An Integrated Machine Learning and Genomic Framework for Precise Detection of Gastric Cancer

Eshmal Iman, Sohail Jabbar, Shabana Ramzan, Ali Raza, Farwa Raoof, Stefania Carvajal Altamiranda, Vivian Lipari, Imran Ashraf

PII: S0002-9440(26)00136-7

DOI: https://doi.org/10.1016/j.ajpath.2026.04.014

Reference: AJPA 4475

To appear in: *The American Journal of Pathology*

Received Date: 31 October 2025

Revised Date: 9 March 2026

Accepted Date: 11 April 2026

Please cite this article as: Iman E, Jabbar S, Ramzan S, Raza A, Raoof F, Altamiranda SC, Lipari V, Ashraf I, An Integrated Machine Learning and Genomic Framework for Precise Detection of Gastric Cancer, *The American Journal of Pathology* (2026), doi: https://doi.org/10.1016/j.ajpath.2026.04.014.

An Integrated Machine Learning and Genomic Framework for Precise Detection of Gastric Cancer

Eshmal Imana, Sohail Jabbarb, Shabana Ramzanc, Ali Razad, Farwa Raoofd, Stefania Carvajal Altamirandah,f,g, Vivian Liparih,1,i, Imran Ashraf*j

aDepartment Computer Science University of Engineering and Technology Taxila Pakistan Rahim Yar Khan Pakistan.

bCollege of Computer and Information Sciences Imam Mohammad Ibn Saud Islamic University (IMSIU) Riyadh 11432 Saudi Arabia

cDepartment of Computer Science and IT Government Sadiq College Women University Bahawalpur 63100 Pakistan

dDepartment of Software Engineering University of Lahore 54000 Lahore Pakistan

eUniversidad Europea del Atlantico. Isabel Torres 21 39011 Santander Spain

fUniversidade Internacional do Cuanza. Cuito Bie Angola.

gFundacion Universitaria Internacional de Colombia Bogota Colombia.

hUniversidad Internacional Iberoamericana Campeche 24560 Mexico.

iUniversidad de La Romana. La Romana Republica Dominicana

jInformation and Communication Engineering Yeungnam University Gyeongsan 38541 Korea

Running Title: Framework for Gastric Cancer Detection

Conflicts of Interests: The authors declare that there is no conflict of interests.

Funding: This research is funded by the European University of Atlantic.

***Corresponding Author**: Imran Ashraf, Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea; imranashraf@ynu.ac.kr

Abstract

This study presents a novel integrative approach for the analysis of high-dimensional gene expression data, leveraging the complementary strengths of unsupervised clustering and supervised classification. Using K-means clustering, the dataset is stratified into three distinct clusters, revealing intrinsic biological patterns and relationships. The resulting cluster assignments are subsequently employed as pseudo-labels to train machine learning models, including support vector machines, random forest, and a stacking ensemble classifier. To validate and enhance the robustness of clustering, complementary methodologies such as hierarchical clustering and DBSCAN are employed, with results visualized through PCA-driven dimensionality reduction. The high predictive accuracy achieved by the classifiers underscores the separability and reliability of the identified clusters. Furthermore, feature importance analysis highlighted key genetic determinants within each cluster, offering actionable insights into potential biomarkers and critical genomic features. This framework bridges the gap between exploratory unsupervised learning and predictive supervised modeling, providing a scalable and interpretable methodology for analyzing complex genomic datasets. Its applicability extends to biomarker discovery, patient stratification, and other precision medicine applications, emphasizing its utility in advancing genomic research and clinical practice.

Keywords: Gastric cancer, histological images, k-means clustering, unsupervised learning, convolutional neural networks, image processing

Introduction

Stomach cancer, commonly referred to as gastric cancer, ranks as the fifth most frequently diagnosed cancer across the globe (1) and remains a significant global health concern. Each year, over one million people are affected by this disease, making it one of the top 3 leading causes of mortality caused by cancer globally. It is crucial to detect gastric cancer in its early stages to reduce mortality rates (2). Its symptoms, such as mild nausea, heartburn, rectal bleeding, and weight loss, can manifest early on (3).

The Cancer Diagnosis Society determined almost 26,500 new cases in 2023 (4) while this figure was estimated at 1,089,103 globally, in 2020 (5). The estimated death toll due to gastric cancer recorded in the United States was 11,130 in 2023 (6), while the worldwide statistics were 768,793 for 2020. It also provides the piece of information that 60% of people identified with gastric cancer in the United States are older than the age of 64. H. pylori chronic infection is considered one of the most significant risk factors contributing to gastric cancer (7). Additionally, the physical and psychological burden on patients, including psychiatric disorders and apprehension, further exacerbates the negative impact of this disease (8).

With the advancement of computational tools and bioinformatics, machine learning (ML) has emerged as an essential tool in handling large-scale biological data, particularly genomic data. This integration has enabled the identification of genetic markers associated with gastric cancer and enhanced the prediction of treatment outcomes. ML has revolutionized cancer diagnosis, disease understanding, and personalized treatment. The integration of ML and data analytics enables the identification of genetic markers associated with the disease and the prediction of treatment progress.

Problem Statement

Gastric cancer is a major global health burden, ranking among the leading causes of cancer-related mortality. Despite significant advances in oncology, the disease remains difficult to diagnose at an early stage, with the majority of cases being identified only after the cancer has progressed to advanced stages. Current diagnostic practices, largely reliant on histopathological examination, are not equipped to fully capture the intricate genetic heterogeneity of gastric cancer. This limitation impairs the precision of subtype classification, which is essential for tailoring personalized treatment strategies.

Moreover, existing therapeutic approaches are often generic and fail to address the unique molecular characteristics of individual patients, leading to suboptimal outcomes. Gastric cancer is a highly complex and multifaceted disease, so there is a pressing need for innovative methods that can integrate genetic insights into clinical practice to enhance early detection, precise classification, and targeted treatment.

The primary challenge lies in identifying the genetic markers that drive gastric cancer and leveraging these markers to refine diagnostic accuracy. This research seeks to bridge this critical gap by employing ML algorithms to uncover genetic subtypes of gastric cancer. Through gene expression profiling and advanced computational models, this study aims to revolutionize

gastric cancer diagnostics, enabling more accurate subtype classification and personalized treatment. Thus, it aims to improve patient survival rates and reduce the global burden of this devastating disease.

Research Contributions

The primary objective is to detect genetic markers and understand genome information related to gastric cancer. This study utilizes genomic data to identify explicit molecular subtypes of gastric cancer, which enables more accurate diagnosis and treatment. Gene expression profiling differs from traditional classification methods that typically rely on histopathological features, offering a dominant and prevailing tool for classifying gastric cancer with exclusive biological characteristics and clinical consequences.

The proposed research combines unsupervised and supervised learning methods to develop a unified approach for gene expression analysis in gastric cancer. The dataset contains genes with significant biological relevance, such as A1CF, ABCA12, ABCC5, and others, which play a crucial role in various cellular processes and disease pathways. K-means clustering is applied for data preprocessing and feature extraction, while classification algorithms like support vector machines (SVM) and random forest (RF) are employed for gene association classification. The stacking classifier, an ensemble learning technique, provides remarkable precision in classifying genome expression profiles.

The primary motivation for this research is the urgent need for more accurate, efficient, and personalized treatment strategies for gastric cancer. Current diagnostic methods largely rely on histopathological features, which, while informative, do not capture the full spectrum of genetic diversity present in gastric cancer. By focusing on gene expression profiling, this study aims to provide a deeper understanding of the molecular subtypes of gastric cancer, offering a path toward more precise diagnostic tools and therapeutic approaches.

The goal of this research is to identify specific genetic markers associated with gastric cancer through gene expression profiling. By leveraging ML techniques, this study aims to classify molecular subtypes of gastric cancer more accurately and effectively. The insights gained from this research can accelerate the discovery of targeted treatments, ultimately improving patient outcomes and contributing to a better understanding of the disease's underlying mechanisms.

Literature Review

The study presented in (9) explores the application of ML algorithms, such as RF, Naive Bayes, and k-nearest neighbor (KNN), in developing a diagnostic model for identifying differentially expressed genes associated with gastric cancer. To construct this model, transcriptional data, and genomic data from the cancer genome atlas database based on gastric cancer and non-gastric cancer tissues were obtained. Meanwhile, the gene expression omnibus database served as a validation source for the proposed gastric cancer diagnostic model. The research pinpointed ten key genes LINC01821, ADGRD1-AS1, CCNB1, AL138826.1, AC022164.1, KIF11, AURKB, NUSAP1, CDK1, and TTK that were essential for the diagnostic model. Results indicated that ML approaches significantly enhanced efficiency, with the RF demonstrating the highest

performance, scoring an accuracy of 92% and an AUC of 97.22%. The model underwent further validation using the GSE54129 dataset, (https://www.ncbi.nlm.nih.gov/geo/) attaining an AUC of 90.4%.

In (2), the authors set up a non-invasive blood-based examination using the Affymetrix gene profiling microarray to scan profiles of gene expression data in peripheral blood samples from gastric cancer patients and compared them with those in healthy controls and subjects having other malignancies. The dataset contained 216 blood samples (36 with gastric cancer, 55 with healthy controls, and the rest were patients with other carcinomas). The study found four-gene signatures of peripheral blood derived from purine-rich elements, structural maintenance of chromosomes 1A (SMC1), DENN domain containing 1B, and cell programmed death protein 4(PDCD4) in gastric cancer. The overall accuracy of the four-gene panel was 95%, and specificity was just under 96%.

The study (10) develops a cost-effective, non-invasive, rapid, and highly accurate diagnostic framework to categorize individuals based on their likelihood of developing gastric cancer, considering various lifestyle factors. Six ML algorithms were utilized to build the model, including multilayer perceptron (MLP), SVM with both linear and RBF kernels, RF, KNN, and extreme gradient boosting. The dataset, comprising 2029 participants, was sourced from the dataset of Ayatollah Taleghani hospital in Iran. Feature selection was performed using the ReliefF algorithm, incorporating attributes such as high salt intake, chronic atrophic gastritis, and fruit consumption. The results demonstrated that gradient boosting achieved the maximum performance, with an accuracy of 83.4%. In contrast, the KNN classifier with K = 7 yielded a mean accuracy of 67.98%. A notable limitation of this research is its reliance on retrospective data, which may impact the generalizability of the findings.

The authors found target miRNA (microRNA) biomarkers in (11) for the diagnosis of gastric cancer, particularly in its early stages. Feature selection of the most important miRNAs was accomplished using a Boruta ML algorithm and balanced with SMOTE random oversampling. Further, fivefold cross-validation detected not only the best hyperparameters of DT, RF, LR, and XGBT but the artificial neural network as well. Specifically, hsa-miR-1343-3p was identified in predicting GC using a ROC cut-off of 8.2 for other types of cancer miRNAs. The study used the GSE106817 dataset, (https://www.ncbi.nlm.nih.gov/geo/) which includes 2,566 miRNAs. The results indicate that among all identified miRNAs, hsa-miR-1343-3p played the most significant role in prediction accuracy, with importance values ranging from a minimum of 6.47 to a maximum of 13.63, a median of 11.44, and a mean of 10.81. Following this, hsa-miR-1290 and hsa-miR-5100 ranked second in importance, exhibiting mean importance values of 8.69 and 8.66, respectively.

The authors in (12) identified ML in combination with bioinformatics that was adopted for identifying some of the potential biomarker candidates responsible for playing a crucial role in the pathogenicity or etiology of gastric cancer. The nanoString profiling transcriptomic approach was conducted on SAMIT resection data samples. After internal cross-validation within SAMIT samples, the model was validated once again with an external cohort. To further validate the constructed nomogram, a single-arm independent group of metastatic GC patients

treated with paclitaxel and ramucirumab was obtained from previous clinical trials as an external validation cohort. The NanoString panel contained 476 genes from pathways of potential relevance to cancer, including cell cycle, DNA damage repair, and immune response. For the Pac-UFT validation cohort, the gene signature of the random forest had an accuracy of 61 %, F-measure was 71%, and AUC was 75% (95% CI 0.50 to 0.99). The RF model identified a 19-gene signature that could predict paclitaxel benefit with an accuracy of 61% in the internal data validation cohort and 64% in the external validation cohort. The F1 measure was 71% and 62% in its internal and external validation cohorts, respectively.

The research conducted in (13) explores the molecular signatures of gastric cancer across various system levels, including hub proteins, receptor transcription factors (TFs), and receptors, utilizing an integrative multi-omics framework combined with ML techniques. Multiple classification algorithms were employed to assess the potential of novel biomarkers. Three microarray datasets, GSE54129, GSE19826, and GSE79973, (https://www.ncbi.nlm.nih.gov/geo/) were obtained from the NCBI-GEO dataset. The features analyzed comprised reporter transcripts and receptor molecules associated with gastric cancer. The results indicated that the proposed biomarkers were less effective in distinguishing between live and dead specimens, with accuracy ranging from 64.6% to 47.7%. In (14), the authors built an interaction network to forecast new biomarkers for gastric cancer using text mining, network analysis, ML, deep learning (DL), and structural bioinformatics approaches. Features include gene expression levels, protein-protein interactions, mutations, epigenetic modifications, and other molecular features related to gastric cancer. The four putatively unique and potential biomarker genes have been identified.

The focus of (15) was the development of a prognostic approach for gastric cancer patients based on immune-related long non-coding RNAs (lncRNAs). The ImmLnc algorithm was employed to identify immune-related lncRNAs, followed by univariate Cox regression analysis. A combination of ten algorithms was utilized to construct the lncRNA prognostic model. Additional analytical methods, such as ssGSEA, the CIBERSORT algorithm, map tools, prophetic, and clusterProfiler, were also implemented. The dataset was obtained from The Cancer Genome Atlas (TCGA) database. The core concept of the algorithm involves evaluating tumor purity using the ESTIMATE model, computing partial correlation coefficients between mRNAs and lncRNAs while accounting for tumor purity, and ranking mRNAs linked to each respective lncRNA. Findings demonstrated that the proposed 18-lncRNA signature performed effectively in both the training and validation datasets.

The research presented in (3) identified potential diagnostic and prognostic miRNAs in gastric cancer. Several ML algorithms, including SVM, RF, and KNN, were applied. The dataset was sourced from the online TCGA database. Among the evaluated models, SVM achieved the maximum accuracy of 93% and an AUC score of 88.5%. In (16), a prognostic approach was developed to enhance the accuracy of stomach cancer diagnosis in post-operative patients. The LASSO regression method was employed to construct the clinical prognostic model. Feature selection was performed using the Boruta algorithm, NNs, SVM, and RF. Data for 955 post-operative gastric cancer patients were retrieved from the SEER database. The findings indicated

that the AUC values for predictions at 1, 3, and 5 years in the training, validation, and external validation datasets consistently remained around 0.8.

In the same way, (17) investigated the predictive capabilities of ML in assessing the recurrence of gastric cancer in post-operative patients. RF, light gradient boosting machine (LightGBM), gradient boosting (GB), decision trees (DT), and logistic regression (LR) were implemented. The dataset, consisting of 2012 patients, was acquired from the BioStudies database. Results revealed that the accuracy of the GB was 89.1%, while AUC values for the models were 96.2%, 92.2%, 89.8%, 79.0%, and 74.8% for RF, LightGBM, GB, DT, and LR, respectively. In (18), researchers developed a predictive model utilizing the gradient-boosting decision tree approach to diagnose gastric cancer based on noninvasive features. The dataset included records from 709 patients at Zhejiang Provincial Hospital. The findings indicated that the model achieved an AUC of 91%, a sensitivity of 87.0%, and a specificity of 84.1% at a threshold value of 0.56, with an overall accuracy of 83.0%. A comparative summary of the reviewed studies is provided in Table 1.

Materials and Methods

Figure 1 shows the methodological workflow. This study incorporates data preprocessing, unsupervised learning approaches such as K-means clustering, and supervised learning algorithms that include SVM, RF, and stacking classifiers. The methodology is based on several sequential steps, illustrated below, to ensure a comprehensive approach to gene expression analysis related to gastric cancer.

Data Acquisition

The study utilized the TCGA dataset (19), a coordinated effort to understand cancer's molecular basis using genome analysis technologies. This study focused on gene expression profiles linked to gastric cancer. The dataset contains information on the expression of 8863 genes, including various genetic alterations and mutations linked to gastric cancer. The data collection phase is critical because it establishes the foundation for the entire research, ensuring that the dataset is robust, diverse, and representative of the gastric cancer population, as shown in Table 2.

Data Preprocessing

Data preprocessing is an important step in ensuring the quality and consistency of the data used in analysis. This step included several key activities.

Removal of unnecessary data

During the preprocessing phase, the removal of unnecessary data is crucial to enhance the quality of the dataset and reduce potential noise. Only non-informative and redundant columns unrelated to gene expression features were removed. Specifically, the columns "Unnamed: 0" (auto-generated index column) were excluded from analysis. No gene expression features were removed at this stage, as shown in Table 3.

Relevant Genes Extraction

Relevant genes were extracted to focus on the most informative genes for gastric cancer classification. Feature selection techniques were employed to identify genes with significant variance and biological relevance.

The analysis began with unsupervised clustering to detect intrinsic patterns and groupings in the gene expression data. We chose the K-means algorithm by applying 3 clusters (20). The features related to gastric cancer are extracted in this procedure. The three associated genes with gastric cancer subtype Adenocarcinomas and adenomas like TCGA-RD-A7BS-01 (class 0), TCGA-HU-A4GD-01(class 1), and TCGA-CG-5719-01(class2) are extracted for further processing.

Training Phase

Following the clustering phase, we moved on to supervised learning with the goal of predicting the clusters identified by K-means. This step involved training classifiers on labeled data obtained during the clustering step. We tested the effectiveness of different ML approaches for the classification of gastric cancer subgroup gene expression clusters. The objective function for training a classifier is formulated as follows:

$$\hat{\theta} = \operatorname{argmin}_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell\left(h_{\theta}(x_i), y_i\right)$$

Where,

$h_{\theta}(x_i)$ is the model's prediction for x_i.

ℓ is the loss function (e.g., cross-entropy or mean squared error).

N is the total number of data points.

In this objective function, $\hat{\theta}$ represents the optimal parameters of the classifier, which are learned by minimizing the average loss across all data points.

Cluster Analysis

The resulting clusters were examined to determine which genes contribute significantly to each cluster. This analysis identified three critical genes: TCGA-RD-A7BS-01 (class 0), TCGA-HU-A4GD-01 (class 1), and TCGA-CG-5719-01 (class 2), which play important roles in expressing the genetic subtypes of gastric cancer.

Biological Interpretation

The clusters were further investigated for their biological relevance, and their relationship with known gastric carcinoma subtypes (e.g., adenocarcinomas and adenomas) was determined.

Applied Clustering Techniques

K-means Clustering

K-Means clustering is a commonly used unsupervised machine learning technique that divides a dataset into k unique and non-overlapping clusters. Each cluster is defined by a centroid, representing the average position of all data points within it. The main goal of the K-Means algorithm is to reduce the within-cluster sum of squares (WCSS), promoting compact and distinct groupings.

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

where C_j is the set of points assigned to cluster j and μ_j is the cluster's centroid.

Selection of K-means for Clustering

The decision to employ K-means clustering in this study stems from its effectiveness in handling large datasets and its ability to produce interpretable clusters. The K-means algorithm minimizes the within-cluster sum of squares (WCSS), defined mathematically as follows.

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in C_i} \| x_j - \mu_i \|^2$$

where k is the number of clusters points, C_i is the set of points in cluster i, x_j is a data point, and μ_i is the centroid of cluster i. Additionally, K-means is computationally efficient, making it suitable for the scale of our data.

Relevant genes are extracted using k-means to focus on the most informative genes for gastric cancer classification. Feature selection techniques were employed to identify genes with significant variance and biological relevance. To determine the optimal number of clusters, silhouette analysis was conducted for k ranging from 2 to 6. Although k = 2 yielded the highest silhouette score (0.276), this configuration was deemed overly coarse and insufficient to capture molecular heterogeneity within gene expression profiles.

Determining the Optimal Number of Clusters

Selecting the appropriate number of clusters (k) is crucial for meaningful clustering results. In this study, we determined ($k = 3$) based on the Elbow method, which involves plotting the WCSS against various values of (k) and identifying the point where the rate of decrease sharply changes, indicating an optimal (k).

For our dataset, after experimenting with multiple values of k, the WCSS plot exhibited a noticeable elbow at $k = 3$. This suggests that the three clusters provided a balanced separation of data points and effectively captured the underlying structure of the gene expression profiles related to gastric cancer.

Implementation Details

The K-means clustering process involved partitioning the dataset into 3 clusters, which were identified based on their genetic relevance to gastric cancer subtypes. The features related to gastric cancer were extracted in this procedure, including genes associated with gastric cancer subtypes like Adenocarcinomas and adenomas. The Algorithm in Table 4 provides working details.

The clustering results were then analyzed to determine the contribution of specific genes to each cluster, focusing on three key genes: TCGA-RD-A7BS-01, TCGA-HU-A4GD-01, and TCGA-CG-5719-01, which were identified as significant markers for gastric cancer subtypes. The expression levels of these genes were quantitatively assessed to evaluate their discriminative power across clusters. These genes were selected for further analysis and used in subsequent supervised learning steps to refine the predictive model.

Supervised Learning Models

Several supervised learning models were used to classify gene expression clusters identified through K-means clustering accurately.

SVM: It was chosen because of its ability to handle high-dimensional data and create a strong decision boundary. The hyperparameters were tuned through grid search and cross-validation methods. To minimize the hinge loss in SVM, we utilize the hinge loss function:

$$L(w) = \frac{1}{N}\sum_{i=1}^{N} \max(0,1 - y_i(w \cdot x_i)) + \frac{\lambda}{2} \parallel w \parallel^2$$

Gradient descent helps in minimizing this loss function by iteratively updating the weights w. The update rule is given by:

$$w = w - \eta \frac{\partial L}{\partial w}$$

where η is the learning rate. This process allows the model to converge towards the optimal weight vector that effectively separates the classes while minimizing the classification error.

RF: his ensemble learning method was chosen because of its ability to handle large, high-dimensional datasets and its resistance to overfitting. The mathematical representation of Gini Impurity for a node is given by.

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i is the probability of a data point belonging to class i. RF minimizes Gini Impurity at each node to build better decision trees. By selecting splits that maximize the purity of the resulting child nodes, the model effectively captures the underlying structure of the data. Additionally, RF's feature importance scores were used to interpret model decisions.

Stacking Classifier: To combine the strengths of multiple models, a stacking classifier (21) was used. This meta-classifier improved overall performance by combining predictions from several base classifiers, including SVM and RF. The stacking model was trained using a meta-learning algorithm, which learned how to best combine the predictions of the base models.

These supervised learning models were trained on clustered data and evaluated using metrics such as accuracy, precision, recall, and F1 score. Cross-validation was used to ensure that the techniques generalize well to new data. The results showed high accuracy and reliability in classifying the genetic subtypes of gastric cancer, demonstrating the effectiveness of the proposed methodology. The Algorithm in Table 4 provides the working details of the proposed approach.

Results

In the results section, the dataset was split into 80% for model training and internal cross-validation, and 20% for independent testing. Cross-validation was applied exclusively within the training set to optimize model parameters and prevent data leakage.

Computational Setup and Reproducibility

All experiments were performed using Python (version 3.10), with machine learning models implemented through the scikit-learn library (version 1.3.2), alongside standard scientific computing packages, including NumPy and Pandas. To ensure reproducibility, a fixed random seed (random_state = 42) was consistently applied across data partitioning, clustering, and model training procedures. Hyperparameter optimization for supervised classifiers was carried out using grid search in conjunction with cross-validation. All preprocessing, clustering, and classification steps were executed within deterministic computational pipelines to minimize variability and maintain methodological consistency.

Evaluation Metrics

The evaluation metrics for the classifiers are detailed in Table 5. These metrics are important indicators of the models' performance, providing a thorough evaluation of their ability to classify genetic subtypes of gastric cancer correctly. The table includes the formulas for accuracy, precision, recall, and F1-score, making it clear how these metrics are calculated. Furthermore, the table displays the actual values of these metrics for each classifier, providing insight into their strengths and potential weaknesses. The evaluation also considers the impact of false positives (FP) and false negatives (FN) (22), which are important in determining the reliability of classifiers. This thorough analysis ensures that the models are not only effective but also reliable in distinguishing between different genetic subtypes, supporting the research's overall goals.

Result Analysis for Random Forest

The RF classifier demonstrated robust performance, achieving an overall accuracy of 97.29%. This high accuracy indicates that the model correctly identified the classes in the majority of

cases. The precision, recall, and F1 scores were consistently high across all classes, showcasing the model's balanced and reliable performance.

Figure 2 illustrates the confusion matrix analysis for RF. For class 0, there are 527 true positives (TP), 1232 true negatives (TN), 25 false positives (FPs), and 25 false negatives (FNs). For class 1, there are 766 TP, 1283 TN, 18 FP, and 9 FN. For class 2, there are 432 TP, 1341 TN, 7 FP, and 14 FN. RF made 1,725 correct predictions alongside 48 wrong predictions.

Table 6 contains the RF result with macro-averaged precision, recall, and F1 scores of 95.23%, 96.67%, and 95.95%, respectively. These metrics indicate that the model is highly precise and reliable in identifying true positives while maintaining a strong ability to correctly classify instances across all classes. The high recall values demonstrate that the model rarely misses instances, and the F1 scores reflect a balanced performance by harmonizing precision and recall.

Result Analysis for Support Vector Machine Classifier

The SVM classifier exhibited exceptional performance, achieving the highest performance among the tested models with a 99.15% accuracy. This indicates the model's superior capability in distinguishing between several classes.

Figure 3 illustrates the confusion matrix analysis. For class 0, there are 544 TPs, 1225 TNs, 5 FPs, and 3 FNs. For class 1, there are 775 TPs, 1300 TNs, 0 FP, and 0 FN. For class 2, there are 443 TPs, 1324 TNs, 3 FPs, and 0 FN. SVM shows better performance than RF with 1,762 correct predictions and made only 11 wrong predictions for three class problems.

Table 6 contains the SVM result having macro-averaged precision, recall, and F1 scores are 98.76%, 98.91%, and 98.83%, respectively. These results underscore the robustness of the SVM classifier, as it not only maintains high accuracy but also effectively minimizes false classifications across all classes.

Result Analysis for Stacking Classifier

The stacking classifier performed exceptionally well with an overall accuracy of 98.87%. It successfully combines the predictions of multiple models to enhance its performance.

Figure 4 illustrates the confusion matrix analysis. For class 0, there are 526 TPs, 760 TNs, 16 FPs, and 10 FNs. For class 1, the classifier produced 760 TPs, 970 TNs, 6 FPs, and 9 FNs. For class 2, it achieved 434 TPs, 1217 TNs, 12 FPs, and 9 FNs. The stacking model performed well with a total of 1,720 correct predictions while 53 samples are incorrectly predicted in this case.

Table 6 contains the stacking classifier result having macro-averaged precision, recall, and F1 scores are 97.65%, 97.89%, and 97.77%, respectively, indicating a strong, balanced performance across all classes. This consistency across metrics highlights the model's robustness and reliability in handling multiple categories. By reducing misclassification and maintaining high precision, the stacking classifier proves itself as a solid choice for complex

classification problems. Its ability to synergize diverse algorithms makes it a versatile and efficient tool in predictive modeling.

Receiver Operating Characteristic Curve

The receiver operating characteristics (ROC) curve (23) was plotted for each classifier to visualize their performance across different threshold settings. The area under the curve (AUC) was calculated for each model, providing an aggregate measure of performance across all classification thresholds. The SVM classifier achieved an AUC of 99%, indicating excellent discriminative ability, followed by the stacking classifier with an AUC of 98.5% and the RF classifier with an AUC of 97.5%.

Feature Importance Analysis

To gain feature insights into the influence of each feature on the model's predictions, two complementary approaches to feature importance analysis were utilized: traditional feature importance derived from RF and Shapley Additive exPlanations (SHAP) values. Each method provides unique advantages in understanding the role of specific features within the model, offering a more comprehensive view when combined.

Feature Importance Using RF

The first approach involved calculating the importance of features using the RF classifier. This method assesses the contribution of each feature based on its ability to reduce the impurity in the model's decision trees, where features frequently used in high-impact splits are deemed more important. The analysis was constrained to the top 10 most important features, which are illustrated in Figure 5.

As depicted, features such as TCGA-BR-8589-01 and TCGA-VQ-A94T-01 emerged as the most significant contributors to the model's predictive accuracy. However, it is worth noting that while traditional feature importance provides a global ranking of features, it lacks the capacity to capture interactions between features and does not explain how feature importance varies across individual predictions.

SHAP-based Feature Importance

To address the limitations of the traditional approach, SHAP values were employed as a second method of feature importance analysis. SHAP values, grounded in cooperative game theory, offer a more granular, instance-level interpretation of feature contributions. This method quantifies the impact of each feature on the model's output for each individual sample, taking into account interactions between features that may influence predictions.

In Figure 6, the SHAP summary plot demonstrates that features such as TCGA-BR-8589-01 and TCGA-VQ-A94T-01 have varying SHAP values across samples, indicating that these features contribute differently to individual predictions. The colors in the plot correspond to feature values (high or low), providing further interpretability regarding the direction and magnitude of

each feature's effect. By capturing the importance of both individual and global features, SHAP delivers deeper insights into the model's behavior.

Significance of the Approaches

By employing both traditional and SHAP-based feature importance methods, this study leverages the strengths of each to achieve a more comprehensive understanding of the model. Traditional feature importance offers a computationally efficient, global view of influential features, while SHAP enhances interpretability by considering feature interactions and providing instance-specific insights. Combining these methods ensures both performance and transparency, building trust in the model's predictions and facilitating better decision-making. This dual approach also enables deeper exploration of local and global feature behavior, uncovering key drivers of model output, detecting biases, and making the model more robust for deployment.

Clustering Analysis of Genomic Data

In order to evaluate the underlying structure of the data and find potential patterns related to the detection of gastric cancer, we used three distinct clustering techniques, K-means, hierarchical clustering, and DBSCAN, on the PCA-transformed genomic dataset.

Figure 7 demonstrates the clustering results from K-means, hierarchical clustering, and DBSCAN. The visualizations illustrate the capabilities and limitations of each clustering algorithm in processing high-dimensional genomic data.

K-means Clustering

As illustrated in the leftmost subplot of Figure 8, K-means splits the data into three unique groups. This method is especially useful in situations when specified clusters are predicted, such as distinguishing between different genetic subtypes of stomach cancer.

Hierarchical Clustering

The middle subplot shows how hierarchical clustering detects three groups in a more hierarchical structure. This method is useful in genomic research because it allows for the investigation of hierarchical links, such as genetic lineage or evolutionary relationships between malignant cells.

DBSCAN Clustering

The subplot on the right-hand side highlights the usefulness of DBSCAN in detecting anomalies and noise. Although the approach yields a single dominant cluster, the red-colored noise points may indicate uncommon yet potentially noteworthy genetic variations, underscoring the usefulness of DBSCAN in detecting outliers in the intricate genomic terrain.

Figure 8 presents a more refined visualization of these clustering results with enhanced label clarity. The better visualization enables a more accurate interpretation of clustering data,

notably in discriminating between noise and actual genomic clusters, which is crucial for comprehending stomach cancer's genetic variability.

Comparison of Classifiers Metrics

The classifiers' comparison based on the evaluation metrics reveals that the SVM outperformed the other models with the highest accuracy, F1, precision, and recall score. The stacking classifier, which combines multiple models, also showed strong performance, indicating the advantage of leveraging different algorithms. The Random Forest classifier, while slightly behind SVM and the stacking classifier, still demonstrated robust performance with high accuracy and balanced metrics, as shown in Figure 9.

Comparison with Existing Approaches

To validate the proposed approach, we compared it with existing studies on gastric cancer diagnosis using gene expression data, as shown in Figure 10. The results reveal that the SVM classifier significantly outperforms traditional methods, setting a new benchmark in diagnostic performance. SVM not only achieved superior accuracy but also excelled in precision, recall, and F1 score, surpassing other techniques with remarkable consistency. These substantial improvements underscore the SVM classifier's exceptional capability to identify true positives and provide a nuanced understanding of genetic subtypes, revolutionizing diagnostic accuracy and offering unprecedented insights.

Overall, the proposed approach represents a significant advancement in diagnostic performance, delivering an exceptionally effective tool for clinicians and researchers and potentially transforming the landscape of gastric cancer detection. Fei Kong, et al. (9) developed a learning model to diagnose gastric cancer. The model, based on the RF, achieved a 92% accuracy and a 91% F1 score. The model was trained on transcriptional, genomic, and clinical data from gastric and non-cancer tissues sourced from the TCGA and GEO (GSE54129, https://www.ncbi.nlm.nih.gov/geo/) databases. This research demonstrates the potential of ML for improving gastric cancer diagnosis. Afrash et al. (10) compared various ML algorithms, finding XGBoost to be the most effective. The XGBoost model achieved an accuracy of 83.4%. Azari et al. (3) used ML to predict miRNA biomarkers for gastric cancer, where SVM outperformed, achieving 93% accuracy. Zhou et al. (17) evaluated ML models to predict the recurrence of gastric cancer in patients after surgery. On the testing set, the logistic regression model scored the highest accuracy of 80.1%.

Training and Validation Accuracy Analysis

Figure 11 illustrates the training and validation accuracy trends for the RF, SVM, and stacking classifier models over a span of 20 epochs. The RF model demonstrates a compatible increase in training accuracy, eventually approaching near-perfect levels. The validation accuracy also improves, albeit at a more moderate rate, indicating effective model learning while avoiding significant overfitting. The steady increase in validation accuracy evidences this.

The SVM model follows a similar upward trajectory for both training and validation accuracies. Although there is a noticeable gap between them, suggesting minor overfitting, the continual rise in validation accuracy implies that the model is successfully extracting meaningful patterns from the dataset.

In contrast, the stacking classifier exhibits a more balanced progression, with training and validation curves that remain relatively close throughout the epochs. The persistent improvement in validation accuracy underscores the model's strong generalization capabilities, highlighting its adaptability to diverse data characteristics. Overall, none of the models displays underfitting, and while slight overfitting is observed, particularly in the SVM model, their generalization performance remains robust and reliable.

Permutation Test Analysis

To evaluate whether the observed classification performance could be attributed to random chance, a permutation test was conducted on the best-performing classifier. In this analysis, class labels were randomly permuted multiple times, and the classification accuracy was recalculated for each permutation to generate a null distribution representing chance-level performance. As shown in Figure 12, the distribution of accuracies obtained under label permutation is concentrated at substantially lower values, whereas the true model accuracy lies far beyond this null distribution. The resulting permutation p-value ($p = 0.009$) indicates that the probability of achieving the observed accuracy by chance is less than 1%.

Feature Stability Analysis and clinical utility

Beyond feature importance magnitude, the stability of selected features across multiple training splits is critical for ensuring robustness and generalizability. To address this, a feature stability analysis was conducted using repeated cross-validation, measuring how frequently top-ranked features appeared across different folds. Figure 13 shows the occurrence frequency of the most stable features across cross-validation folds. Several features consistently appeared in a majority of folds, indicating strong stability and reducing the likelihood that their importance arises from random sampling effects. This consistency suggests stability of feature rankings across different training splits. The inclusion of feature stability analysis provides an additional assessment of robustness across training splits.

Receiver Operating Characteristic (ROC) Curve Analysis

The discriminative capability of the supervised classifiers was further assessed using macro-averaged Receiver Operating Characteristic (ROC) curves, providing a threshold-independent evaluation of classification performance. As shown in Figure 14, the ROC curves for all three classifiers closely approach the upper-left region of the coordinate space, indicating a high degree of separability among the cluster-derived classes. The corresponding macro-averaged Area Under the Curve (AUC) values were 0.9992 for Random Forest, 0.9998 for Support Vector Machine (SVM), and 0.9998 for the Stacking classifier. These near-unity AUC values demonstrate strong discriminative performance within the constructed feature space.

Discussions

The results underscore the pivotal role of machine learning in advancing genomic analysis, particularly in enhancing the accuracy of gastric cancer diagnosis and treatment. By identifying genetic markers and molecular subtypes, machine learning algorithms offer valuable insights into disease progression, which are essential for developing personalized treatment strategies. The integration of both unsupervised and supervised learning methods provides a holistic approach to gene expression analysis, facilitating a deeper understanding of the molecular mechanisms. This dual approach not only enhances our ability to decipher complex genomic data but also holds significant promise for developing more targeted, effective therapeutic interventions.

Conclusions and Future Directions

This research presents the integration of bioinformatics, machine learning, and genomic analysis, marking a significant step forward in enhancing the diagnostic accuracy of gastric cancer. Employing a dual approach involving unsupervised and supervised techniques showed better results. We successfully identified pivotal genetic markers and molecular subtypes, thus contributing to the advancement of personalized medicine and improved patient prognostication. Through the application of unsupervised clustering techniques, particularly K-means, distinct genetic signatures associated with various gastric cancer subtypes have been uncovered, providing deeper insights into the disease's molecular heterogeneity. On the other hand, supervised algorithms like support vector machines (SVM) and random forest (RF) have enabled precise classification, reinforcing the effectiveness of these models in tackling complex genomic data.

The findings of this study not only demonstrate the power of advanced analytical techniques in refining gastric cancer diagnostics but also highlight the crucial role of machine learning in unraveling the genetic intricacies of the disease. The impressive accuracy of the classifiers showcases their robustness in reliably identifying and characterizing genetic subgroups, which is essential for tailoring treatment strategies to individual patients. Ultimately, this research sets a new benchmark in gastric cancer diagnosis, offering promising pathways for more accurate, targeted therapies that hold the potential to revolutionize patient care.

Future research will prioritize the expansion of the dataset to include a broader array of genetic profiles, with a particular focus on identifying novel biomarkers that could further refine the early detection of gastric cancer. Beyond enhancing the performance of existing machine learning models, the investigation will extend to advanced methodologies, such as deep learning and ensemble learning techniques, aimed at improving both the interpretability and clinical applicability of the models. This comprehensive approach, combining dataset expansion with the exploration of cutting-edge algorithms, will not only provide deeper insights into the genetic underpinnings of gastric cancer but also facilitate the development of more precise diagnostic tools and personalized therapeutic strategies, ultimately advancing clinical outcomes.

References

1. Smyth EC, Nilsson M, Grabsch HI, Grieken NC van, Lordick F. Gastric cancer. The Lancet. 2020;396(10251):635-648.

2. Shi J, Cheng C, Ma J, Liew CC, Geng X. Gene expression signature for detection of gastric cancer in peripheral blood. Oncology Letters. 2018;15(6):9802-9810.

3. Azari H, Nazari E, Mohit R, et al. Machine learning algorithms reveal potential miRNAs biomarkers in gastric cancer. Scientific Reports. 2023;13(1):6147.

4. Adams A, Gandhi A, In H. Gastric cancer: A unique opportunity to shift the paradigm of cancer disparities in the united states. Current Problems in Surgery. 2023;60(10):101382.

5. Thuler LCS. The epidemiology of stomach cancer. Exon Publications. Published online 2022:101-110.

6. Kendrick P, Kelly YO, Baumann MM, et al. The burden of stomach cancer mortality by county, race, and ethnicity in the USA, 2000–2019: A systematic analysis of health disparities. The Lancet Regional Health–Americas. 2023;24:100547.

7. Thrift AP, El-Serag HB. Burden of gastric cancer. Clinical gastroenterology and hepatology. 2020;18(3):534-542.

8. Rupp SK, Stengel A. Influencing factors and effects of treatment on quality of life in patients with gastric cancer—a systematic review. Frontiers in Psychiatry. 2021;12:656929.

9. Kong F, Yan Z, Lan N, Wang P, Fan S, Yuan W. Construction and validation of gastric cancer diagnosis model based on machine learning. Exploration of Medicine. 2022;3(3):300-313.

10. Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. BMC gastroenterology. 2023;23(1):6.

11. Gilani N, Arabi Belaghi R, Aftabi Y, Faramarzi E, Edgünlü T, Somi MH. Identifying potential miRNA biomarkers for gastric cancer diagnosis using machine learning variable selection approach. Frontiers in genetics. 2022;12:779455.

12. Sundar R, Kumarakulasinghe NB, Chan YH, et al. Machine-learning model derived gene signature predictive of paclitaxel survival benefit in gastric cancer: Results from the randomised phase III SAMIT trial. Gut. 2022;71(4):676-685.

13. Kori M, Gov E. Bioinformatics prediction and machine learning on gene expression data identifies novel gene candidates in gastric cancer. Genes. 2022;13(12):2233.

14. Saha S., Vyas R, Computational analysis of gastric canceromics data to identify putative biomarkers. Current Topics in Medicinal Chemistry. 2024; 24(2):128-156.

15. Li G, Huo D, Guo N, et al. Integrating multiple machine learning algorithms for prognostic prediction of gastric cancer based on immune-related lncRNAs. Frontiers in Genetics. 2023;14:1106724.

16. Liu D, Wang X, Li L, et al. Machine learning-based model for the prognosis of postoperative gastric cancer. Cancer Management and Research. Published online 2022:135-155.

17. Zhou C, Hu J, Wang Y, et al. A machine learning-based predictor for the identification of the recurrence of patients with gastric cancer after operation. Scientific reports. 2021;11(1):1571.

18. Zhu SL, Dong J, Zhang C, Huang YB, Pan W. Application of machine learning in the diagnosis of gastric cancer based on noninvasive characteristics. Plos one. 2020;15(12):e0244869.

19. Tomczak K, Czerwińska P, Wiznerowicz M. Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia. 2015;2015(1):68-77.

20. Na S, Xumin L, Yong G, 2010 Third International Symposium on Intelligent Information Technology and Security Informatics. Research on k-means clustering algorithm: an improved k-means clustering algorithm. 2010:63-67.

21. Alexandropoulos SAN, Aridas CK, Kotsiantis SB, Vrahatis MN. Stacking strong ensembles of classifiers. In: Artificial Intelligence Applications and Innovations: 15th IFIP WG 12.5 International Conference, AIAI 2019, Hersonissos, Crete, Greece, May 24–26, 2019, Proceedings 15. Springer; 2019:545-556.

22. Russell J. Machine learning fairness in justice systems: Base rates, false positives, and false negatives. In: 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE; 2020:817-820.

23. Hoo Z. H, Candlish J, Teare D. What is an ROC curve? Emergency Medicine Journal. 2017;34(6):357-359.

Figure Legends

Figure 1: The methodology workflow analysis.

Figure 2: Confusion matrix for random forest classifiers.

Figure 3: Confusion matrix for SVM classifiers.

Figure 4: Confusion matrix for stacking classifiers.

Figure 5: Top 10 feature importances as per random forest classifier.

Figure 6: SHAP summary plot for feature importance.

Figure 7: Clustering comparison using K-means, hierarchical, and DBSCAN results.

Figure 8: Clustering on PCA-Reduced Data: KMeans, DBSCAN, and Hierarchical.

Figure 9: Performance metrics of supervised learning algorithms.

Figure 10: Performance metrics of existing approaches.

Figure 11: Training and validation accuracy graphs for, (A) Random forest, (B) Stacking classifier, and (C) Support vector machines.

Figure 12: Permutation test accuracy distribution for the best-performing classifier.

Figure 13: Feature stability analysis across cross-validation folds.

Figure 14: Macro-averaged Receiver Operating Characteristic (ROC) curves. (A) Random forest, (B) Stacking classifier, and (C) Support vector machines.

Table 1: Summary of approaches for gastric cancer diagnosis and prognosis: an overview of problem statements, datasets, methodologies, performance metrics, and limitations. GSE accession numbers can be located at https://www.ncbi.nlm.nih.gov/geo/.

Ref. (Year)	Problem Statement	Dataset	Solution	Performance Evaluation Metrics	Limitations
(9) (2022)	Construction of diagnosis model for gastric cancer.	TCGA, GEO	RF, NB, KNN	Accuracy: 92%, AUC: 97.22%, External validation AUC: 90.4%	Retrospective data from public sources limits external validity.
(2) (2018)	Non-invasive blood-based gene expression profile screening for gastric cancer diagnosis.	216 blood samples from gastric cancer patients and controls	LR model	Accuracy: 95%, AUC: 99%, Specificity: 96%	Small sample size affected ROC AUC values.
(10) (2023)	Lifestyle factor-based gastric cancer risk classification using ML algorithms.	2029 individuals from Iranian hospital	XGBoost, SVM, MLP, KNN	Best accuracy: XGBoost 83.4%, AUC: 84.9%, Specificity: 85.9%	Retrospective data reduces generalizability.
(11) (2022)	miRNA biomarkers for GC diagnosis, especially in the early stages.	GSE106817 dataset, 2,566 miRNAs	Boruta feature selection, SMOTE, DT, RF, ANN	hsa-miR-1343-3p identified as most important, mean importance: 10.81	Small sample size, lack of clinical validation.
(13) (2022)	Integrative multi-omics approach to identify gastric	GSE19826, GSE54129,	Multi-omics classification models	Accuracy: 64.6% to 47.7%	Limited by accuracy in distinguish-ing

Ref. (Year)	Problem Statement	Dataset	Solution	Performance Evaluation Metrics	Limitations
	cancer biomarkers.	GSE79973 datasets		depending on feature	live and dead specimens.
(14) (2023)	Identification of new biomarkers for gastric cancer using network analysis, ML, DL, and bioinformatics.	Gene expression, protein-protein interactions, mutations	Text mining, network analysis, structural bioinformatics	Identified 4 unique biomarker genes	No experimental validation or clinical trials performed.
(15) (2023)	Prognostic model for gastric cancer patients based on immune-related lncRNAs.	TCGA dataset	ImmLnc algorithm, Cox regression analysis	Developed 18-lncRNA signature	Lack of external validation.
(3) (2023)	Identification of miRNA biomarkers for GC detection.	TCGA dataset	SVM, RF, KNN	SVM, accuracy: 93%, AUC: 88.5%	Limited by the lack of early-stage diagnosis performance.
(16) (2022)	Predictive model for diagnosing gastric cancer based on non-invasive characteristics.	SEER dataset, 709 patients	GBDT model	AUC: 91%, Sensitivity: 87%, Specificity: 84%	Accuracy slightly lower: 83%, external validation required.

Table 2. Relevant gene extraction.

	TCGA-BR4279-01	TCGA-VQA94T-01	TCGA-BR8589-01	TCGA-BRA4IV-01	TCGA-VQA8E2-01	...	GTEX-WZTO2126-SM-4PQYW	GTEX-Q2AH1126-SM-48TZM	GTEX-P4QT1526-SM-3NMCT	TCGA-CG5734-11	GTEX-QDVJ1426-SM-48U1Y
0	10.4647	6.6146	10.4016	4.6554	9.8054	…	3.2506	5.4593	0	12.1036	5.1895
1	0	4.3985	2.4817	1.3475	3.4159	…	3.0795	5.4593	3.6917	1.9619	3.7267
2	0	1.915	0	2.7364	1.9642	…	1.6459	1.6888	0.8033	0	2.2239
3	9.7639	8.25	6.0767	10.3051	10.1232	…	10.5893	9.3984	11.9008	9.7787	9.4305
4	2.0828	0.7594	9.2494	7.6121	0.9762	…	5.3361	8.2453	0.8033	6.6976	9.1078
…	…	…	…	…	…	…	…	…	…	…	…
8858	10.6801	11.341	11.5792	11.8576	12.5407	…	11.7969	11.9503	12.0408	11.9479	11.9589
8859	4.749	4.3504	3.7667	5.3813	6.5375	…	7.7835	6.217	7.4168	0	6.3961
8860	10.4614	10.7943	10.6264	8.5261	11.0916	…	6.9724	8.2488	7.2643	8.5766	8.0781
8861	9.9374	10.9393	11.6924	7.97	11.3222	…	4.408	9.8712	4.4987	9.4988	9.5601
8862	2.0828	1.2543	4.4353	3.5624	8.6491	…	1.0459	0.8002	0	1.9619	1.7856

Table 3: The removal of non-informative columns.

Column Name	Description	Reason for Removal
Unnamed: 0	Auto-generated index column during CSV export	Not a gene expression feature

Table 4: Proposed Methodology.

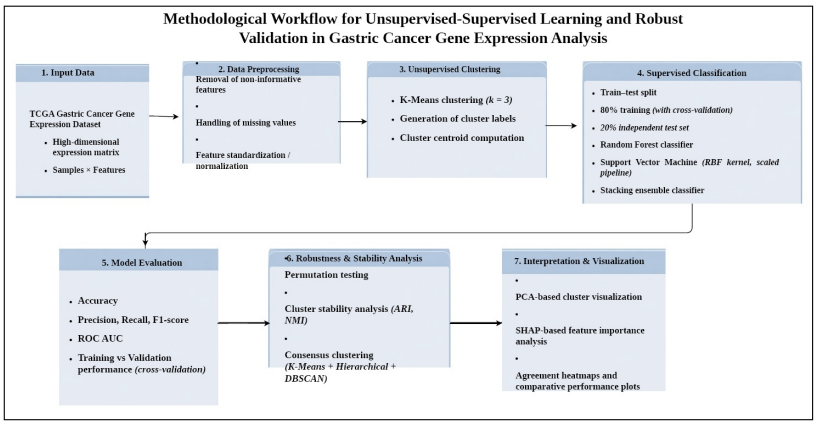
Step	Phase	Formulation & Logic
1	K-Means Clustering	Compute centroids μ_j where $C_j = \{x_i : \min_j \parallel x_i - \mu_j \parallel\}$: $$\mu_j = \frac{1}{\lvert C_j \rvert} \sum_{x_i \in C_j} x_i$$
2	Cluster Assignment	Assign label y_i for each data point x_i: $$y_i = \operatorname{argmin}_j \parallel x_i - \mu_j \parallel$$
3	Classifier Training	Optimize parameters $\hat{\theta}$ using loss function ℓ: $$\hat{\theta} = \operatorname{argmin}_\theta \left[\frac{1}{N} \sum_{i=1}^{N} \ell \left(h_\theta(x_i), y_i \right) \right]$$
4	Prediction	Predict class for new input x': $$\hat{y}' = h_{\hat{\theta}}(x')$$

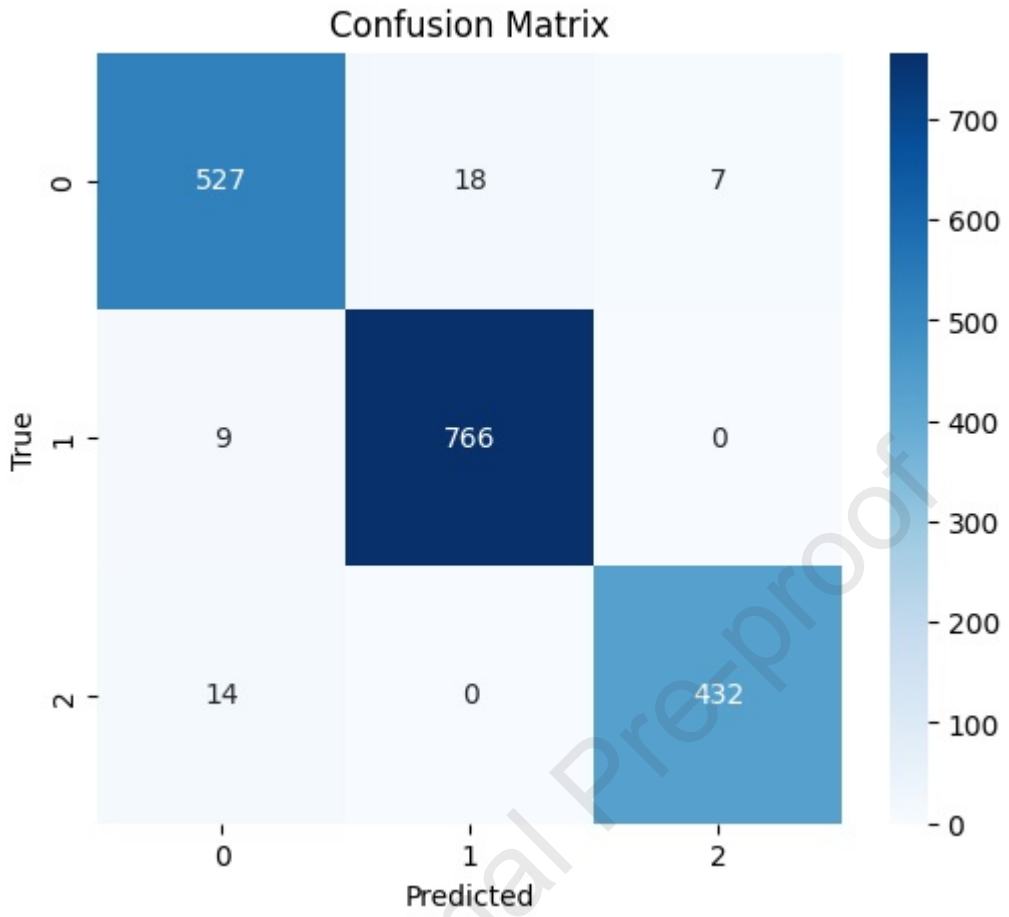
Table 5: Evaluation metrics formula.

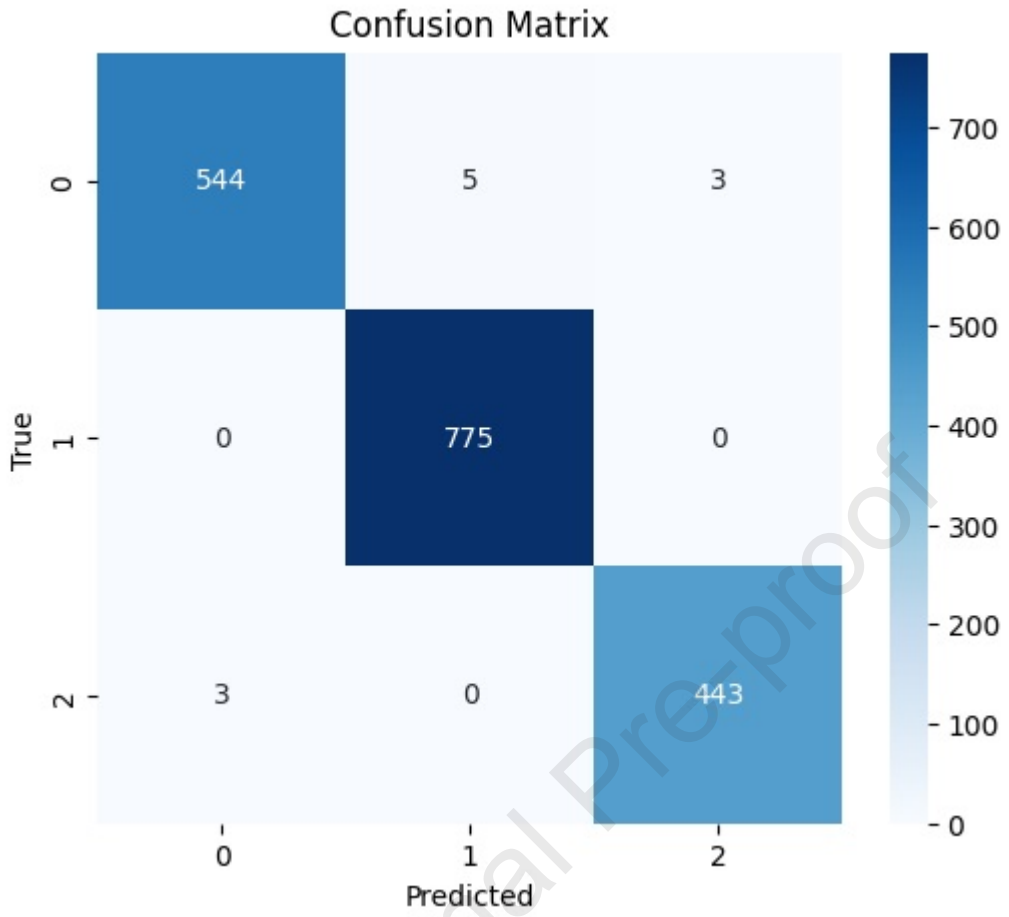
Metric	Formula
Accuracy	Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	Precision=$\frac{TP}{TP+FP}$
Recall	Recall = $\frac{TP}{TP+FN}$
F1-score	F1-score=2 $\times \frac{Precision \times Recall}{Precision+Recall}$

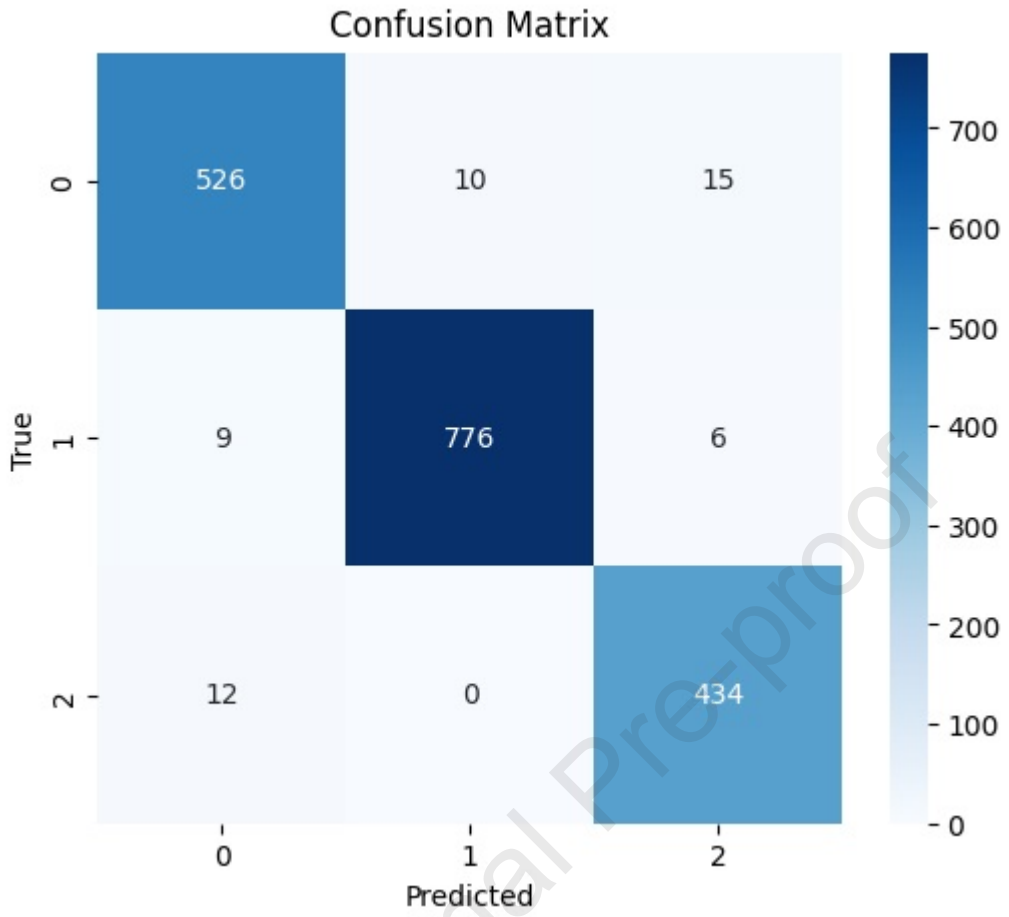
Table 6: Evaluation metrics for supervised learning algorithms.

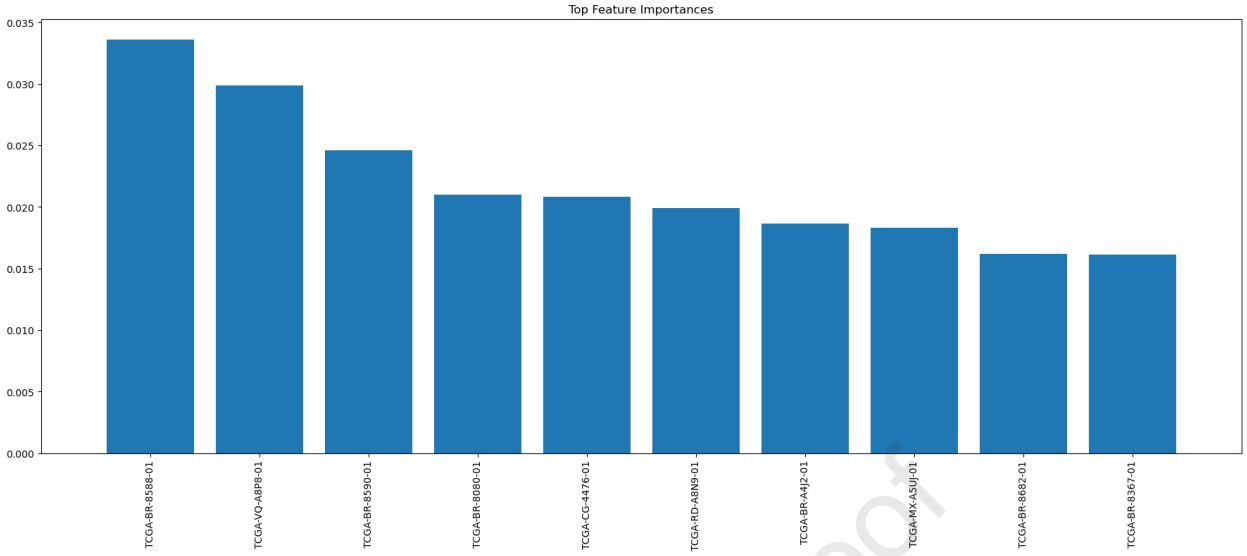
Algorithm	Accuracy	Precision	Recall	F1 score
RF	97.01%	95.23%	96.67%	95.95%
SVM	99.15%	98.76%	98.91%	98.83%
Stacking Classifier	98.87%	97.65%	97.89%	97.77%

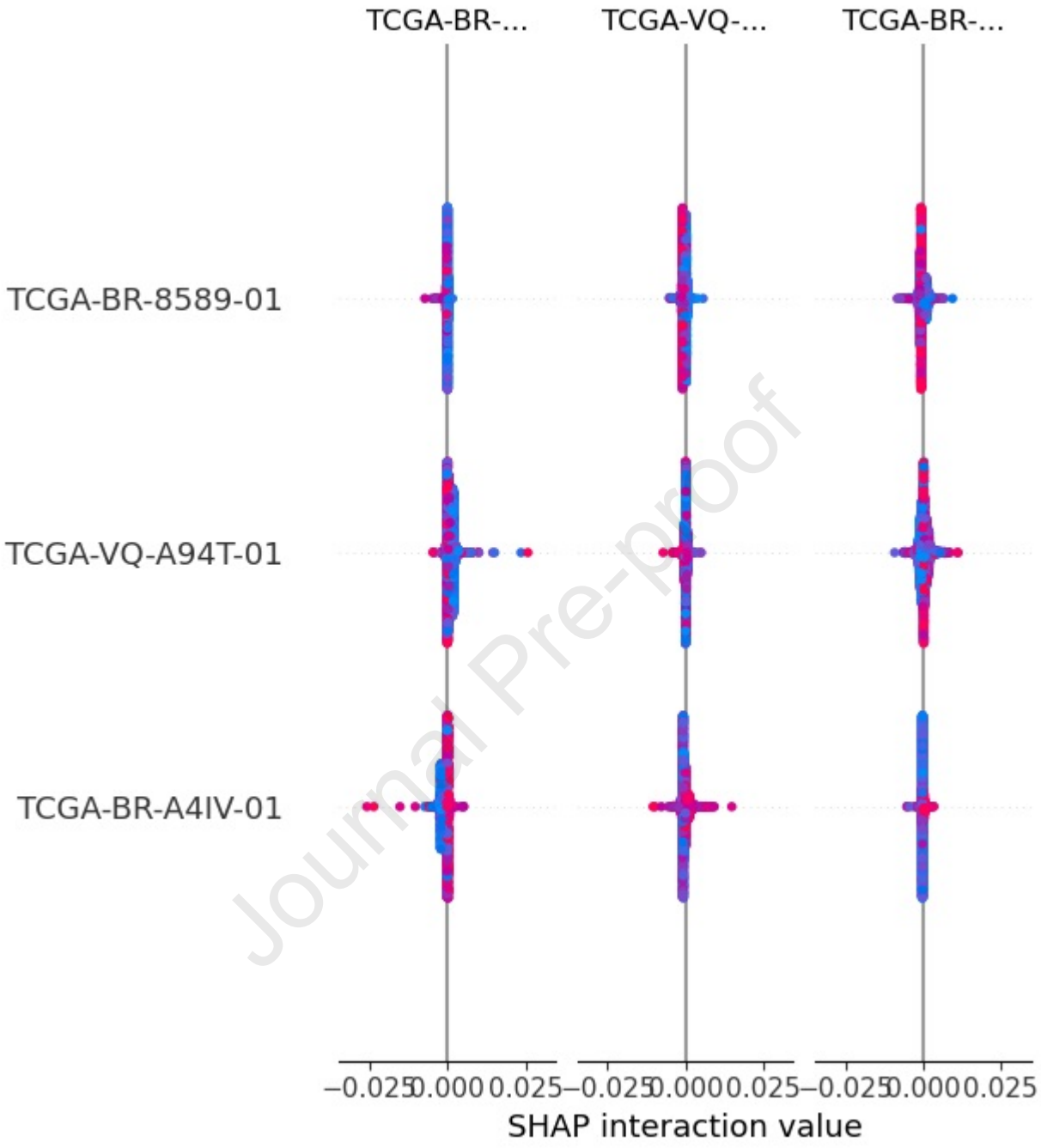




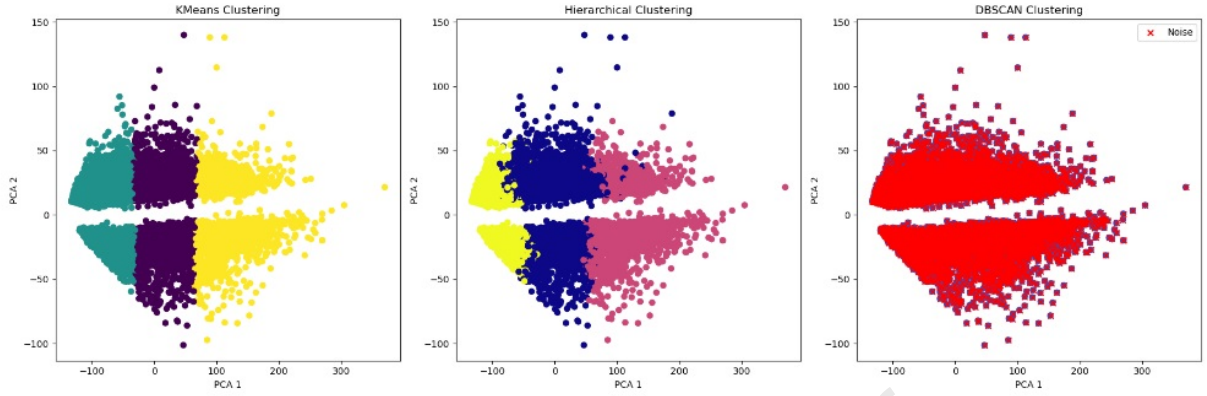


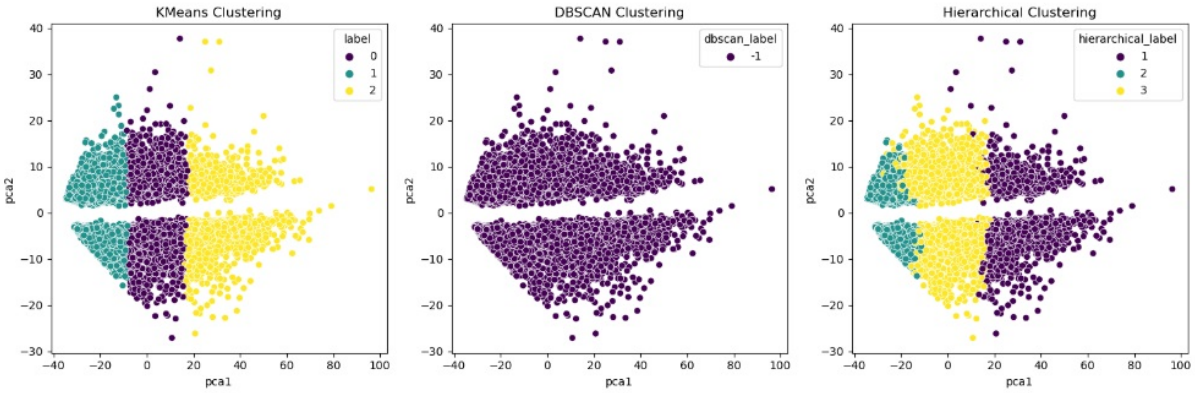


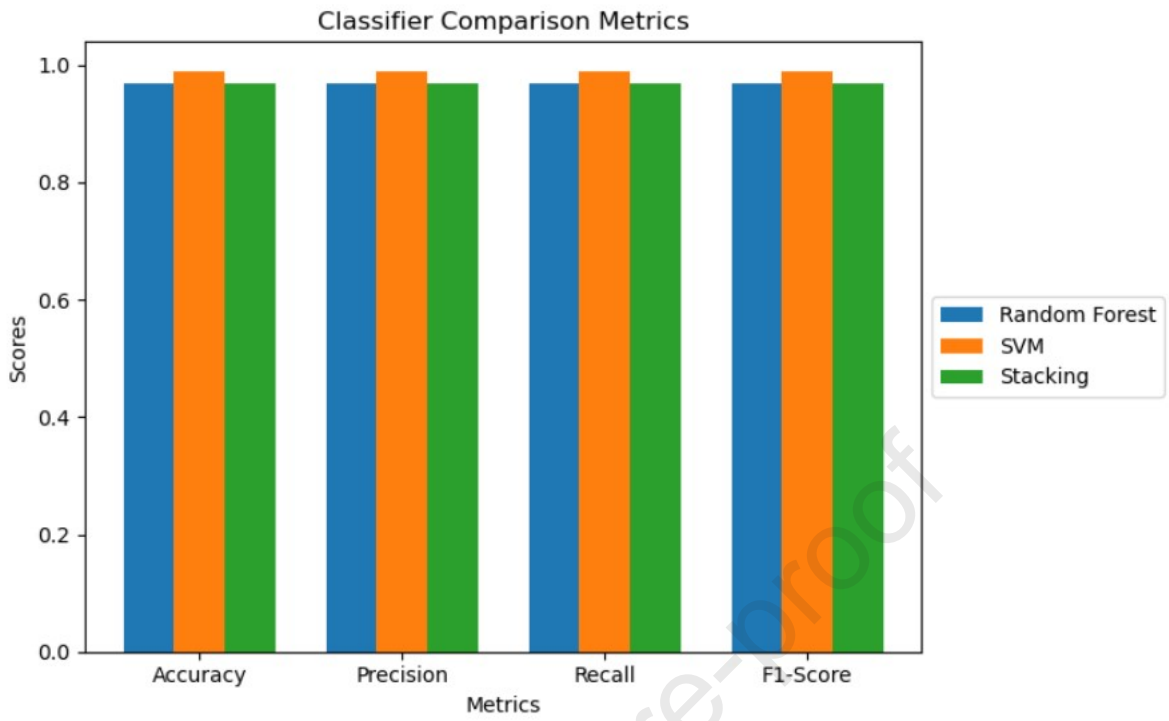


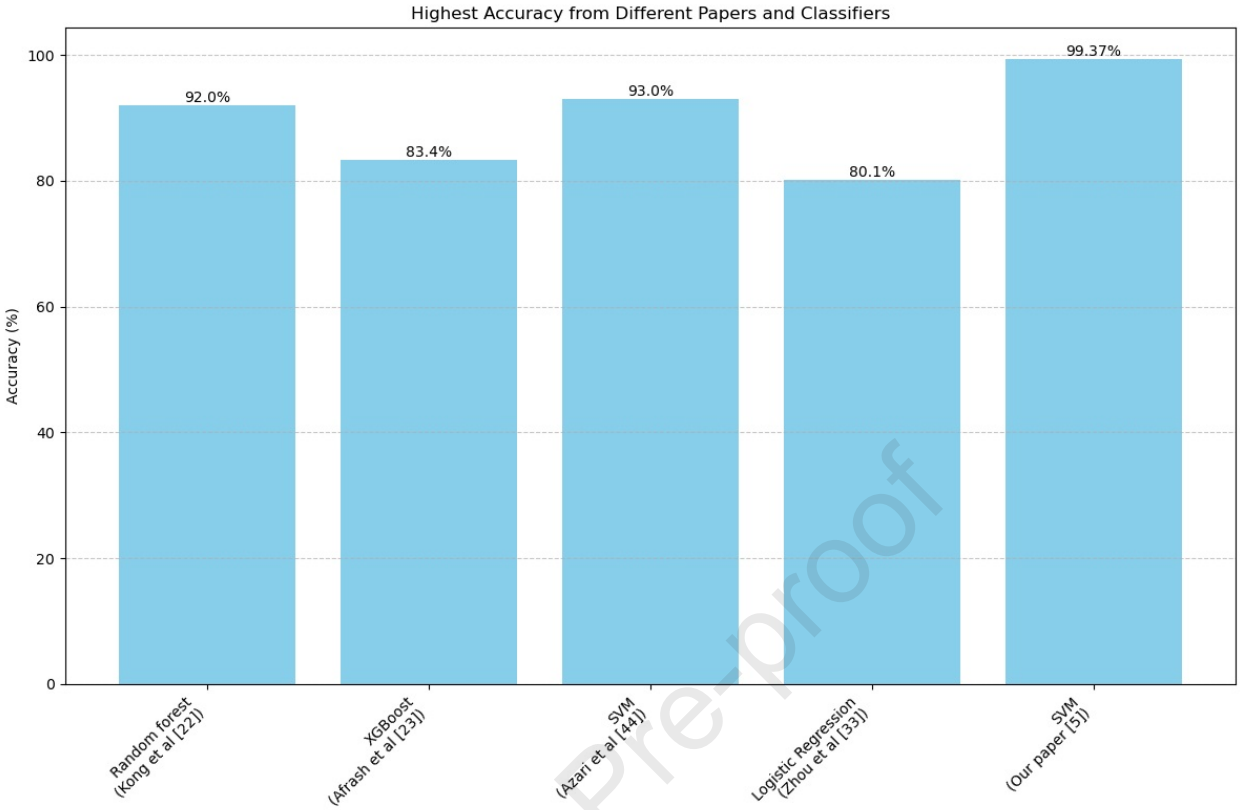


```
Unique labels found by DBSCAN: [-1]
```

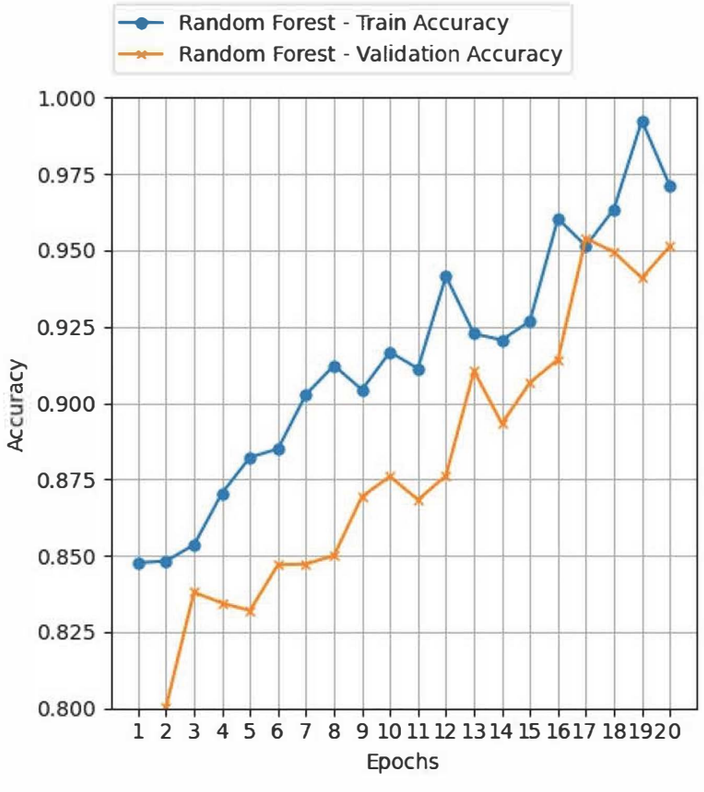




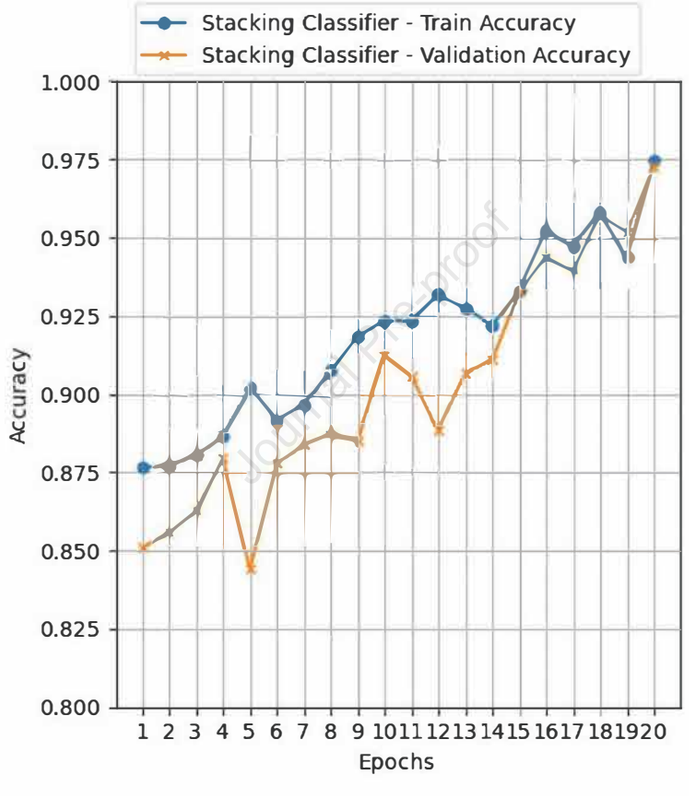




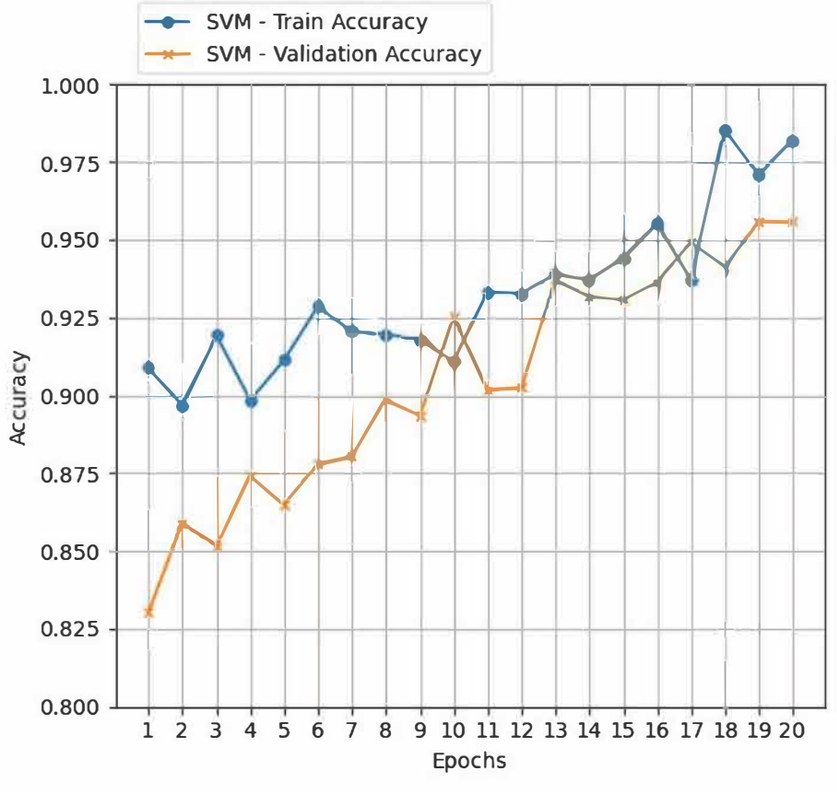
A

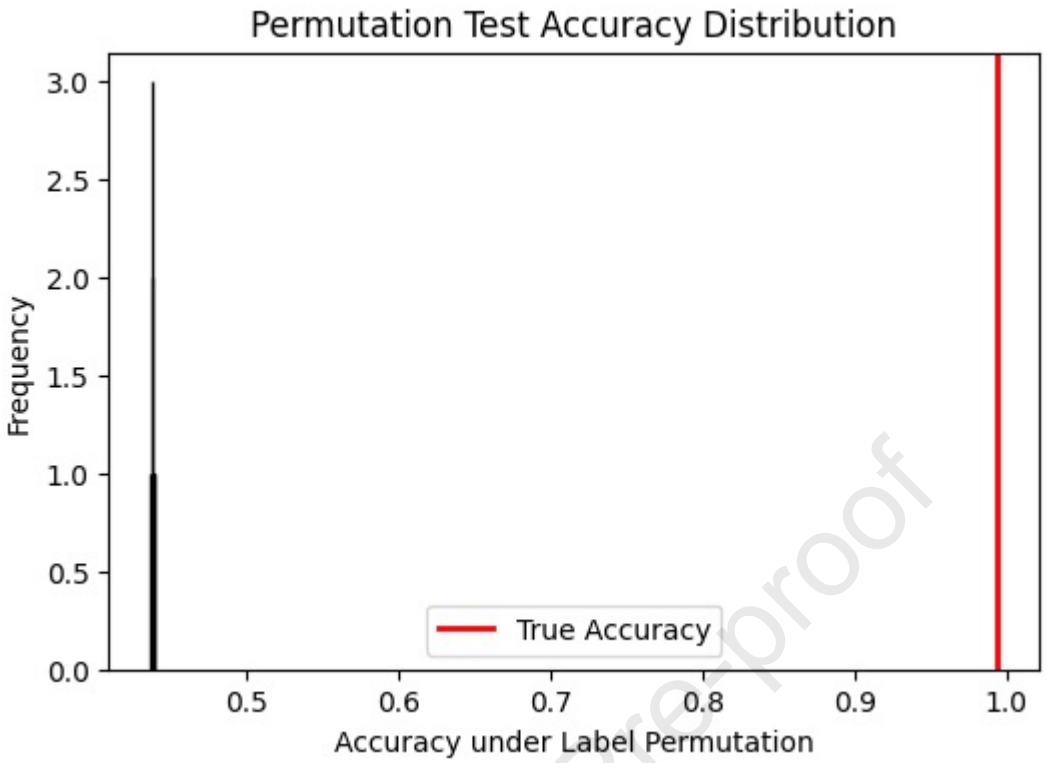


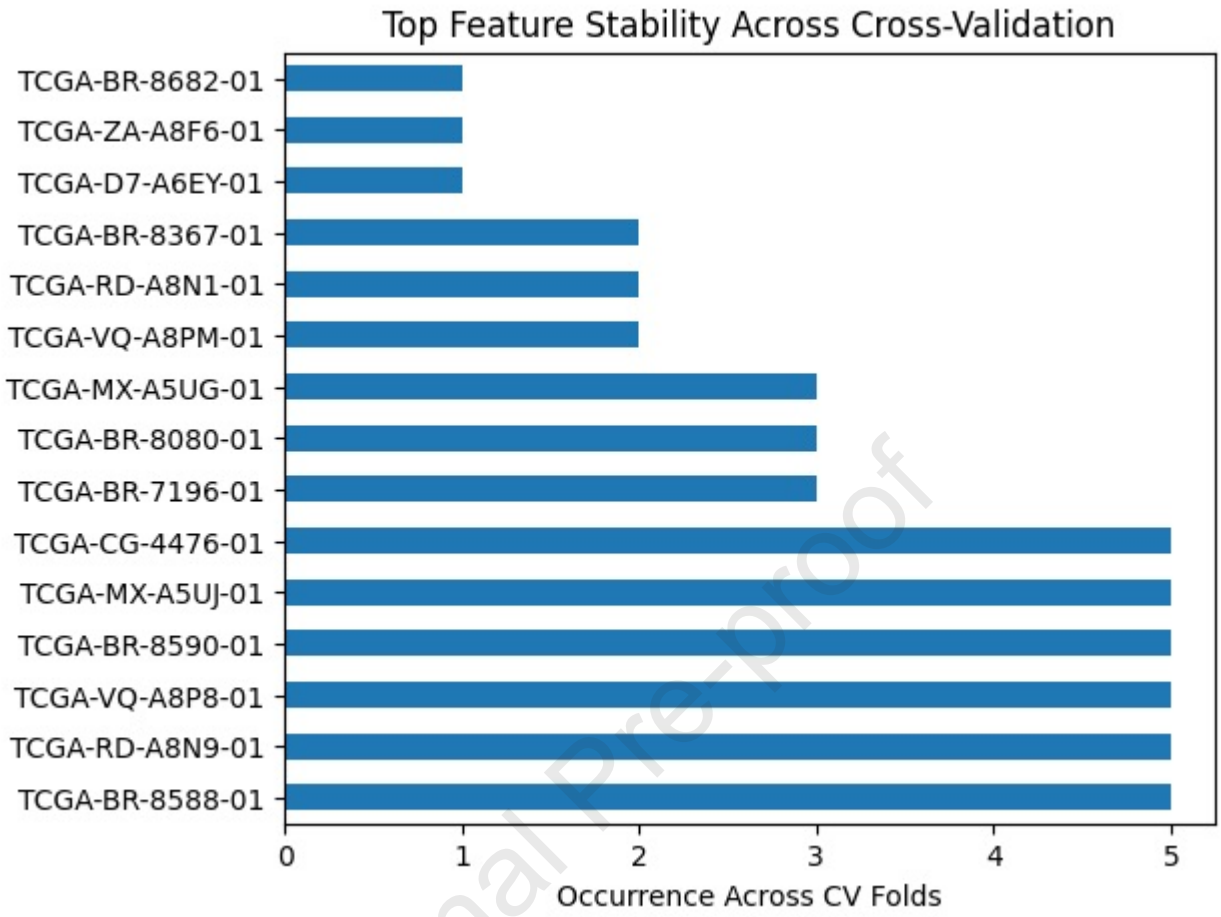
B



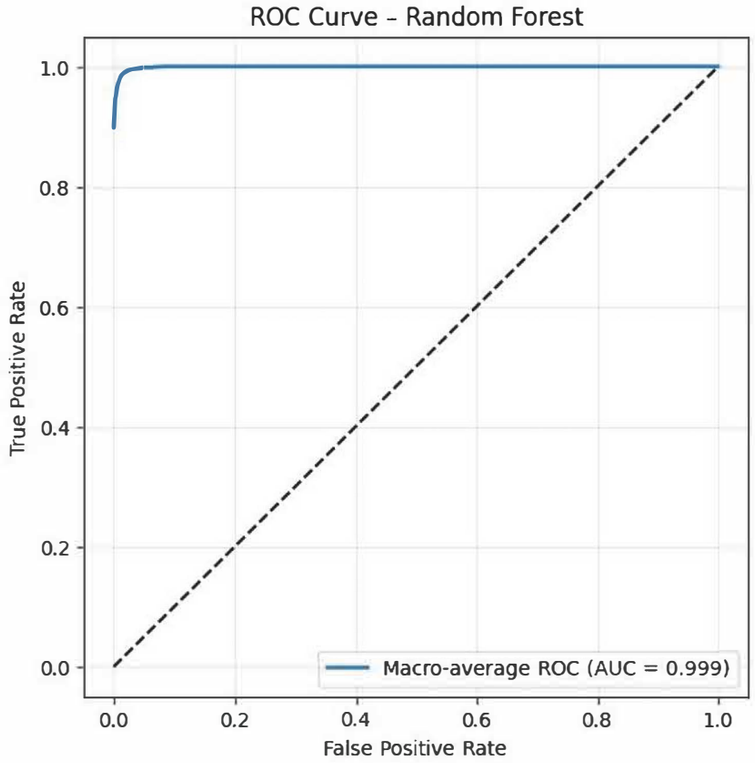
C



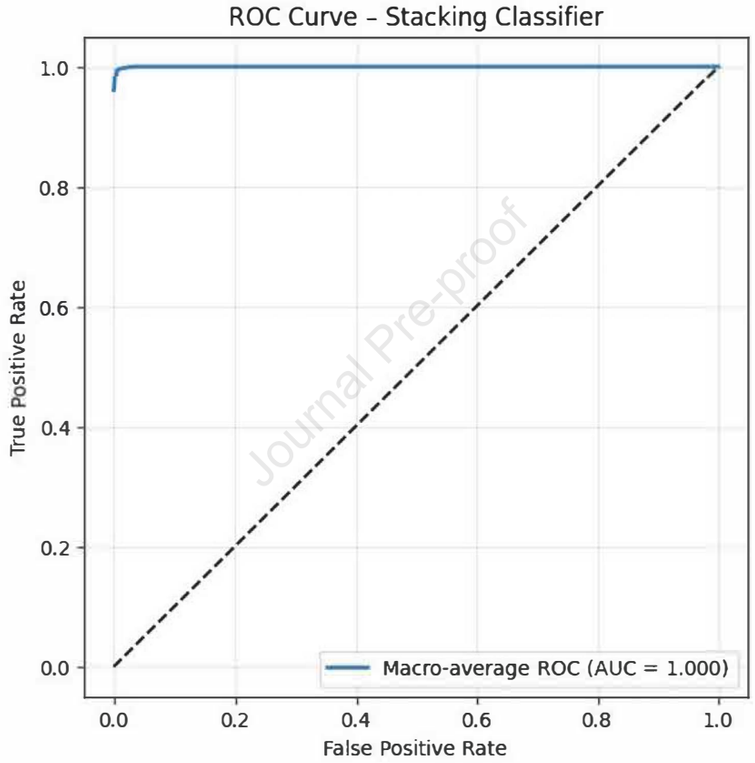




A



RF Macro AUC: 0.9992386493964617



Stacking Macro AUC: 0.9998685923069193

C