ISSN 2007-9737

Improving Statistical Learning Methods Via Features Selection without Replacement Sampling and Random Projection

Sulaiman Khan, Muhammad Ahmad, Fida Ullah, Carlos Fernando Aguilar Ibañez*, José Eduardo Valdez Rodriguez

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

{skhan2024; mahmad2024, fullah-2022, carlosaguilari, jvaldezr2018}@cic.ipn.mx

Abstract. Cancer is fundamentally a genetic disease characterized by genetic and epigenetic alterations that disrupt normal gene expression, leading to uncontrolled cell growth and metastasis. High-dimensional microarray datasets pose challenges for classification models due to the "small n, large p" problem, which can lead to overfitting. This study makes three different key contributions: 1) We propose a machine learning-based approach integrating the Feature Selection Without Replacement (FSWOR) technique and a projection method to improve classification accuracy. 2) We apply the Kendall statistical test to identify the most significant genes from the brain cancer microarray dataset (GSE50161), reducing the feature space from 54,675 to 20,890 genes. 3) We apply machine learning models using k-fold cross-validation techniques in which our model incorporates ensemble classifiers with LDA projection and Naïve Bayes, achieving a test score of 96%, outperforming existing methods by 9.09%. The results demonstrate the effectiveness of our approach in high-dimensional gene expression analysis, improving classification accuracy while mitigating overfitting. This study contributes to cancer biomarker discovery, offering a robust computational method for analyzing microarray data.

Keywords. Brain cancer, gene expression, machine learning, SVM, NB, LR, DT, KNN, dimension reduction, PCA, LDA, GRP, SRP.

1 Introduction

Cancer is fundamentally a genetic disease resulting from genetic or epigenetic alterations within somatic cells. These alterations disrupt normal gene expression patterns [1,2,3], affecting genes that regulate cell growth, survival, and invasion while suppressing growth-inhibiting genes. The primary mechanism involves the accumulation of mutations; however, epigenetic changes, such as DNA methylation, also play a crucial role. The resulting aberrant gene expression leads to the hallmarks and enabling characteristics of cancer [4]. Most common cancers arise from acquired mutations in somatic cells, whereas rare hereditary cancer syndromes result from specific germline mutations.

Cancer-associated genes are categorized into oncogenes (activated, phenotypically dominant) and tumor suppressor genes (inactivated, phenotypically recessive). Oncogene activation can occur through point mutations, gene amplification, or DNA translocation, whereas tumor suppressor genes are inactivated by mutations or promoter silencing [5]. The uncontrolled growth and spread of these abnormal cells define the disease [6]. Cancer is a subset of neoplasms [7], characterized by unregulated cell growth that forms a mass or tumor, potentially spreading diffusely.

Metastasis, the spread of cancer [8], is a multi-step process. It begins with the detachment of tumor cells from neighboring cells and the surrounding stroma at the primary site. Enzymatic degradation of the extracellular matrix facilitates the directional movement of individual cells or cell clusters. These cells then invade blood or lymphatic vessels (intravasation) and travel through the circulatory system [9]. Survival during circulation is crucial until they reach a suitable metastatic site, which is often determined by the availability of growth factors. At the metastatic site,

cells attach to blood vessel endothelium, exit the vessel (extravasation), proliferate, invade, and recruit a new blood supply.

Epithelial-to-mesenchymal transition (EMT) is a key process in the invasion and metastasis of epithelial tumors [10], often followed by a mesenchymal-to-epithelial transition (MET) at the metastatic site [11]. Metastasis patterns to specific organs are not random but are influenced by the expression of chemokine receptors on tumor cells, guiding them to favorable environments for colony establishment [12]. In short, cancer cells escape their origin and establish new colonies in distant parts of the body. Distant metastases account for 90% of cancer deaths [13].

The spread can occur through the lymphatic system or bloodstream, leading to the development of new tumors in areas such as the lymph nodes (neck, underarms, groin) or distant organs such as liver, bones, brain, lungs [14,28,31]. If cancer spreads from its origin, it is termed metastatic cancer of the original site, not the site of spread [15].

A microarray is a molecular biological method in which tens of thousands of probes, each demonstrating a specific DNA sequence, are examined and enumerated to provide a comprehensive gene expression profile of multiple biological samples [16]. From a computational viewpoint, single-cell RNA sequencing studies generate a large amount of data, encompassing several cells and thousands of gene dimensions [17]. Most single-cell RNA sequence data so far belong to the small n large p category, where n is the number of samples and p is the number of dimensions (Genes) [18,19,20]. This violates the Gaussian Markov assumption, i.e, $n > p$. Classification models' effects from overfitting due to large features and a smaller number of samples.

This study makes the following contributions:

- We apply the Kendall statistical test to identify the most significant genes from the brain cancer microarray dataset (GSE50161), reducing the feature space from 54,675 to 20,890 genes.
- We propose, implement and evaluated a machine learning-based approach that integrates the Feature Selection without Replacement (FSWOR) technique and a projection method to enhance classification accuracy.
- We apply machine learning models using k-fold cross-validation techniques, where our model incorporates ensemble classifiers with LDA projection and Naïve Bayes, achieving a test score of 96%. This outperforms existing methods by 9.09% (Bruno et al.).

2 Literature Review

Zhang et al. [21] proposed a random projection enhancement (RPE) method to improve surrogate model performance. They applied RPE to least squares support vector regression (LSSVR) and conducted numerical experiments. The results showed improved predictive accuracy, robustness, and optimization performance, even for high-dimensional problems. They further validated RPE's effectiveness on other models and real-world engineering applications.

Hu et al. [22] proposed a hybrid pre-computation-based Heterogeneous Graph Neural Network (RpHGNN) that balances efficiency and low information loss. By using this, they introduced a Random Projection Squashing step for linear complexity and a Relation-wise Neighbor Collection component for finer-grained information aggregation. Their experimental results demonstrated state-of-the-art performance on seven benchmark datasets, achieving a 230% speedup over the best baseline. Their method outperformed both pre-processing-based and end-to-end models.

Fabiani et al. [23] investigate Best Approximation for Feedforward Neural Networks (FNNs) using Random Projection Neural Networks (RPNNs). They show that RPNNs, with fixed internal weights and non-polynomial activation functions, achieve exponential convergence in function approximation. Five benchmark tests demonstrate their effectiveness, achieving performance comparable to Legendre Polynomials. This highlights the potential of RPNNs for efficient and accurate function approximation.

Li et al. [24] propose a probabilistic framework for sequential random projection, addressing challenges in sequential decision-making under uncertainty. They construct a stopped process to analyze sequential concentration events and derive a non-asymptotic probability bound. This extends the Johnson-Lindenstrauss lemma to a martingale setting, contributing to random projection and sequential analysis.

Asi et al. [25] propose the Projection Unit, a framework for efficient and private mean estimation by projecting inputs into random low-dimensional subspaces. This method achieves near-optimal error with reduced communication and computational costs. Experiments show it performs similarly to optimal algorithms in private mean estimation and federated learning.

McDonnell et al. [26] propose a novel approach for continual learning (CL) with pre-trained models, addressing catastrophic forgetting by using frozen Random Projection layers and class-prototype accumulation. This method enhances linear separability and decorrelates class-prototypes to reduce distribution gaps. Experiments show that their approach reduces error rates by 20%-62% on class-incremental datasets without using rehearsal memory.

Kumaran et al. [27] explore the use of local random quantum circuits for dimensionality reduction of large low-rank datasets, leveraging the random projection method. They demonstrate that quantum circuits with short depths perform comparably to classical principal component analysis on image datasets such as MNIST and CIFAR-100. Benchmarking quantum random projection against classical methods, they demonstrate its effectiveness in reducing dimensions and computing von Neumann entropies, as well as implementing singular value decomposition for large matrices.

Ahmed et al. [30] explored the machine learning models using Support Vector machine to perform the diagnosis of breast cáncer which is the leading cause of death using the WBCD dataset. They employed different kernels enhanced with many paramteres using the holdout method and evaluated accuracy, sensitivity, specificity, and predictive values. Their proposed model shows the results that the Cauchy and Rational Quadratic kernels behave similarly and converge to the same performance.

Table 1. Prior studies on brain cancer detection

Ref.	Method	Techniques	C.V score
[29]	**SVM**, NB, RF, DT, MLP	**PCA**	0.88
proposed	DT, **NB**, SVM, LR, KNN	FSWOR(PCA), **FSWOR(LDA)**, FSWOR(GRP), FSWOR(SRP)	0.96

Table 1 compare previous and proposed studies on brain cancer detection using supervised learning methods. Feltes et al applied classifiers like SVM, NB, RF, DT, and MLP with PCA, achieving a 0.88 cross-validation score. The proposed study utilized DT, NB, SVM, LR and KNN with advanced FSWOR-based techniques (PCA, LDA, GRP, SRP). These hybrid approaches improved the performance of classifiers. As a result, the proposed method achieved a higher C.V score 0.96, showing superior classification performance.

3 Methodology

This section outlines the methodological framework adopted to conduct the present study. The approach was carefully designed to ensure reliability, validity, and reproducibility of results. In this study, we utilized a publicly available brain cancer dataset to perform effective classification using high-dimensional gene expression data. To address the inherent challenges of microarray datasets—such as extremely large feature spaces and limited sample sizes—we designed a multi-stage machine learning pipeline tailored to address the challenges posed by microarray datasets, such as high feature dimensionality and a limited number of samples. Our methodology begins with statistical filtering using the Kendall rank correlation test to identify the most informative genes, thereby reducing the dimensionality and improving model interpretability. This is followed by the application of a novel Feature Selection without Replacement (FSWOR) approach in combination with projection techniques to further refine the input space. Finally, we evaluate the performance of

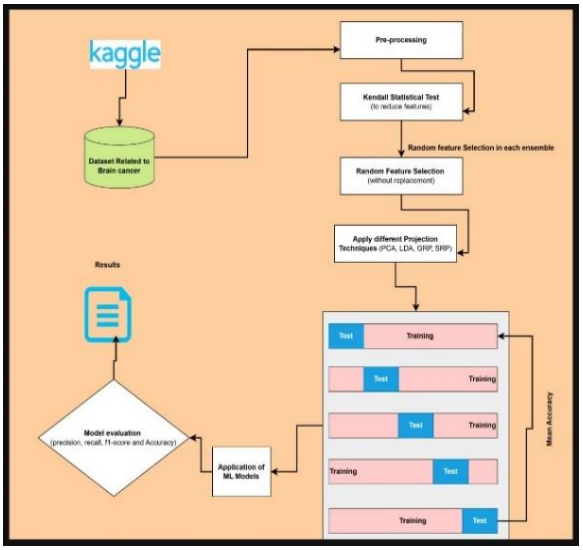

Fig. 1. Proposed architecture and design

several classifiers, including an ensemble model combining Linear Discriminant Analysis (LDA) and Naïve Bayes, using k-fold cross-validation to ensure robustness and generalizability. The following subsections detail each component of our proposed framework. Figure 1 shows the overall architecture of our proposed methodology.

3.1 Brain Cancer Dataset

We sourced our dataset from Kaggle[1], which contains gene expression data for brain cancer. The dataset comprises a total of 130 samples, with 13 classes representing normal tissues and 117 classes representing abnormal tissues.

In abnormal cases, there are four different types of cancer: Ependymoma, Glioblastoma, Medulloblastoma, and Pilocytic Astrocytoma. The corpus is in a well-structured format, which contains 54,675 independent gene expression (features). This extensive data allows for in-depth analysis of different brain cancer types. We apply the Kendall statistical test to this dataset. We obtained 20,890 significant genes. We apply the normalization scaling technique after the Kendall statistical test.

[1] https://www.kaggle.com/datasets/brunogrisci/brain-cancer-gene-expression-umida

3.2 Kendall Statistical Test

In this section, we discuss the application of the Kendall rank correlation test on our dataset to perform initial feature selection. The GSE50161 brain cancer microarray dataset contains expression values for 54,675 genes across multiple samples. Given the high dimensionality of the data and the relatively small number of samples, applying a statistical method to reduce the feature space was essential to improve model performance and reduce the risk of overfitting. The Kendall test, a non-parametric measure of correlation based on the ranks of the data rather than their actual values, was chosen due to its robustness in handling non-Gaussian distributions and its suitability for detecting monotonic relationships.

We calculated the Kendall tau correlation coefficient between each gene's expression values and the corresponding class labels, which represent different types of brain cancer. Genes with higher absolute correlation values were considered more relevant for distinguishing between cancer types. Based on this analysis, we selected the top-ranked genes that showed statistically significant correlations with the class labels. As a result, we reduced the feature space from 54,675 to 20,890 genes. This step was critical for removing irrelevant or weakly associated genes, thereby simplifying the dataset and improving the efficiency and accuracy of downstream machine learning models.

3.3 Randomly Features Selection

The figure 1 outlines a structured methodology for developing machine learning models to analyze brain cancer data. The workflow begins with pre-processing, where raw data is refined through techniques such as normalization and handling of missing values to ensure robustness.

Feature selection is rigorously addressed via the Kendall Statistical Test, which identifies statistically significant biomarkers, and random feature selection (without replacement) to enhance

model diversity in each ensemble and reduce dimensionality.

Subsequently, projection techniques are applied to distill high-dimensional data into interpretable subspaces. These include:

- Principal Component Analysis (PCA) for variance maximization,
- Linear Discriminant Analysis (LDA) to optimize class separability,
- Gaussian Random Projection (GRP) and Sparse Random Projection (SRP) to preserve pairwise distances while improving computational efficiency.

The machine learning models such as Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bays (NB) and K Nearest Neighbour (KNN) are trained, and validated using a k-fold cross-validation framework, ensuring rigorous assessment of generalizability as shown in figure 1. Ensemble methods are employed, where individual learners are trained on distinct feature subsets to mitigate overfitting and enhance predictive stability.

Performance metrics—precision, recall, F1-score, and accuracy—are systematically evaluated to quantify diagnostic efficacy. Precision reflects the model's ability to minimize false positives, while recall captures sensitivity in detecting true cancer cases. The F1-score harmonizes these metrics, and accuracy provides an aggregate measure of correctness.

This pipeline integrates statistical rigor, algorithmic diversity, and iterative validation, aiming to advance computational tools for brain cancer diagnosis. By bridging data-driven insights with clinical relevance, the framework underscores the potential of machine learning to augment oncological decision-making, ultimately contributing to precision medicine initiatives.

3.4 Applied Projection

Following the initial reduction of the feature space using the Kendall statistical test, we applied a projection technique to further enhance the discriminative capacity of the selected gene expression features. Projection methods aim to transform high-dimensional data into a lower-dimensional subspace while preserving class separability and the intrinsic structure of the data. In our study, we employed Linear Discriminant Analysis (LDA) as the projection technique due to its effectiveness in supervised settings, particularly when the goal is to maximize between-class variance and minimize within-class variance.

LDA was applied to the subset of 20,890 genes obtained after Kendall-based filtering. By projecting the data onto a new axis defined by the directions of maximum class separation, LDA transformed the high-dimensional gene space into a compact, informative representation that facilitated more effective classification. This transformation not only reduced computational complexity but also helped to improve the generalization ability of the subsequent classifiers by emphasizing the features most relevant for distinguishing between brain cancer subtypes. The combination of Kendall filtering followed by LDA projection created a streamlined and interpretable feature space, setting the stage for accurate and robust classification.

3.5 Application of Machine Learning, Training and Testing

After dimensionality reduction and projection, the refined dataset was used to train and evaluate several machine learning models for brain cancer classification. To ensure robust performance and mitigate the risk of overfitting, we implemented a stratified k-fold cross-validation strategy, where the dataset was divided into k equal parts, ensuring each fold preserved the original class distribution. During each iteration, one fold was held out as the test set while the remaining folds were used for training, and this process was repeated k times so that each sample served as a test instance exactly once. The final performance was calculated as the average of the scores across all folds.

We experimented with various classifiers, including NB, SVM, and RF. However, the best results were achieved with an ensemble model that integrated Linear Discriminant Analysis (LDA) for feature projection and Naïve Bayes for classification. The ensemble framework benefited from LDA’s dimensionality reduction and Naïve Bayes' probabilistic modeling, resulting in high classification accuracy. Our proposed model

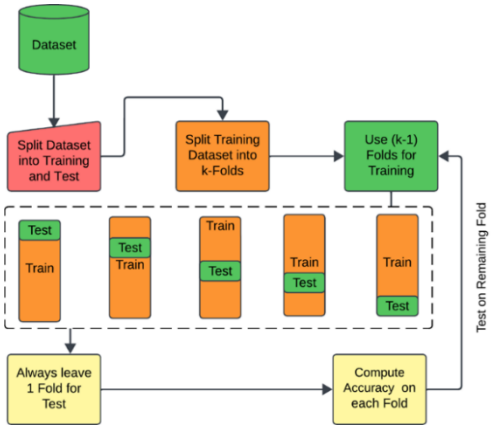

Fig. 2. Application of models training and testing

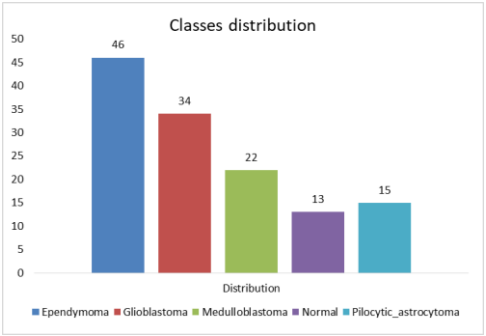

Fig. 3. Class distribution of brain cancer dataset

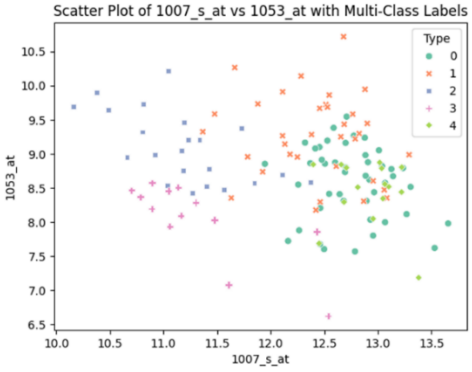

Fig. 4. Scatter plot before Kendall Statistical Test

achieved a test accuracy of 96%, outperforming previous benchmarks such as Bruno et al. by 9.09%.

To ensure fairness and reproducibility, all models were trained and tested under identical conditions using the same data splits and evaluation metrics. Accuracy, precision, recall, and F1-score were computed for comprehensive assessment. This pipeline demonstrated that combining statistical filtering, supervised projection, and ensemble classification could effectively tackle the challenges posed by high-dimensional microarray datasets in cancer genomics. Figure 2 shows the overall process of application of models, training and testing phase.

4 Results and Analysis

This section presents the experimental results and a detailed analysis of the performance of our proposed machine learning framework for brain cancer classification. After applying the Kendall statistical test for initial feature selection and LDA for projection, the resulting dataset was used to train various classifiers. The effectiveness of each model was evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score. We also compared our results with existing approaches in the literature to highlight the improvements achieved through our method. The analysis provides insights into how each component of the pipeline—feature selection, projection, and classification—contributed to the overall performance and generalization capability of the model.

4.1 Statistical Analysis

We have mentioned the class distribution and scatterplot before the Kendall statistical test, and after the Kendall statistical test, we obtained the most significant features, i.e., 20890. These features are used in Machine learning models with a total sample size of 130.

Figure 3 illustrates the various classes of Brain cancer types. Ependymoma, Glioblastoma, Medulloblastoma, Normal Astrocytoma, and Pilocytic astrocytoma are classified into 46, 34, 22, 13, and 15 classes, respectively. 0, 1, 2, 3, and 4 denote Ependymoma, Glioblastoma, Medulloblastoma, Normal, and Pilocytic astrocytoma classes, respectively. But Normal class is healthy tissue.

Figure 4 visualizes the relationship between two variables, 1007_s_at and 1053_at, categorized into multiple classes (labeled 0 to 4).

ISSN 2007-9737

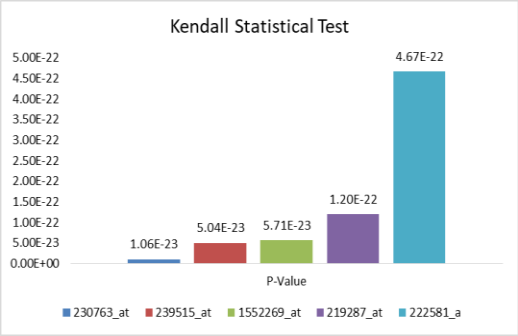

Fig. 5. Kendall statistical test for brain cancer dataset

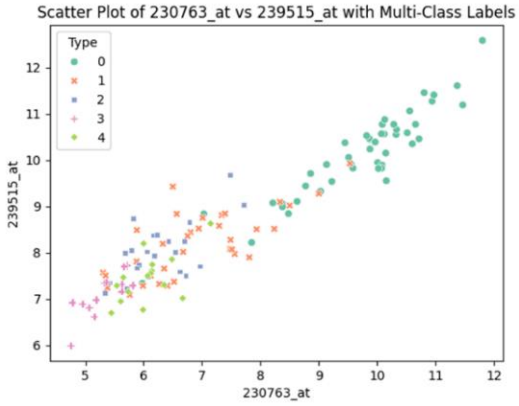

Fig. 6. Scatter plot after Kendall Statistical Test

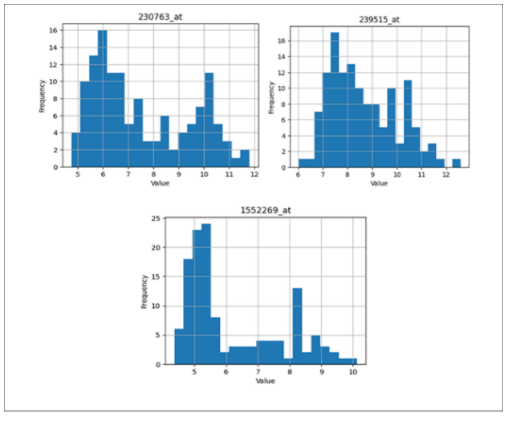

Fig. 7. Histogram of genes expression levels

The plot helps in understanding how these classes (Types) are not linearly separable into different groups in the data before the Kendall statistical test. In our dataset, classes were not linearly separable in different assemblies. To handle this task, we employed the Kendall statistical test and reduced the feature size to make classes linearly separable.

Figure 5 represents the results of a Kendall Statistical Test, which is a non-parametric test used to assess the association between two measured quantities. The y-axis likely represents the p-values obtained from the test, with values ranging from 5.00E-22 (a minimal number indicating a highly significant result) down to 0.00E+00. The x-axis lists different identifiers, possibly representing genes or probes, such as "230763_at" and "239515_at," among others.

The p-values are extremely low, suggesting that the associations being tested are highly statistically significant. For example, the p-value of 1.06E-22 is astronomically small, indicating that the likelihood of the observed association occurring by chance is virtually zero. The identifiers on the x-axis correspond to specific data points or samples being analyzed, and the p-values associated with each suggest that a strong, statistically significant relationship is being measured.

In simpler terms, this figure indicates that the analyzed data provide strong evidence of a relationship, suggesting that the results are not due to random chance. This could be crucial in fields like genetics or bioinformatics, where identifying significant associations can lead to important discoveries about gene functions or disease mechanisms.

Figure 6 visualizes the relationship between two variables, 230763_at and 239515_at, categorized into multiple classes (labeled 0 to 4). The plot helps in understanding how these classes (Types) interact across features, which can aid in identifying patterns or clusters in the data after applying the Kendall statistical test.

Figure 7 demonstrates the histogram of genes, i.e., 230763_at, 239515_at, and 1552269_at. The 230763_at expression level is very high between 5 and 7, and moderate between 9 and 11. The 239515_at expression level peaks between 7 and 9. The 1552269_at expression level is very low between 6 and 8, and between 9 and 10. The specific and peak range indicate which genes are more active (higher expression) or deactivated (lower expression) across samples.

4.2 Machine Learning Analysis

This section shows the experimental results that were made using the multi-stage machine learning pipeline on the brain cancer dataset. The results of various classifiers are tabulated and compared to understand the efficacy of the developed framework to deal with high dimensional gene expressions. The results are discussed in the context of cross validation score, precision, recall and F1-score among other pertinent evaluation measures. As a supplement, comparative discussions are also presented to point out the strengths and shortcomings of the models applied in the solution to the challenges presented by microarray datasets.

4.2.1 Classifier Performance

Table 2 presents the performance metrics of five models: Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB) after applying Principal Component Analysis (PCA) for dimensionality reduction. Among the models, SVM stands out with the highest average precision (0.900), recall (0.884), F1-score (0.887), and test accuracy (0.915), showcasing its superior ability to classify data accurately. KNN and Naive Bayes also perform well, achieving balanced metrics and strong accuracy scores (0.896 and 0.892, respectively). Logistic Regression and Decision Tree, while slightly lower in performance, still demonstrate reliable outcomes with precision, recall, F1-scores, and test accuracy values that reflect solid classification capabilities. Overall, the table highlights SVM as the most effective model, followed by KNN and NB, in this PCA-optimized classification task.

Table 3 summarizes the performance of five classification models: Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB) after applying Linear Discriminant Analysis (LDA) for feature extraction. Naive Bayes (NB) achieves the best results overall, with the highest average precision (0.938), recall (0.916), F1-score (0.923), and accuracy (0.938). KNN also performs exceptionally well, closely following NB in precision (0.934), recall (0.900), F1-score (0.905), and accuracy (0.927). SVM and Logistic Regression also show robust performance, with SVM attaining an accuracy of 0.919, slightly outperforming LR. The Decision Tree, while effective, lags behind the other models, particularly in terms of F1-score (0.845) and accuracy (0.873). Overall, the results emphasize that NB and KNN benefit the most from LDA, offering the most consistent and accurate classifications.

Table 4 highlights the performance metrics of five models: Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB) after using Gaussian Random Projection (GRP) for feature transformation. SVM emerges as the top performer, achieving the highest average precision (0.893), recall (0.844), F1-score (0.853), and accuracy (0.873), making it the most reliable model

Table 2. Machine learning Results with PCA

Models	Precision	Recall	F1-score	Accuracy
DT	0.837	0.834	0.827	0.862
SVM	0.900	0.884	0.887	0.915
LR	0.856	0.838	0.840	0.885
KNN	0.873	0.873	0.871	0.896
NB	0.868	0.859	0.861	0.892

Table 3. Machine learning results with LDA

Models	Precision	Recall	F1-score	Accuracy
DT	0.862	0.857	0.845	0.873
SVM	0.923	0.898	0.901	0.919
LR	0.910	0.877	0.881	0.896
KNN	0.934	0.900	0.905	0.927
NB	0.938	0.916	0.923	0.938

Table 4. Machine learning results with GRP

Models	Precision	Recall	F1-score	Accuracy
DT	0.636	0.600	0.592	0.658
SVM	0.893	0.844	0.853	0.873
LR	0.821	0.813	0.806	0.838
KNN	0.789	0.788	0.778	0.800
NB	0.809	0.776	0.772	0.815

in this context. Logistic Regression and Naive Bayes follow with balanced metrics, delivering accuracy scores of 0.838 and 0.815, respectively. KNN shows moderate performance, with a precision of 0.789 and an accuracy of 0.800. In contrast, the Decision Tree performs poorly compared to the other models, with the lowest precision (0.636), recall (0.600), F1-score (0.592), and accuracy (0.658). These results suggest that SVM is the most effective model under GRP transformation, while DT struggles to adapt effectively.

Table 5 presents the performance of five models: Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Naive Bayes (NB) after applying Sparse Random Projection (SRP) for feature transformation. SVM achieves the highest average performance across all metrics, with precision (0.915), recall (0.855), F1-score (0.866), and accuracy (0.885), making it the standout model in this setup. Naive Bayes (NB) and Logistic Regression (LR) yield similar and balanced results, with NB slightly outperforming in accuracy (0.835) and LR demonstrating consistent performance across all metrics. KNN demonstrates moderate results, maintaining acceptable scores but falling behind the top-performing models. Decision Tree (DT), however, exhibits the weakest performance, with the lowest precision (0.576), recall (0.567), F1-score (0.559), and accuracy (0.619). Overall, the table underscores the robustness of SVM and highlights the difficulty of DT in adapting to the SRP transformation.

Table 6 highlights the top-performing models for each feature transformation technique: PCA, LDA, GRP, and SRP. Across these methods, the Support Vector Machine (SVM) consistently demonstrates superior performance, achieving the highest average scores in three techniques. Under PCA, SVM achieves a precision of 0.900, a recall of 0.884, an F1-score of 0.887, and an accuracy of 0.915. For LDA, Naive Bayes (NB) emerges as the leader with outstanding results, including a precision of 0.938, a recall of 0.916, an F1-score of 0.923, and an accuracy of 0.938. In both GRP and SRP, SVM once again excels, showcasing its adaptability and reliability across different transformations. This table highlights the

Table 5. Machine learning results with SRP

Models	Precision	Recall	F1-score	Accuracy
DT	0.576	0.567	0.559	0.619
SVM	0.915	0.855	0.866	0.885
LR	0.831	0.804	0.799	0.827
KNN	0.799	0.773	0.766	0.792
NB	0.849	0.802	0.807	0.835

Table 6. Top performing models in each Projection

Models	Precision	Recall	F1-score	Accuracy
		PCA		
SVM	0.900	0.884	0.887	0.915
		LDA		
NB	0.938	0.916	0.923	0.938
		GRP		
SVM	0.893	0.844	0.853	0.873
		SRP		
SVM	0.915	0.855	0.866	0.885

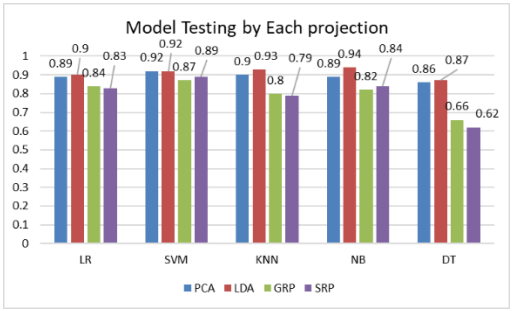

Fig. 8. Comparisons of different machine learning models using different projection techniques

robustness and versatility of SVM, while also acknowledging the strong performance of NB under LDA.

Figure 8 illustrates the testing performance of various models, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes (NB), and Decision Tree (DT), using different projection techniques: PCA, LDA, GRP, and SRP.

Naive Bayes (NB) with LDA achieves the highest score (0.94), highlighting its effectiveness

with this technique. SVM also performs consistently well, particularly with LDA and PCA, achieving top scores of 0.92 for each. KNN and Logistic Regression deliver stable results across the projection methods, with scores hovering near 0.90 under PCA and LDA but slightly lower with GRP and SRP. The Decision Tree lags significantly behind, with its highest score of 0.87 (under PCA and LDA), and its weakest performance under GRP (0.66) and SRP (0.62).

Overall, the figure highlights the strength of SVM and NB across different projections, as well as the notable gap in performance for DT under GRP and SRP.

Figure 9 illustrates the accuracy of an SVM model, but using Principal Component Analysis (PCA) for dimensionality reduction. The accuracy trends suggest that PCA might be more stable or practical compared to Sparse Random Projection, as indicated by the higher accuracy values.

Figure 10 represents the accuracy of a Naive Bayes model using Linear Discriminant Analysis (LDA) for dimensionality reduction. The accuracy trends presented here demonstrate how the model performs as the number of projections changes, providing insight into the effectiveness of LDA in this context.

Figure 11 depicts the accuracy of an SVM model using Gaussian Random Projection. The graph helps in understanding how this particular projection method affects the model's accuracy across different numbers of projections.

Figure 12 shows the accuracy of a Support Vector Machine (SVM) model using Sparse Random Projection for dimensionality reduction. The accuracy is plotted against the number of projections. As the number of projections increases, the accuracy fluctuates, indicating that the optimal number of projections is crucial for achieving the best model performance.

Figure 13 shows the mean confusion matrix for four different machine learning models such as SVM with PCA technique, Naïve Bayes with LDA, SVM with GRP, and SVM with SRP. Each figure shows the actual versus predicted values across five classes, highlighting both accurate classifications and misclassifications.

SVM using PCA technique achieved nearly perfect classification for classes 0, 2, and 3. For class 1, most instances were correctly identified,

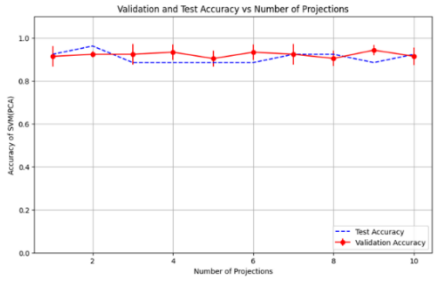

Fig. 9. Validation and Test accuracy comparison vs number of projections

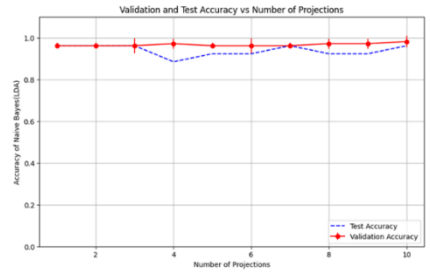

Fig. 10. Validation and Test accuracy comparison vs number of projections

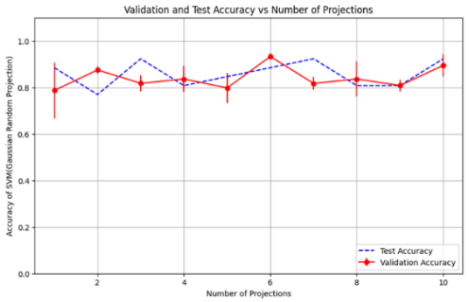

Fig. 11. Validation and Test accuracy comparison vs number of projections

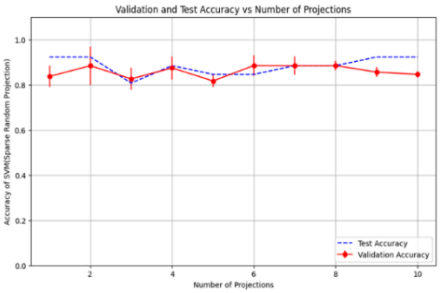

Fig. 12. Validation and Test accuracy comparison vs number of projections

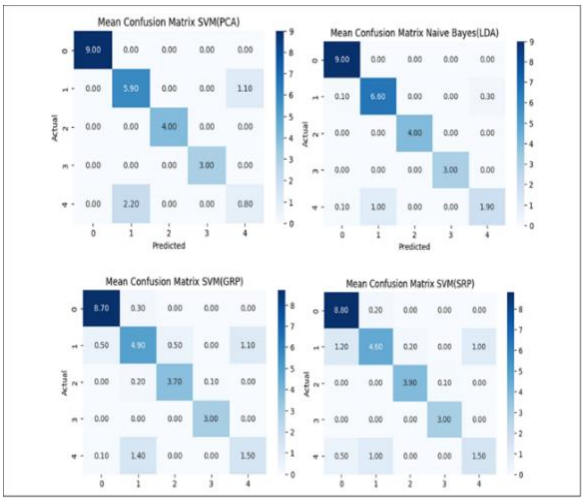

Fig. 13. Confusion matrix for brain cancer classification

though some were misclassified as class 4. Similarly, while class 4 was mostly predicted correctly, a portion of its instances were confused with class 1. This suggests PCA was effective at separating certain classes, but some overlap remained between classes 1 and 4.

Naïve Bayes using LDA technique also performed well for classes 0, 2, and 3, with high correct accuracy rates. While the Class 1 was also predicted correctly, but there is a noticeable number of its samples were also misclassified same as class 4. Conversely, class 4 achieved moderate accuracy but suffered from confusion with class 1. This shows that NB model handled linearly separable classes well, but struggled when class boundaries overlapped.

SVM using the GRP technique provided balanced performance across all four classes. While class 0 achieved high accuracy, classes 1, 2, 3, and 4 showed moderate levels of misclassification. Notably, confusion occurred between class 1 and classes 2 and 4, as well as between class 2 and class 3. This suggests GRP preserved useful global structure but introduced overlaps that reduced separability between neighboring classes.

SVM based on SRP technique reported good result on class 0 and reasonable results on class 1, 2, 3, and 4. Nevertheless, certain misclassifications were observed, the most common between class 1 and 4, between class 2 and 3. Although it was still competitive in accuracy across all classes, despite this, SRP generally showed adequate discriminatory features to survive projection.

In general, PCA and LDA proved superior at providing distinct groupings amongst some classes, resulting in an almost perfect classification of classes 0, 2 and 3. But the two techniques were not able to handle class 1 and class 4 due to overlap. GRP and SRP gave more evenly distributed predictions of all classes but at the expense of more cross-classification errors. These findings demonstrate the trade-off bias in favor of certain classes (PCA, LDA) versus the preference to be more uniform in the case of all classes (GRP, SRP).

5 Conclusion

In this study, we proposed a machine learning-based model utilizing the FSWOR technique and a projection method to address overfitting in high-dimensional brain cancer microarray data (GSE50161).

We applied the Kendall statistical test to identify 20,890 significant genes from a total of 54,675 genes across 130 samples. Our approach integrated ensemble classifiers with feature selection without replacement and projection techniques to improve classification accuracy.

The proposed model, leveraging LDA projection with Naïve Bayes, outperformed existing methods, achieving a cross-validation score of 96%.

This result demonstrates the effectiveness of our approach in high-dimensional gene expression analysis. The ability to extract biologically relevant features enhances interpretability in cancer classification.

Our findings highlight the importance of statistical feature selection and dimensionality reduction in microarray data analysis. The proposed method significantly mitigates overfitting while maintaining high predictive performance. Future work may explore deep learning techniques and other projection methods to improve the results further.

Overall, our study contributes to the advancement of computational techniques for cancer biomarker discovery and classification.

Data Availability Statement: The dataset utilized in this study is publicly available on Kaggle[2].

Reference

1. **Sartori, F., Codicè, F., Caranzano, I., Rollo, C., Birolo, G., Fariselli, P., Pancotti, C. (2025).** A Comprehensive Review Deep Learning Applications Multi-Omics Data Cancer Research. Genes, Vol. 16, No. 6, pp. 648. doi: 10.3390/genes16060648.
2. **Martinez, L.F., Gamboa, D., Valle, P.A. (2025).** Mathematical Model Gastric Cancer Immunotherapy: Global Dynamics Tumor Clearance Conditions. Computación y Sistemas, Vol. 29, No. 3.
3. **Nenclares, P., Harrington, K.J. (2020).** Biology Cancer. Medicine, Vol. 48, No. 2, pp. 67–72. doi: 10.1016/j.mpmed.2019.11.001.
4. **Wang, Y., Liu, C., Fang, C., Peng, Q., Qin, W., Yan, X., Zhang, K. (2025).** Engineered Cancer Nanovaccines: New Frontier Cancer Therapy. Nano-Micro Letters, Vol. 17, No. 1, pp. 30. doi: 10.1007/s40820-024-01533-y.
5. **Hsu, P.Y., Liou, C.F. (2025).** Impact Patient Resourcefulness Cancer Patients' Pain Management Medical Opioid Use: Cross-Sectional Study. European Journal of Oncology Nursing, Vol. 74, pp. 102–771. doi: 10.1016/j.ejon.2024.102771.
6. **Lu, X., Jin, J., Wu, Y., Lin, J., Zhang, X., Lu, S., Luan, X. (2025).** Self-Assembled PROTACs Enable Protein Degradation Reprogram Tumor Microenvironment Synergistically Enhanced Colorectal Cancer Immunotherapy. Bioactive Materials, Vol. 43, pp. 255–272. doi: 10.1016/j.bioactmat. 2024.09.022.
7. **Murugan, K., Dhivya, R., Sangeetha, C.N., Alagarsamy, M. (2025).** Design Analysis Highly Sensitive Terahertz Biosensor Early Cancer Detection Using Silver Surface Plasmon Resonance Metasurfaces Elastic Reflection Starling Murmuration Equivariant Quantum Decision Networks. ECS Journal of Solid State Science and Technology, Vol. 14, No. 1. doi: 10.1149/2162-8777/ada4da.
8. **Cai, G., Huang, F., Gao, Y., Li, X., Chi, J., Xie, J., Liu, J. (2024).** Artificial Intelligence-Based Models Enabling Accurate Diagnosis Ovarian Cancer Using Laboratory Tests China: Multicentre Retrospective Cohort Study. The Lancet Digital Health, Vol. 6, No. 3, pp. 176–186. doi: 10.1016/S2589-7500(23)00245-5.
9. **Baima, G., Minoli, M., Michaud, D.S., Aimetti, M., Sanz, M., Loos, B.G., Romandini, M. (2024).** Periodontitis Risk Cancer: Mechanistic Evidence. Periodontology 2000, Vol. 96, No. 1, pp. 83–94. doi: 10.1111/prd.12540.
10. **Abdolahi, M., Ghaedi Talkhounche, P., Derakhshan Nazari, M.H., Hosseininia, H.S., Khoshdel-Rad, N., Ebrahimi Sadrabadi, A. (2024).** Functional Enrichment Analysis Tumor Microenvironment–Driven Molecular Alterations Facilitate Epithelial-To-Mesenchymal Transition Distant Metastasis. Bioinformatics and Biology Insights, Vol. 18. doi: 10.1177/11779322241227722.
11. **Liaghat, M., Ferdousmakan, S., Mortazavi, S.H., Yahyazadeh, S., Irani, A., Banihashemi, S., Nabi-Afjadi, M. (2024).** Impact Epithelial-Mesenchymal Transition (EMT) Induced Metabolic Processes Intracellular Signaling Pathways Chemo-Resistance Metastasis Recurrence Solid Tumors. Cell Communication and Signaling, Vol. 22, No. 1, pp. 1–24. doi: 10.1186/s12964-024-01957-4.
12. **Agraval, H., Kandhari, K., Yadav, U.C. (2024).** MMPs Potential Molecular Target Epithelial-To-Mesenchymal Transition Driven COPD Progression. Life Sciences, Vol. 349, pp. 122–874. doi: 10.1016/j.lfs.2024.122874.
13. **Schiliro, C., Firestein, B.L. (2021).** Mechanisms Metabolic Reprogramming Cancer Cells Supporting Enhanced Growth Proliferation. Cells, Vol. 10, No. 5, pp. 1056. doi: 10.3390/cells10051056.

[2] https://www.kaggle.com/datasets/brunogrisci/brain-cancer-gene-expression-cumida

14. **Holá, A. (2023).** English Medical Students. Charles University Prague, Karolinum Press.
15. **Ishibashi, K., Hirata, E. (2024).** Multifaceted Interactions Cancer Cells Glial Cells Brain Metastasis. Cancer Science, Vol. 115, No. 9, pp. 2871–2878. doi: 10.1111/cas.16241.
16. **Visentin, L., Scarpellino, G., Chinigò, G., Munaron, L., Ruffinatti, F.A. (2022).** BioTEA: Containerized Methods Analysis Microarray-Based Transcriptomics Data. Biology, Vol. 11, No. 9, pp. 1346. doi: 10.3390/biology11091346.
17. **Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., Luo, Y. (2022).** Single-Cell RNA Sequencing Technologies Applications: Brief Overview. Clinical and Translational Medicine, Vol. 12, No. 3, pp. 694. doi: 10.1002/ctm2.694.
18. **Zanella, L., Facco, P., Bezzo, F., Cimetta, E. (2022).** Feature Selection Molecular Classification Cancer Phenotypes: Comparative Study. International Journal of Molecular Sciences, Vol. 23, No. 16, pp. 9087. doi: 10.3390/ijms23169087.
19. **Caraway, C.A., Gaitsch, H., Wicks, E.E., Kalluri, A., Kunadi, N., Tyler, B.M. (2022).** Polymeric Nanoparticles Brain Cancer Therapy: Review Current Approaches. Polymers, Vol. 14, No. 14, pp. 2963. doi: 10.3390/polym14142963.
20. **Feng, C., Liu, S., Zhang, H., Guan, R., Li, D., Zhou, F., Feng, X. (2020).** Dimension Reduction Clustering Models Single-Cell RNA Sequencing Data: Comparative Study. International Journal of Molecular Sciences, Vol. 21, No. 6, pp. 2181. doi: 10.3390/ijms21062181.
21. **Zhang, S., Pang, Y., Liu, F., Wang, M., Kan, Z., Song, X. (2024).** Random Projection Enhancement: Novel Method Improving Performance Surrogate Models. Swarm and Evolutionary Computation, Vol. 89, pp. 101–645.
22. **Hu, J., Hooi, B., He, B. (2024).** Efficient Heterogeneous Graph Learning via Random Projection. IEEE Transactions on Knowledge and Data Engineering. doi: 10.1109/TKDE.2024.3434956.
23. **Fabiani, G. (2024).** Random Projection Neural Networks Best Approximation: Convergence Theory Practical Applications. arXiv preprint. doi: 10.48550/arXiv. 2402.11397.
24. **Li, Y. (2024).** Probability Tools Sequential Random Projection. arXiv preprint. doi: 10.48550/arXiv.2402.14026.
25. **Asi, H., Feldman, V., Nelson, J., Nguyen, H., Talwar, K. (2024).** Fast Optimal Locally Private Mean Estimation via Random Projections. Advances in Neural Information Processing Systems, Vol. 36.
26. **McDonnell, M.D., Gong, D., Parvaneh, A., Abbasnejad, E., van den Hengel, A. (2024).** Ranpac: Random Projections Pre-Trained Models Continual Learning. Advances in Neural Information Processing Systems, Vol. 36.
27. **Kumaran, K., Sajjan, M., Oh, S., Kais, S. (2024).** Random Projection Using Random Quantum Circuits. Physical Review Research, Vol. 6, No. 1.
28. **Ahmad, M., Usman, S., Batyrshin, I., Muzammil, M., Sajid, K., Hasnain, M., Sidorov, G. (2025).** Automated Diagnosis Lung Diseases Using Vision Transformer: Comparative Study Chest X-Ray Classification. arXiv preprint. doi: 10.48550/arXiv.2503.18973.
29. **Feltes, B.C., Chandelier, E.B., Grisci, B.I., Dorn, M. (2019).** Cumida: Extensively Curated Microarray Database Benchmarking Testing Machine Learning Approaches Cancer Research. Journal of Computational Biology, Vol. 26, No. 4, pp. 376–386. doi: 10.1089/cmb.2018.0238.
30. **Ahmed-Medjahed, S., Boukhatem, F. (2024).** Applying Support Vector Machines Different Kernel Breast Cancer Diagnosis. Computación y Sistemas, Vol. 28, No. 2, pp. 659–667. doi: 10.13053/cys-28-2-4207.
31. **Medjahed, S.A., Boukhatem, F. (2025).** WOA-SVM: Whale Optimization Algorithm Support Vector Machine Hyperspectral Band Selection 2D Images Feature Selection. Computación y Sistemas, Vol. 29, No. 3.

ISSN 2007-9737

Article received on 15/07/2025; accepted on 20/11/2025.
**Corresponding author is Carlos Fernando Aguilar Ibañez.*