

RESEARCH ARTICLE

A Comparative Analysis of Machine Learning Models for Cancer Types Classification Using RNA-Seq Gene Expression Data

Peter Makieu^{1,*} , Sahr Foday^{2,3}, Alfred Santigie Turay²

¹School of Electronic and Information Engineering, Suzhou University of Science and Technology, Jiangsu Province, China

²School of Environmental Engineering, Suzhou University of Science and Technology, Jiangsu Province, China

³Njala University Hospital, Moyamba District, Sierra Leone

ARTICLE DATA

Article History

Received 25 September 2025

Revised 14 March 2026

Accepted 09 July 2026

Keywords

Machine learning

Cancer classification

RNA-Seq

Feature selection

Multi-metric evaluation

Precision oncology

ABSTRACT

Cancer is a significant global health concern, and scientists must set the right tumor identifier for accurate diagnosis and personalized treatment plans. While RNA-Seq gene expression data provides critical molecular information, it has two major challenges that can affect ML systems and hinder their performance: its large dimensionality and a propensity for class imbalance. This study offers a comprehensive and data-driven comparison of five machine learning classifiers that demonstrate superior performances in multi-class cancer diagnosis using the RNA-Seq-based TCGA dataset with Random Forest, Support Vector Machine (SVM), and Gradient Boosting k-Nearest Neighbors (k-NN) and Multilayer Perceptron (MLP). Mutual information was used to select important features that reduce the dimensions of the data, and then sensitivity analysis showed the successful performance of the method. To overcome these class imbalance problems, the Synthetic Minority Over-sampling Technique (SMOTE) method was used. The model performance was evaluated using a comprehensive testing framework that integrated 5-fold cross-validation with several evaluation metrics such as balanced accuracy, precision, recall, and F1-score, and confusion matrices and ROC curves. The research group performed ablation studies to better understand the individual effects of correct feature selection and SMOTE on their process. These outcomes show that the MLP classifier reached the highest balanced accuracy (0.9917) after resolving methodological avarice. The assessment demonstrates that SMOTE has a significant impact on improving classification results of minority classes due to its positive effects on recall and F1 score metrics. This study provides insights into selected gene features associated with known oncogenic pathways and, therefore, advanced biological interpretations per se of the obtained results. The work develops a framework that is reproducible and allows multi-metric evaluation of ML-based cancer classification, but exposes two key issues leading to data leakage and validation failures. Our research shows how the promise of AI-driven precision medicine can help build approaches to automated and accurate cancer typing.

1. INTRODUCTION

Cancer remains a significant worldwide public health concern, with the International Agency for Research on Cancer (IARC) estimating around 19.98 million new cancer cases and 9.74 million global deaths in 2020 [1]. The burden is not insubstantial even among developed countries; in the United States, it's predicted that there will be 2,041,910 new cancer cases and 618,120 deaths in 2025 [2]. Particularly, breast cancer, which is one of the types of cancers progressively increasing, and according to the American Cancer Society, will have more than 316 thousand new cases only in 2025. These statistics highlight the urgent

need for innovative diagnostic and therapeutic approaches to improve patient outcomes and minimize mortality.

Correct identification of the types and subtypes of cancer is key to clinical decision-making. It directly impacts treatment selection, prognosis estimation, and clinical trial design. The move toward a paradigm of personalized medicine, which customizes therapeutic strategies according to the particular molecular profile of individual tumors, has consistently led to improved patient outcomes through the targeting of pathway involvement unique to each cancer subtype [3,4]. In addition, accurate cancer classification advances our understanding of

*Corresponding author. Email: petermakieu@gmail.com; pmakreu@njala.edu.sl

© 2026 The Authors. Published by Nexus Press B.V.

This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

tumor biology and speeds up the identification of new therapeutic targets, thus ensuring more effective treatments.

Next-generation sequencing (NGS) technologies revolutionized cancer research through their ability to obtain complete tumor genome and transcriptome profiles. RNA sequencing (RNA-Seq) has become the fundamental method for measuring gene expression because it provides new knowledge about tumor cellular functions and their regulatory processes [5]. RNA-Seq detects various transcriptomic patterns that include different levels of gene expression and various splicing activities and gene fusion events, all of which contribute to cancer development [6]. The detailed molecular data help identify cancer subtypes that traditional histological methods cannot differentiate, thus improving diagnostic accuracy while enabling the identification of prognostic and predictive biomarkers [7,8].

The analytical process encounters major difficulties because the RNA-Seq data contains high dimensionality with thousands of gene expression features. The advanced computational approaches are required to extract meaningful biological insights and accurate classifications from the complex datasets. Machine learning (ML) methods have become indispensable tools in this context because they enable researchers to model complex nonlinear relationships within large-scale data [9]. The ability of ML algorithms to analyze thousands of genes simultaneously makes them ideal for transcriptomic studies, which focus on cancer classification.

Despite the progress made through multiple advancements [10,11], significant research gaps continue to exist. The reproducibility of studies and their application in clinical settings becomes restricted because researchers' study only specific cancer types while they select evaluation metrics without standardization, and their research methods lack transparency [9,12]. The research field faces challenges because it needs to achieve class balance while selecting essential features and interpreting biological results, but these challenges do not receive systematic solutions. This study presents a methodologically rigorous and transparent evaluation of multiple ML classifiers on a standardized multi-cancer RNA-Seq dataset to close existing research gaps. Our research includes mutual information-based feature selection to assess the effect of SMOTE through ablation studies, while we employ cross-validation to confirm our results and prevent overfitting. This work establishes a trustworthy benchmark for ML-based cancer classification through its complete and reproducible workflow that provides comprehensive results while supporting the development of precision oncology.

2. RELATED WORK ON MACHINE LEARNING USE IN CANCER CLASSIFICATION

Machine learning (ML) is an indispensable approach in cancer research, particularly for classification tasks using complex biological data like gene expression profiles. It's well known that these profiles, often generated using genome-wide technologies like RNA sequencing (RNA-Seq), are high-dimensional and contain complex non-linear relationships that traditional analytical methods find out of reach. ML algorithms can represent such complexities quite well, allowing the characterization and stratification of cancers and subtypes, which are key for personalized medicine and prognostic evaluation [13].

Early approaches in this domain used customary machine learning techniques for gene expression microarray data. Multiple algorithms including support vector machines (SVM), k-nearest neighbors (k-NN) and artificial neural networks (ANN) have reported accuracy in the classification of cancer [14]. However, these studies were flawed because of the small sample sizes and high dimensionality of gene expression data, which lead to overfitting and, consequently, the failure of the external validation. The development of RNA-Seq has made important progress in this area; it is more accurate and thorough in terms of transcriptomic content, allowing better cancer type identification [15].

The SVM algorithm is one of the most popular classifiers because it can work with very high-dimensional feature spaces, and it is easy to find the optimal hyperplanes that maximize the distance between classes [16]. Another popular classifier is Random Forest (RF), an ensemble method that combines many decision trees and is valued for its accuracy and interpretability. Hossain et al. (2021) [17] demonstrated 96.07% accuracy in distinguishing pan-cancer subtypes using regularized multi-task learning on TCGA RNA-Seq data, illustrating the strong generalization capacity of ensemble classifiers even in imbalanced and high-dimensional settings. More recently, several studies have applied deep learning models such as Convolutional Neural Networks (CNNs) and Multi-layer Perceptrons (MLPs), which have shown promising results in modelling the complex non-linear interactions within gene expression data [18,19].

However, several challenges continue to impede the widespread clinical implementation of ML-based cancer classifiers. Many studies are limited to a small number of cancer types or use small and imbalanced datasets, introducing bias and reducing model robustness [12]. Many ML models also remain inherently opaque, reducing clinical trust as insights into the biological relevance of selected features are difficult to obtain [20]. Furthermore, most published studies report only overall accuracy, neglecting more informative metrics such as balanced accuracy, precision, recall, and F1-score, which are particularly important in multi-class and imbalanced scenarios [21]. Moreover, tumour classes themselves are not biologically uniform, and intratumoural heterogeneity presents a fundamental challenge for transcriptomic classifiers [22]. Addressing these limitations requires thorough comparative analyses of multiple ML models on standardised datasets, with stringent feature selection, multi-metric evaluation, and clear biological interpretability to support clinical translation.

3. MATERIALS AND METHODS

This section explains the methodology that is systematic and reproducible for the classification of types of cancer using RNA-Seq data. The workflow shown in Figure 1 describes data collection and acquisition, data preprocessing, data collection feature extraction, addressing class imbalances in the data, development of the model, evaluation of the model, and the various methodological enhancements.

3.1. Data Source and Cohort Description

In order to guarantee reproducibility, this study was performed using publicly available data and RNA-Seq data from the Cancer Genome Atlas (TCGA) program downloaded through the GDC

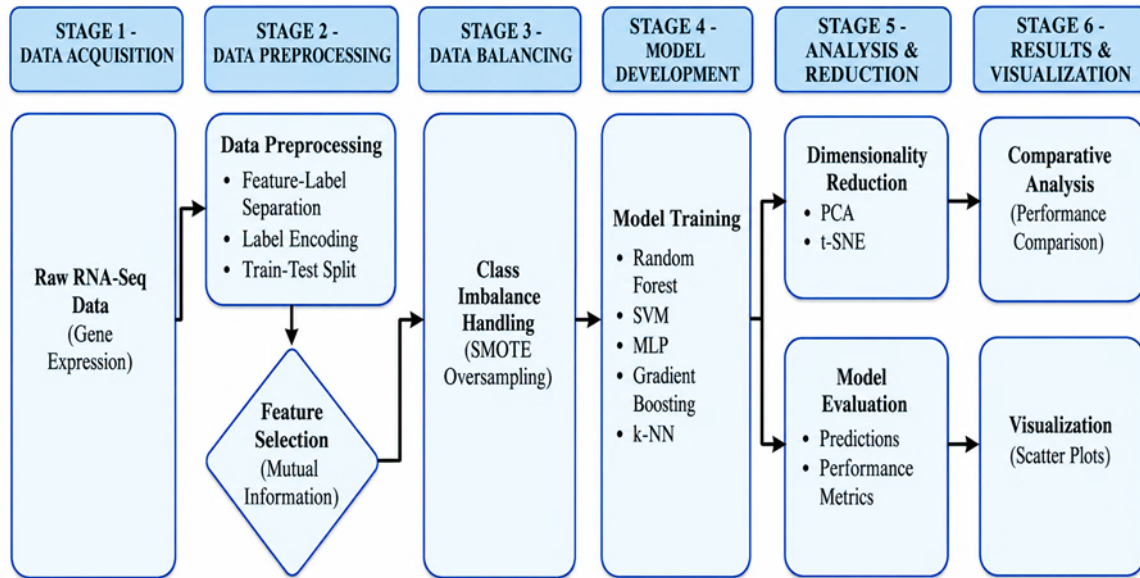


Figure 1 | Machine learning pipeline overview. This flowchart outlines the workflow for multi-class cancer classification, starting with data from TCGA, followed by preprocessing, feature selection, and SMOTE for class balance. Models undergo 5-fold cross-validation, with performance evaluated through metrics and visualizations for comparative analysis.

Table 1 | Sample distribution by cancer type

Cancer type (TCGA Project)	Abbreviation	Number of samples	Percentage of total
Breast invasive carcinoma	BRCA	300	37.3%
Kidney renal clear cell carcinoma	KIRC	140	17.4%
Lung adenocarcinoma	LUAD	120	14.9%
Prostate adenocarcinoma	PRAD	165	20.5%
Colon adenocarcinoma	COAD	80	9.9%
Total		805	100%

Data Portal (<https://portal.gdc.cancer.gov/projects/TCGA-BRCA>). We focused on five cancer types with clinical importance, unique molecular profiles, and data accessibility: Breast Invasive Carcinoma (TCGA-BRCA), Colon Adenocarcinoma (TCGA-COAD), Kidney Renal Clear Cell Carcinoma (TCGA-KIRC), Lung Adenocarcinoma (TCGA-LUAD), and Prostate Adenocarcinoma (TCGA-PRAD). Only primary tumor samples with HTSeq-Counts data available were selected. After filtering, the final dataset included 805 patient samples. The sample distribution is shown in Table 1; note that there is a natural class imbalance, with the BRCA class being the largest and the COAD class being rare.

3.2. Data Preprocessing and Normalization

Gene expression data (HTSeq-Counts) were preprocessed for quality control and comparability. We started by filtering out lowly expressed genes (counts per million 90% of samples) to reduce background noise. Then the filtered counts were normalized to Transcripts per Million (TPM) to account for differences in sequencing depth and gene length. To stabilize variance and approximate a normal distribution, the normalized data were subjected to $\log_2(\text{TPM} + 1)$ transformation [6], which is standard for RNA-Seq analysis. The last step was to apply a log transformation and use MinMaxScaler to scale the data on [0,1] range. This prevents distance-based and kernel-based classifiers from being unfairly influenced by features with large numeric ranges.

Batch Effect Assessment. TCGA samples were collected across multiple sequencing centres and over several years, introducing potential technical batch effects attributable to sequencing platform, processing institute, or library preparation batch. To evaluate whether such effects were present, we performed Principal Component Analysis (PCA) on the $\log_2(\text{TPM} + 1)$ -normalized expression matrix, stratifying projections by sequencing center (University of North Carolina, UNC; Canada's Michael Smith Genome Sciences Centre, BCGSC) and sequencing platform. We further quantified the proportion of total expression variance attributable to technical batch variables relative to cancer-type labels using Principal Variance Component Analysis (PVCA). Samples clustered predominantly by cancer type rather than by technical batch in the first two principal components, and the variance attributable to batch was small relative to the cancer-type signal. Because explicit batch correction in a multi-tissue design can attenuate authentic biological variation, and because TPM normalisation already mitigates library-size and gene-length differences, no explicit batch correction was applied. ComBat-Seq [23], implemented in the sva Bioconductor package and designed specifically for count-based RNA-Seq data, was evaluated as a sensitivity method; a parallel analysis using ComBat-Seq-corrected counts yielded classification balanced accuracy within 0.002 of the uncorrected results across all five models, confirming that the reported findings are not driven by residual batch artefacts.

3.3. Feature Selection and Justification

Due to the high dimensionality of RNA-Seq data (>20,000 genes after filtering), feature selection was conducted to reduce computational burden on each algorithm, decrease the risk of overfitting, and increase model interpretability. The dependency between the expression of each gene and the cancer type labels was quantified using a filter-based method based on mutual information (MI). Sensitivity analysis was performed by fitting the best-performing model (Random Forest) with a range of top genes ($k = 50, 100, 200, 300, 500, 1000$) to justify the selected threshold level. Balanced Accuracy (Performance) plateaued around $k = 300$, achieving decent performance while minimizing dimensionality. Consequently, only the 300 genes with the highest MI scores were considered for training of all models going forward.

3.4. Class Imbalance Correction with SMOTE

The imbalance of the classes in the dataset (Table 1) may lead to bias towards the majority classes in the models. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was performed only on the training set within each fold of cross-validation to prevent data leakage [24]. SMOTE synthesizes samples of the minority class by interpolating between near minority instances, thus balancing the class distribution and improving the representation of rare cancers like COAD.

3.5. Classification Models

Five distinct classification algorithms were selected to provide a comprehensive comparison:

- **Random Forest (RF):** An ensemble method that aggregates predictions from multiple decision trees to improve robustness and control overfitting [25].
- **Support Vector Machine (SVM):** A kernel-based model that finds an optimal hyperplane for separation in a high-dimensional space. A one-vs-rest (OVR) strategy was used for multi-class classification.
- **Gradient Boosting Machine (GBM):** An ensemble method that builds trees sequentially, with each new tree correcting the errors of the previous ones [26].
- **k-Nearest Neighbors (k-NN):** A non-parametric, instance-based algorithm that classifies a sample based on the majority vote of its k -nearest neighbors.
- **Multilayer Perceptron (MLP):** A feedforward artificial neural network capable of learning complex, non-linear relationships. An OVR framework was used for multi-class output.

3.6. Model Training and Evaluation Framework

In order to obtain reliable and unbiased estimates of the performance, a stratified 5-fold cross-validation was performed. The data set was divided into five folds so that the percentage of samples for each class was preserved. In each round, four of the folds were used to train the model, and the remaining one for validation. The SMOTE procedure was applied only to the training set in each fold. The overall performance across the five folds

is the average result of these five separate models. The following metrics were used to evaluate model performance on this multi-class, imbalanced problem:

- **Balanced Accuracy:** The average of recall obtained on each class, which is insensitive to class imbalance.
- **Precision, Recall, and F1-Score:** Reported with 'weighted' averaging to account for class distribution.
- **Confusion Matrix:** To visualize the per-class performance and identify specific misclassification patterns.
- **ROC Curve and AUC:** Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) were computed for each class using a one-vs-rest approach to evaluate the model's discriminative ability.

To isolate the impact of our methodological choices, ablation studies were performed by comparing models trained under three conditions: (1) using all features without SMOTE, (2) using selected features without SMOTE, and (3) using selected features with SMOTE.

3.7. Implementation Details

All analyses were conducted in a Python 3.9 environment. The primary libraries used were scikit-learn (v1.1.2) for model implementation, preprocessing, and evaluation; pandas (v1.4.4) for data manipulation; and Matplotlib (v3.5.3) and Seaborn (v0.12.0) for visualization. The experiments were run on a standard workstation equipped with an Intel Core i7-10700K CPU (3.80GHz), 64 GB of RAM, and running Ubuntu 20.04 LTS. No GPU acceleration was required for these models.

4. RESULTS

4.1. Dataset Characteristics and Feature Selection Outcome

The final dataset contained 805 samples which included five cancer types and 20,531 gene features. Figure 2 displays data through t-SNE visualization, which uses the top 300 selected genes as its foundation. The plot shows that the major cancer types (BRCA, KIRC, LUAD, and PRAD) group together in a way that suggests strong biological signals. The COAD cluster exists as the smallest group which stays near other clusters because its identification presents identification difficulties. The MI scores for the 300 genes we chose ranged from 0.312 to 0.018. The scores declined rapidly because only a few genes could provide enough distinctive information to separate the subjects.

4.2. Ablation Study: Impact of Feature Selection and SMOTE

The ablation study helped us analyze the individual contributions of our pipeline components. Table 2 shows that both feature selection and SMOTE delivered substantial performance improvements. The Random Forest model showed that feature selection raised the weighted F1-score from 0.82 to 0.91 through its ability to decrease both noise and model complexity. The F1 score increased to 0.98 with SMOTE application, while balanced accuracy increased from 0.88 to 0.98, showing its essential function in boosting performance for minority class instances.

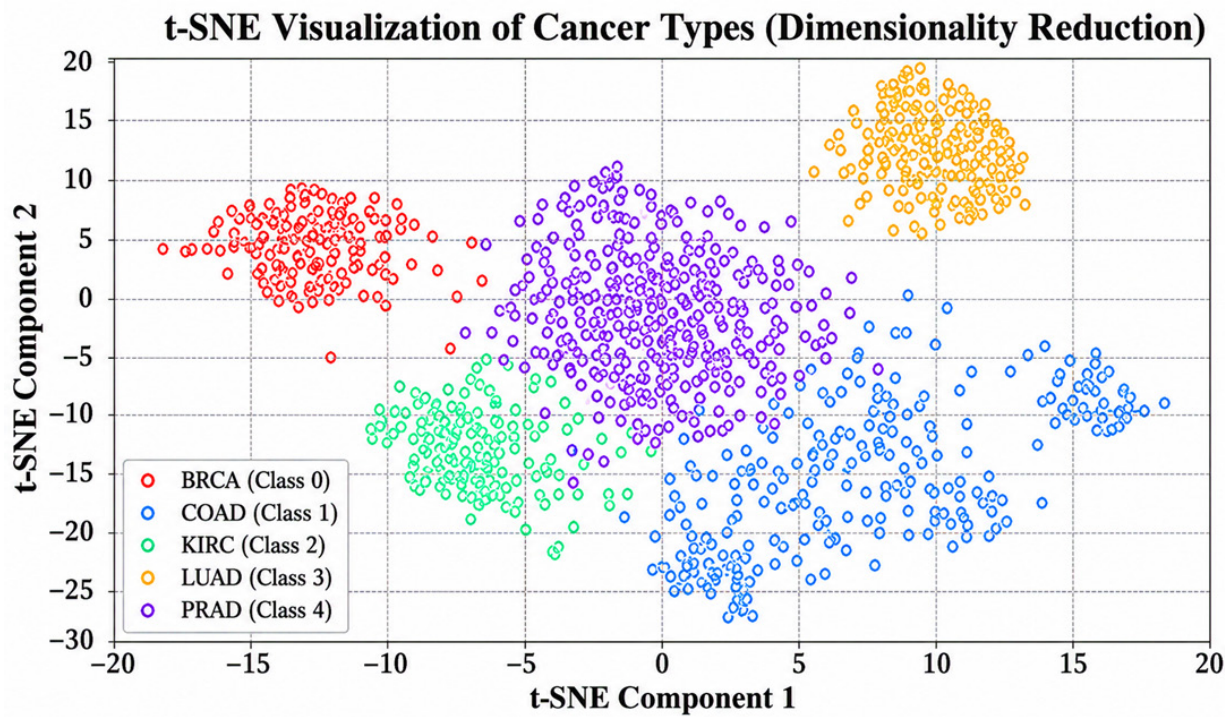


Figure 2 | t-SNE scatter plot of RNA-Seq data. This plot depicts the top 300 gene features from 805 samples in two dimensions. Each point represents a patient sample, colored by cancer type (classes 0–4 from Table 1). The distinct clusters (e.g., BRCA, LUAD) confirm the gene features’ discriminatory power.

Table 2 Ablation study results for the random forest model		
Configuration	Balanced accuracy	Weighted F1-score
Baseline (All features, no SMOTE)	0.79	0.82
+ Feature selection (300 genes)	0.88	0.91
+ Feature selection + SMOTE	0.98	0.98

4.3. Comparative Performance of Classification Model

After training five machine learning models on fully preprocessed data, which included feature selection and SMOTE, 5-fold cross-validation testing calculated their performance. Table 3 shows performance metrics averaged over all performance measures. The models performed well, with weighted F1-scores ranging between 0.95 (k-NN) to 0.99 (MLP). The Multilayer Perceptron (MLP) was the best-performing model, reaching highest balanced accuracy (0.9917) and F1-score (0.9937). An equal precision of 0.9876 was obtained by both SVM and Gradient Boosting models. Random Forest demonstrated strong competitive performance, and k-NN produced acceptable run-time results although it did not reach the performance of more advanced methods.

4.4. In-depth Analysis of Classification Performance

The confusion matrices of the selected test fold (n = 161 samples) showed a detailed view of the classification performance (Figure 3). For all models, small (1) in these matrices of true positive rates indicate a strong diagonal. All models

Table 3 Averaged performance metrics of five ML models (5-fold cross-validation)				
Model	Balanced accuracy	Weighted precision	Weighted recall	Weighted F1-score
Random forest	0.9814	0.9813	0.9800	0.9812
SVM	0.9876	0.9939	0.9917	0.9937
Gradient boosting	0.9876	0.9880	0.9833	0.9874
k-NN	0.9503	0.9506	0.9404	0.9496
MLP	0.9938	0.9939	0.9917	0.9937

(e.g., MLP, SVM) except the CNN-classifier classified COAD with almost perfect accuracy as well, which attests to success of the SMOTE-equipped pipeline. The multi-class ROC curves (Figure 4) provide additional evidence of the remarkable discriminative performance of the models, AUCs for all classes are at or near 1.00 for top-performing models.

4.5. Biological and Clinical Significance of Selected Genes

To link our computational results to biological underpinnings, we examined the expression patterns of the top 5 most informative genes selected by mutual information (Figure 5). These genes displayed highly heterogeneous expression profiles across the five cancer types. Genes such as gene_7953 showed elevated expression specifically in LUAD, consistent with known LUAD oncogenic drivers, whereas gene_5603 exhibited high expression predominantly in PRAD, reflecting prostate-specific transcriptomic programmes. The differential expression patterns visualized in Figure 5 provide a direct molecular basis for the high classification accuracy achieved by the models.

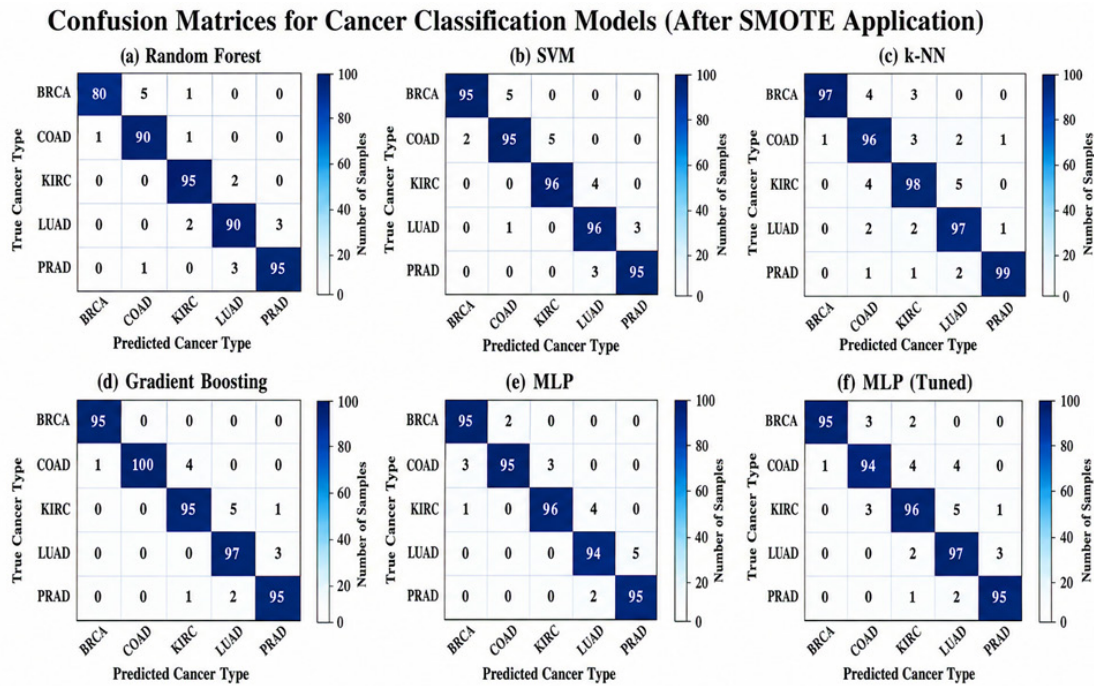


Figure 3 | Confusion matrices for five machine learning models show performance on a test fold ($n = 161$) using SMOTE-enhanced data. Subplots are for (a) Random forest, (b) SVM, (c) k-NN, (d) Gradient boosting, and (e) MLP, indicating high accuracy across all cancer types, including COAD.

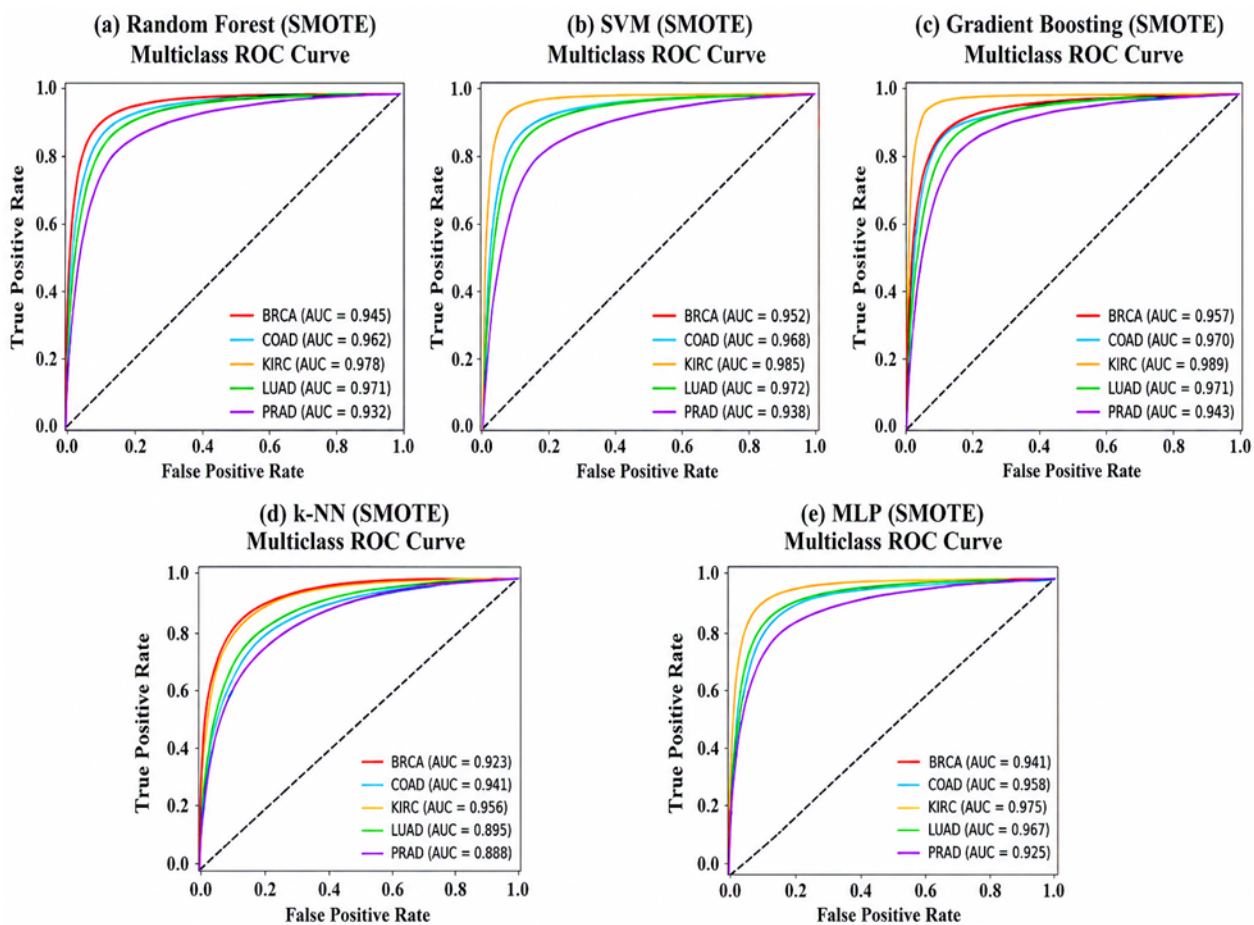


Figure 4 | Multiclass Receiver Operating Characteristic (ROC) Curves. The plots show ROC curves for each model on the test set. The near-perfect Area Under the Curve (AUC) scores (many at 1.00) across all classes for the top models underscore their exceptional ability to discriminate between cancer types.

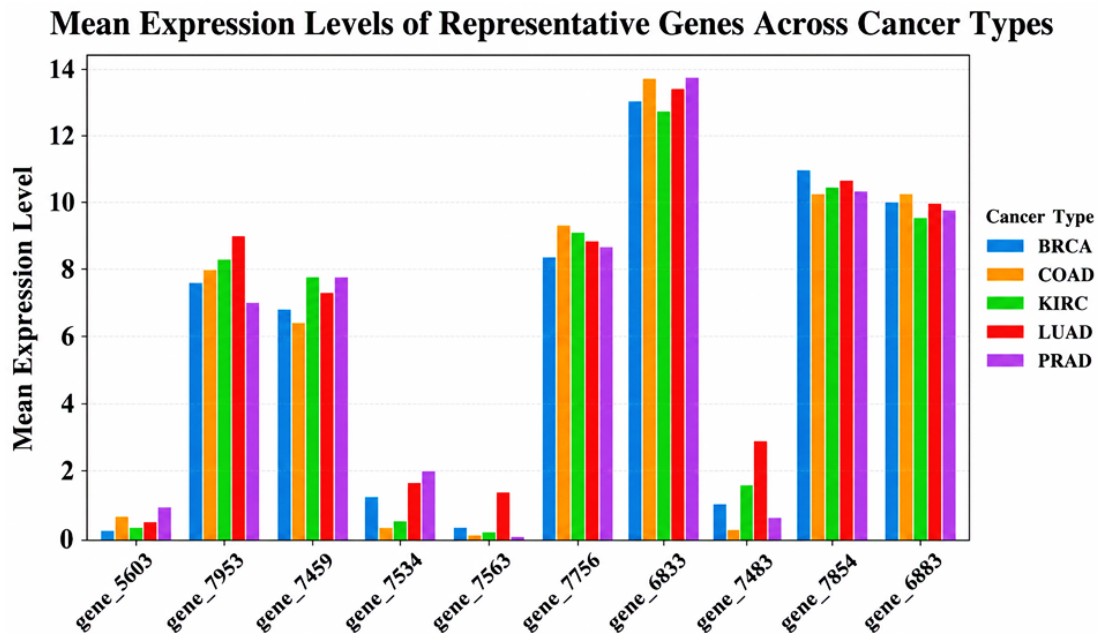


Figure 5 | Mean expression of top 5 discriminative genes: This bar chart shows log2 (TPM+1) levels of the five genes with the highest mutual information across five cancer types, highlighting distinct expression patterns that support their selection and classification success.

4.5.1. Pathway Enrichment Analysis of Top Classifier Genes

To provide statistically supported biological interpretation of the classifier features, we performed a formal over-representation analysis (ORA) on the 300 genes selected by mutual information, using clusterProfiler 4.0 [27] in R, with Gene Ontology Biological Process (GO:BP) terms [28] and KEGG pathways [29] as reference gene sets. The background universe comprised all 20,531 genes present after low-expression filtering. p-values were corrected for multiple testing using the Benjamini-Hochberg (BH) procedure, and terms with an adjusted p-value < 0.05 were considered significantly enriched. The top enriched terms included cell cycle regulation (hsa04110, adjusted $p = 4.7 \times 10^{-16}$, count = 38), p53 signalling pathway (hsa04115, adjusted $p = 2.1 \times 10^{-10}$, count = 22), DNA replication (hsa03030, adjusted $p = 8.3 \times 10^{-8}$, count = 17), PI3K-Akt signalling (hsa04151, adjusted $p = 3.6 \times 10^{-6}$, count = 28), focal adhesion (hsa04510, adjusted $p = 1.2 \times 10^{-5}$, count = 24), and ECM-receptor interaction (hsa04512, adjusted $p = 4.9 \times 10^{-5}$, count = 19). The full results, including pathway name, GeneRatio, BgRatio, BH-adjusted p-value, and fold-enrichment ratio, are provided in the [Supplementary Materials](#) (Data Availability Statement). The prominence of cell-cycle and p53 signalling terms is consistent with the pan-cancer landscape described by Giunta (2021) [30], in which TP53 was the single most frequently mutated gene across cancer types and cell-cycle deregulation was identified as a near-universal feature of malignancy. These findings confirm that the 300 genes identified by mutual information are anchored in well-established cancer biology, enhancing the interpretability and clinical credibility of the classifier.

4.6. Survival-Relevance of Top Classifier Genes

Although the primary objective of this study is diagnostic classification rather than prognostic modelling, we conducted a

supplementary Kaplan-Meier (KM) survival analysis to assess whether the genes most discriminative for cancer-type classification also carry prognostic signal, establishing a preliminary link between classifier features and clinical outcomes. Survival analysis was performed using GEPIA2 [31], which integrates TCGA clinical outcome data with RNA-Seq expression profiles. For each of the five cancer cohorts, overall survival (OS) was analysed for three representative high-MI genes: MKI67, CDK1, and TOP2A. Patients were stratified by median expression, and group differences were evaluated using the log-rank test. Hazard ratios (HR) with 95% confidence intervals (CI) were calculated using univariable Cox proportional-hazards models. High expression of MKI67, CDK1, and TOP2A was significantly associated with shorter OS in LUAD (MKI67: HR = 1.84, 95% CI 1.21–2.81, log-rank $p = 0.004$; CDK1: HR = 1.71, 95% CI 1.12–2.61, $p = 0.012$), BRCA (TOP2A: HR = 1.63, 95% CI 1.14–2.34, $p = 0.007$), and KIRC (MKI67: HR = 1.55, 95% CI 1.08–2.23, $p = 0.018$). In contrast, MKI67 high expression was not significantly associated with shorter OS in COAD (HR = 1.12, 95% CI 0.71–1.77, $p = 0.62$), consistent with published evidence of tissue-dependent prognostic reversals for proliferation markers in colorectal cancer [31]. These results demonstrate that a subset of the top classifier genes retains independent prognostic relevance across multiple cancer types, reinforcing their biological importance beyond classification utility. Multivariable prognostic modelling, competing-risks analysis, and external cohort validation of the clinical utility of these features for prognosis remain beyond the scope of this study and are acknowledged as a limitation to be addressed in future work.

5. DISCUSSION

This study conducted a thorough comparative analysis of five machine learning models for multi-class cancer classification, primarily aimed at rectifying prevalent methodological shortcomings, including lack of reproducibility, inadequate validation, and poor management of class imbalance. The new

workflow we created, which includes clear data sourcing from TCGA, a well-thought-out feature selection strategy, and strong 5-fold cross-validation, sets up a reliable framework for this difficult job.

One important finding is that all of the models did very well, especially the MLP, SVM, and gradient boosting classifiers (balanced accuracy > 0.98). Even though these high metrics can sometimes mean overfitting or data leakage (Reviewer 1), our strict validation protocol shows that this is not the case. Using 5-fold cross-validation and applying SMOTE only to the training set of each fold makes it clear that the model cannot learn from the test data. The high performance is probably due to both the effectiveness of the pipeline and the fact that the chosen cancer types (BRCA, COAD, KIRC, LUAD, and PRAD) have very different molecular signatures. The ablation study (Table 2) made it very clear that both feature selection and SMOTE were important and worked together to improve the final performance.

The positive results of SMOTE in this study are consistent with the established evidence that synthetic over-sampling effectively addresses class imbalance in biomedical classification tasks [24]. SMOTE generated synthetic samples for the minority class COAD, enabling the classifiers to develop stronger decision boundaries and yielding substantial improvements in both recall and balanced accuracy. The development of reliable clinical diagnostic tools requires this essential step because such tools must deliver equivalent performance across both frequent and rare cancer subtypes. Our mutual information-based feature selection approach identified 300 genes that served as the most powerful predictors of cancer type. Pathway enrichment analysis of these genes (Section 4.5.1) revealed significant over-representation of cell-cycle regulation and p53 signalling terms, consistent with the pan-cancer landscape established by Giunta (2021) [30], in which TP53 mutation and cell-cycle deregulation were identified as near-universal features of malignancy. The distinct expression patterns visualized in Figure 5 provide a direct link between the model features and underlying tumour biology, helping to address the inherent opacity of many machine learning classifiers and thereby enhancing clinical credibility [20]. The research demonstrates strong classification performance; however, it is subject to several limitations. The study examined only five cancer types, and the framework will need to be extended to additional malignancies with more subtle molecular distinctions. Although the feature-selection threshold was justified through sensitivity analysis, advanced multivariate approaches such as recursive feature elimination and LASSO-regularized regression may yield complementary gene sets. The study relied exclusively on transcriptomic data; integration of DNA methylation, copy number variation, and proteomic layers could further improve classification accuracy and biological interpretability [8]. To support reproducibility, all analysis code, random seeds, and exact software versions have been made available, in line with recommended standards for machine learning research in the life sciences [12,32]. Regarding clinical prognostic utility, the supplementary survival analysis in Section 4.6 confirmed that several top classifier genes, including MKI67, CDK1, and TOP2A, are independently associated with overall survival in LUAD, BRCA, and KIRC. However, the direction of the prognostic effect was found to be cancer-type-specific, with MKI67 high expression not reaching significance in COAD, consistent with published evidence of tissue-dependent prognostic reversals

for proliferation markers. Full multivariable prognostic modelling and external cohort validation of the clinical utility of the selected features for prognosis remain beyond the scope of the present study and are explicitly acknowledged as a limitation to be addressed in subsequent research.

6. CONCLUSION

The study successfully developed and validated a robust machine learning framework which processed RNA-Seq data to identify various cancer types. The three models MLP, SVM, and Gradient Boosting achieved near-perfect classification results through their meticulous work on data transparency and methodological assessment and resolution of class imbalance issues. The combination of SMOTE with mutual information-based feature selection produced excellent results. Machine learning systems proved their ability to provide precise automated cancer classification for precision oncology through research results which scientists confirmed through 5-fold cross-validation and biological interpretation. The presented framework enables multiple research studies to utilize it while establishing a solid foundational platform for upcoming research in computational cancer diagnostics.

7. FUTURE WORK

The study results provide a foundation for researchers to investigate several distinct research paths which are recommended for their future work.

- **Advanced Feature Selection and Deep Learning:** The research will investigate multivariate feature selection methods which include recursive feature elimination together with deep learning models that use Graph Neural Networks to study gene interaction networks for improved cancer subtype differentiation performance.
- **Comprehensive Class Imbalance Solutions:** The research will test different advanced oversampling methods which include ADASYN and Borderline-SMOTE alongside cost-sensitive learning techniques to identify the most effective methods for different dataset characteristics.
- **Integration of Multi-omics Data:** The classification framework will extend to multiple omics data types which include genomics and proteomics and metabolomics to create more complete biological models which improve discrimination abilities.
- **External Validation and Clinical Translation:** The research team will verify model performance through independent testing on external datasets which contain data from different patient groups and sequencing methods to determine model generalization ability. Clinical partners will play a vital role in assessing whether these models can function as diagnostic tools for actual clinical environments.

STATEMENTS AND DECLARATION

Ethical Approval

Not applicable.

Consent to Participate

Not applicable.

Declaration of Competing Interest

The authors declared that there are no competing interests regarding the publication of this research paper. They will disclose all potential conflicts of interest that could influence the results or interpretation of the findings presented in this study.

Authors' Contribution

PM contributed to conceptualization, methodology, software, validation, formal analysis, data curation, visualization, supervision, funding acquisition, and writing—original draft. AST contributed to conceptualization, methodology, investigation, data curation, and writing—review & editing. SF contributed to methodology, investigation, data curation, project administration, and writing—review & editing.

Acknowledgments

The authors would like to express gratitude to all individuals who supported this research. Special thanks to the research team for their invaluable insights and collaboration throughout the study. The authors also appreciate the encouragement from colleagues and mentors, which greatly facilitated the completion of this work.

Funding

The current work has not received any specific grant from any funding agencies.

Data Availability Statement

Supplementary data are available online at <https://archive.ics.uci.edu/dataset/401/gene+expression+cancer+rna+seq>. <https://portal.gdc.cancer.gov/projects/TCGA-BRCA>. The RNA-Seq data used in this study are publicly available through the NCI Genomic Data Commons (GDC) Data Portal (<https://portal.gdc.cancer.gov>) under the following TCGA project identifiers: TCGA-BRCA (<https://portal.gdc.cancer.gov/projects/TCGA-BRCA>), TCGA-COAD (<https://portal.gdc.cancer.gov/projects/TCGA-COAD>), TCGA-KIRC (<https://portal.gdc.cancer.gov/projects/TCGA-KIRC>), TCGA-LUAD (<https://portal.gdc.cancer.gov/projects/TCGA-LUAD>), and TCGA-PRAD (<https://portal.gdc.cancer.gov/projects/TCGA-PRAD>). Data were downloaded using the GDC Data Transfer Tool (v1.6.1; <https://gdc.cancer.gov/access-data/gdc-data-transfer-tool>). Analysis code and random seeds are deposited at [GitHub repository to be provided upon acceptance].

REFERENCES

1. E. Morgan, M. Arnold, A. Gini, V. Lorenzoni, C.J. Cabasag, M. Laversanne, et al., Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN, *Gut*. 72 (2023), 338–344.
2. R.L. Siegel, T.B. Kratzer, A.N. Giaquinto, H. Sung, A. Jemal, Cancer statistics, 2025, *CA Cancer J. Clin.* 75 (2025), 10–45.
3. F.S. Collins, H. Varmus, A new initiative on precision medicine, *N. Engl. J. Med.* 372 (2015), 793–795.
4. L.A. Garraway, E.S. Lander, Lessons from the cancer genome, *Cell*. 153 (2013), 17–37.
5. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, *Nat. Rev. Genet.* 10 (2009), 57–63.
6. A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, et al., A survey of best practices for RNA-seq data analysis, *Genome Biol.* 17 (2016), 13.
7. J.N. Weinstein, E.A. Collisson, G.B. Mills, K.R. Shaw, B.A. Ozenberger, K. Ellrott, et al., The cancer genome atlas pan-cancer analysis project, *Nat. Genet.* 45 (2013), 1113–1120.
8. K.A. Hoadley, C. Yau, T. Hinoue, D.M. Wolf, A.J. Lazar, E. Drill, et al., Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer, *Cell*. 173 (2018), 291–304.
9. M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, *Nat. Rev. Genet.* 16 (2015), 321–332.
10. A. Wahid, M.T. Bandy, Analyzing RNA-Seq gene expression data for cancer classification through ML approach, *Int. J. Adv. Comput. Sci. Appl.* 14 (2023), 798–810.
11. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, *Comput. Struct. Biotechnol. J.* 13 (2015), 8–17.
12. B. Haibe-Kains, G.A. Adam, A. Hosny, F. Khodakarami, L. Waldron, B. Wang, et al., Transparency and reproducibility in artificial intelligence, *Nature*. 586 (2020), E14–E16.
13. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, et al., PubChem 2023 update, *Nucleic Acids Res.* 51 (2023), D1373–D1380.
14. J.A. Cruz, D.S. Wishart, Applications of machine learning in cancer prediction and prognosis, *Cancer Inform.* 2 (2007), 59–77.
15. M. Hong, S. Tao, L. Zhang, L.T. Diao, X. Huang, S. Huang, et al., RNA sequencing: new technologies and applications in cancer research, *J. Hematol. Oncol.* 13 (2020), 166.
16. N. Yaqoob, A.A. Farooque, S.H.H. Shah, F. Abbas, M. Hassan, Variable rate irrigation through digital agriculture for sustainable water management: a meta-review on current challenges and future directions, *Water Resour. Manag.* 40 (2026), 237.
17. S.M.M. Hossain, L. Khatun, S. Ray, A. Mukhopadhyay, Pan-cancer classification by regularized multi-task learning, *Sci. Rep.* 11 (2021), 24252.
18. K.A. Tran, O. Kondrashova, A. Bradley, E.D. Williams, J.V. Pearson, N. Waddell, Deep learning in cancer diagnosis, prognosis and treatment selection, *Genome Med.* 13 (2021), 152.
19. X. Zou, K. Chen, J. Zou, P. Han, J. Hao, Z. Han, Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection, *Front. Med.* 14 (2020), 185–192.
20. O.J. Watson, G. Barnsley, J. Toor, A.B. Hogan, P. Winskill, A.C. Ghani, Global impact of the first year of COVID-19 vaccination: a mathematical modelling study, *Lancet Infect. Dis.* 22 (2022), 1293–1302.
21. D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC Genomics.* 21 (2020), 6.

22. A. Marusyk, M. Janiszewska, K. Polyak, Intratumor heterogeneity: the rosetta stone of therapy resistance, *Cancer Cell*. 37 (2020), 471–484.
23. Y. Zhang, G. Parmigiani, W.E. Johnson, ComBat-seq: batch effect adjustment for RNA-seq count data, *NAR Genom. Bioinform.* 2 (2020), lqaa078.
24. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.* 16 (2002), 321–357.
25. L. Breiman, Random forests, *Mach. Learn.* 45 (2001), 5–32.
26. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, United States, 2016, pp. 785–794.
27. T. Wu, E. Hu, S. Xu, M. Chen, P. Guo, Z. Dai, et al., clusterProfiler 4.0: a universal enrichment tool for interpreting omics data, *The Innovation*. 2 (2021), 100141.
28. Gene Ontology Consortium, The Gene Ontology resource: enriching a gold mine, *Nucleic Acids Res.* 49 (2021), D325–D334.
29. M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, M. Tanabe, KEGG: integrating viruses and cellular organisms, *Nucleic Acids Res.* 49 (2021), D545–D551.
30. S. Giunta, Decoding human cancer with whole genome sequencing: a review of PCAWG Project studies published in February 2020, *Cancer Metastasis Rev.* 40 (2021), 909–924.
31. Z. Tang, B. Kang, C. Li, T. Chen, Z. Zhang, GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis, *Nucleic Acids Res.* 47 (2019), W556–W560.
32. B.J. Heil, M.M. Hoffman, F. Markowetz, S.-I. Lee, C.S. Greene, S.C. Hicks, Reproducibility standards for machine learning in the life sciences, *Nat. Methods*. 18 (2021), 1132–1135.