, doi: 10.1038/s41598-023-27548-w.

[14] S. R. Waheed et al., “Melanoma Skin Cancer Classification based on CNN Deep Learning Algorithms,” Malaysian J. Fundam. Appl. Sci., vol. 19, no. 3, pp. 299–305, May 2023, doi:10.11113/mjfas.v19n3.2900.

[15] B. Abunasser, M. R. AL-Hiealy, I. Zaqout, and S. Abu-Naser, “Convolution Neural Network for Breast Cancer Detection and Classification Using Deep Learning,” Asian Pacific J. Cancer Prev., vol. 24, no. 2, pp. 531–544, Feb. 2023, doi:10.31557/APJCP.2023.24.2.531.

[16] 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.

[17] W. Rahman, M. G. G. Faruque, K. Roksana, A. H. M. S. Sadi, M. M. Rahman, and M. M. Azad, “Multiclass blood cancer classification using deep CNN with optimized features,” Array, vol. 18, p. 100292, Jul. 2023, doi: 10.1016/j.array.2023.100292.

[18] S. Almutairi, M. S., B.-G. Kim, M. M. Aborokbah, and N. C., “Breast cancer classification using Deep Q Learning (DQL) and gorilla troops optimization (GTO),” Appl. Soft Comput., vol. 142, p. 110292, Jul. 2023, doi: 10.1016/j.asoc.2023.110292.

[19] Rahmanul Hoque, Suman Das, Mahmudul Hoque, and Mahmudul 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.

[20] 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.

[21] M. D. Ali *et al.*, “Breast Cancer Classification through Meta-Learning Ensemble Technique Using Convolution Neural Networks,” *Diagnostics*, vol. 13, no. 13, p. 2242, Jun. 2023, doi:10.3390/diagnostics13132242.