





Article

Optimizing Model Performance and Interpretability: Application to Biological Data Classification

Zhenyu Huang ^{1,2} , Xuechen Mu ^{2,3}, Yangkun Cao ⁴ , Qiufen Chen ², Siyu Qiao ², Bocheng Shi ^{2,4} , Gangyi Xiao ¹, Yan Wang ^{1,*}  and Ying Xu ^{2,*}

¹ College of Computer Science and Technology, Jilin University, Changchun 130012, China; zhenyuh19@mails.jlu.edu.cn (Z.H.); xiaogy19@mails.jlu.edu.cn (G.X.)

² Systems Biology Lab for Metabolic Reprogramming, Department of Human Genetics and Cell Biology, School of Medicine, Southern University of Science and Technology, Shenzhen 518055, China; m250921296@gmail.com (X.M.); chenqf829@foxmail.com (Q.C.); shibc22@mails.jlu.edu.cn (B.S.)

³ School of Mathematics, Jilin University, Changchun 130012, China

⁴ School of Artificial Intelligence, Jilin University, Changchun 130012, China; caoyk20@mails.jlu.edu.cn

* Correspondence: wy6868@jlu.edu.cn (Y.W.); xuy9@sustech.edu.cn (Y.X.)

Abstract: This study introduces a novel framework that simultaneously addresses the challenges of performance accuracy and result interpretability in transcriptomic-data-based classification. **Background/objectives:** In biological data classification, it is challenging to achieve both high performance accuracy and interpretability at the same time. This study presents a framework to address both challenges in transcriptomic-data-based classification. The goal is to select features, models, and a meta-voting classifier that optimizes both classification performance and interpretability. **Methods:** The framework consists of a four-step feature selection process: (1) the identification of metabolic pathways whose enzyme-gene expressions discriminate samples with different labels, aiding interpretability; (2) the selection of pathways whose expression variance is largely captured by the first principal component of the gene expression matrix; (3) the selection of minimal sets of genes, whose collective discerning power covers 95% of the pathway-based discerning power; and (4) the introduction of adversarial samples to identify and filter genes sensitive to such samples. Additionally, adversarial samples are used to select the optimal classification model, and a meta-voting classifier is constructed based on the optimized model results. **Results:** The framework applied to two cancer classification problems showed that in the binary classification, the prediction performance was comparable to the full-gene model, with F1-score differences of between −5% and 5%. In the ternary classification, the performance was significantly better, with F1-score differences ranging from −2% to 12%, while also maintaining excellent interpretability of the selected feature genes. **Conclusions:** This framework effectively integrates feature selection, adversarial sample handling, and model optimization, offering a valuable tool for a wide range of biological data classification problems. Its ability to balance performance accuracy and high interpretability makes it highly applicable in the field of computational biology.

Keywords: feature gene selection; model selection; machine learning; interpretability



Academic Editor: Hongyan Xu

Received: 19 January 2025

Revised: 11 February 2025

Accepted: 24 February 2025

Published: 28 February 2025

Citation: Huang, Z.; Mu, X.; Cao, Y.; Chen, Q.; Qiao, S.; Shi, B.; Xiao, G.; Wang, Y.; Xu, Y. Optimizing Model Performance and Interpretability: Application to Biological Data Classification. *Genes* **2025**, *16*, 297. <https://doi.org/10.3390/genes16030297>

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Machine learning encompasses a variety of techniques aimed at identifying a mapping from a feature space to labeled outputs while optimizing a given objective function. These methods have found wide application in data analysis across a range of domains that

generate large volumes of data. Various machine-learning algorithms, such as neural network-based methods [1], support vector machine (SVM) [2], regression models [3], and decision tree-based methods [4,5], are commonly employed for addressing different data-analysis challenges. It is widely recognized that the effectiveness of these algorithms can vary depending on the problem at hand, as highlighted by the “No Free Lunch Theorem” [6]. Hence, a common issue faced by data-analysis practitioners, particularly those who are not familiar with the detailed mathematics underlying the relevant algorithms, is: which of these algorithms makes the best choice for his/her problem?

Effective algorithm selection must address three key issues: (1) interpretability—whether the model provides biologically meaningful insights into the features [7]; (2) predictive performance—whether the model achieves robust accuracy [8]; and (3) data suitability—whether the model aligns well with the characteristics of the data [9]. Current approaches to feature interpretability often rely on feature importance metrics [10,11], which tend to overlook interactions among features. In biological systems, synergistic interactions can be more informative than the contribution of individual features [12,13]. Similarly, while performance metrics such as the F1 score and AUC are useful and serve as good indicators for model selection, they may not adequately capture the true predictive capability for omics data, where high label noise is common [14].

To address this challenge, we propose a framework that integrates interpretable feature selection and model selection, specifically designed for omics data analyses. Our approach is grounded in the principle that feature selection should be based on target-related pathways, which are not only strongly discriminative but also enable meaningful interpretation of the features selected [15], as high interpretability is crucial for understanding the biological significance of the models built. Additionally, the framework incorporates adversarial samples [16,17] to assess the sensitivity of both the features and models, effectively addressing the inherent uncertainties that omics data may have [18,19].

Overall, our framework covers feature selection and model selection with the goal of identifying a model that can differentiate samples sharing common yet hidden labels, based on a concise set of interpretable features. The framework consists of two main components: one for selecting interpretable features and another for evaluating models that best align with the data. To achieve the framework’s objectives, we ultimately develop a stacking meta-classifier using the outcomes of model selection. In this work, we explore five machine learning models, with the flexibility to incorporate additional models in future applications.

This study focuses on analyzing transcriptomic data in the context of human disease research, specifically utilizing cancer tissue transcriptomic data from the TCGA database [20]. We illustrate the effectiveness of our framework by applying it to various cancer data analysis problems. For simplicity, throughout the paper, we use the terms machine-learning algorithms and models interchangeably. In conclusion, the key contributions of our work are as follows:

1. We propose a novel framework that integrates interpretable feature selection and robust model selection by incorporating adversarial samples, thereby enhancing both predictive performance and biological interpretability.
2. We introduce a domain-specific feature selection strategy based on target-related pathways in transcriptomic data, which outperforms conventional general-purpose methods.
3. We develop a stacking meta-classifier that demonstrates superior performance in both binary and ternary classification problems, underscoring its potential for broad applications in omics data analysis.

2. Materials and Methods

2.1. Datasets

This study presents two classification tasks to showcase the effectiveness of the proposed method. The first task involves identifying genes whose expression levels can effectively differentiate between 368 melanoma samples that have metastasized and 102 samples that have not, with data sourced from the TCGA [21]. Due to the occurrence of distant non-lymph node metastasis, some of the primary cancer samples in this dataset may actually be metastasized, but they have been labeled as primary cancer, making them adversarial samples [22]. The second task focuses on distinguishing genes whose expression levels can separate 436 primary cancer samples based on their respective metastatic sites: 207 to the liver, 177 to the lung, and 52 to the bone, with all data also from TCGA. For these sites, the metastatic cancer cells might remain dormant, introducing some uncertainty in the accuracy of metastatic site identification [23]. For clarity, we refer to these two problems as the binary and ternary classification tasks, respectively, in the Section 3.

Each dataset includes expression data for 60,499 genes, as provided by GENCODE v23 [24], while enzyme-related genes are sourced from the HumanCyc database, totaling 2453 genes [25]. All gene expression data are standardized to TPMs (transcripts per million) for consistency.

2.2. Basic Machine Learning Models

This study evaluates the predictive performance using five existing classification models: logistic regression (LR) [3], LightGBM (LGBM) [5], random forest (RF) [26], support vector machine (SVM) [27], and XGBoost (XGB) [4], which is readily extendable to include more classification models. To ensure classification generality, the dataset for each classification problem is split with a random seed, currently using 42, and all models undergo 5-fold cross-validation (CV).

2.3. Feature Selection for Transcriptomic-Data-Based Classification

We begin the feature selection process by focusing on the 2453 enzyme genes [25]. The objective is to identify a small subset of genes whose expression profiles possess sufficient discriminative ability to group the labeled samples into categories, each corresponding to a specific label, while ensuring that the biological roles of these genes provide meaningful insight into the separation.

Assessing a gene (feature)'s importance to a classification problem: Consider a set of samples $\mathbf{S} = \{s_1, \dots, s_n\}$, each having a label from $L = \{l_1, \dots, l_m\}$, denoted as $L(\mathbf{S})$, with each s_i representing the expressions of K genes. A classification problem is to find a subset of k genes $G = \{g_1, \dots, g_k\}$ so that a logistic regression of G 's expressions can accurately predict $L(\mathbf{S})$. The level of contribution by gene g_i to the performance by the regression model is measured using the standard *importance score* [28], defined as the absolute value of the coefficient of the g_i term in the model. Using this approach, one can assess the importance level of each gene in G to the classification problem and remove those with no or little importance when solving the classification problem.

While the approach has been used widely, its weakness has been pointed out and alleviated by multiple authors using different techniques [29,30]; where the issue lies in that it tends to underestimate the importance of informative genes whose expressions weakly and nonlinearly correlate with $L(\mathbf{S})$ [31]. A general strategy for overcoming the weakness is through applying a technique called *non-importance scoring* [32], which is based on the following premise: The importance value of a gene having true discerning power to the classification problem as defined above, called \mathbf{C} , generally goes down in a modified classification problem with mislabeled samples, denoted as \mathbf{C}' , which is the same as \mathbf{C}

except that the labels of a subset S' of S are permuted. An execution of this strategy can be conducted through comparing the importance value of the gene in C vs. the importance-score distribution across a set $\{C'\}$, each having a different permutation of the labels of S' , and the gene is considered as *important* if its importance score to C is generally higher than those with partially permuted labels.

Differential gene filtering: DESeq2 [33] was used to identify, among genes deemed to be important, differentially expressed genes between samples with one label and those with others. Genes with a fold change ($|FC|$) ≥ 1.5 and an adjusted p -value (adj. p -value) < 0.05 were considered differentially expressed. Then, a pathway enrichment analysis was conducted using ClusterProfiler software v3.20 [34] with differentially expressed genes. Only pathways with an adjusted p -value < 0.05 for the enrichment were selected for further analyses.

Representative pathway identification: For a given enriched pathway, let X represent the gene expression matrix, where each column corresponds to a gene and each row represents a sample. The covariance matrix of X , denoted as C , is calculated as:

$$C = \frac{1}{n-1} (X - \bar{X})^T \cdot (X - \bar{X}) \quad (1)$$

where n is the total number of samples, and X is the vector of mean expressions for each gene across all samples in X . Significantly enriched pathways are required to contain no fewer than five genes.

Let PC_1 be the first principal component of X , which captures the maximum variance in the pathway's gene expression data. Let V be the variance of PC_1 . In our pathway-based feature selection, we select pathways first, whose V values are sufficiently large. However, the V value is pathway size-dependent, which tends to go down with the increase in the number of genes of a pathway. Hence a V value-based selection procedure requires normalization with respect to the pathway size. Based on our preliminary analyses, we first consider only pathways with $V > 0.7$. To derive an empirical relationship between V values and pathway sizes, we have considered the density distribution of pathway size k across these pathways, revealing that it follows a power law distribution [35] $\alpha = 0.5$. Let k_m be the median of this distribution. For a pathway with size k and variance V , we normalize the V value as follows:

$$V' = V \times \left(\frac{k_m}{k} \right)^\alpha \quad (2)$$

Selection of representative genes for a selected pathway: For each gene g_i within the pathway, denote v_i as the loading of gene i on the first principal component (PC_1), and $|v_i|$ represents the contribution of i to PC_1 . The total contribution to PC_1 is given by the sum $\sum_{i=1}^t |v_i|$, where t is the total number of genes in the pathway. For simplicity, we assume that the values of v_1, v_2, \dots, v_t are sorted in descending order based on the magnitude of $|v_i|$. Let s be the smallest index such that:

$$\sum_{i=1}^s |v_i| > 95\% \sum_{i=1}^t |v_i|. \quad (3)$$

We leave out genes $v_{s+1} \dots v_t$ from further consideration. To address redundancy in the enriched pathways, we remove parent pathways when parent-child relationships are detected, retaining only the more specific child pathways.

Filtering of genes insensitive to adversarial samples: Similarly, a gene-sensitivity testing method was employed to assess the robustness of the selected genes against *adversarial* samples, defined as samples with their labels randomly permuted. Across

100 random permutations, we obtained the probability density distribution of each gene's importance score. Genes that showed important scores lower than those under true labels in at least 95 of these 100 iterations were retained for final analysis, ensuring insensitivity to adversarial samples.

2.4. Model Selection for a Given Classification Problem

For a variety of classification tasks, such as cancer subtyping based on gene expression data, we propose a method to evaluate the appropriateness of different models for the given problem. The following criteria are used to select the most suitable model: (1) *Classification accuracy*, which is quantified using the F1 score, calculated as:

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

where precision and recall are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \text{ and } \text{Recall} = \frac{TP}{TP + FN}. \quad (5)$$

In these formulas, TP and FP represent the true positive and false positive counts, respectively. (2) *Classification stability* (CS), which assesses the model's sensitivity to random mislabeling of N% of the samples. Specifically, it evaluates the difference in F1 score between the original and mis-labeled datasets. We calculate this as:

$$CS = TP - TP', \quad (6)$$

where TP' is the same as TP but on the dataset with mislabeling. A small CS indicates that the model performs similarly on both original and altered data, implying robustness to noisy labels. (3) *Classification robustness* (CR), which evaluates the model's ability to correctly classify mislabeled samples. This is calculated as:

$$CR = TP' - (TP - k), \quad (7)$$

where k is the number of mislabeled samples that belong to the true positives. CR reflects the model's capacity to recover from mislabeling and maintain accurate classification. A higher CR value signifies greater robustness to label noise.

To ensure reliable results, the process of randomly mislabeling 20% of the samples is repeated 100 times, and the average performance across all 100 modified datasets is used to derive the final metrics. In summary, the model selection process considers these factors: classification accuracy, stability, and robustness, each evaluated through specific performance metrics. Hence, for the selected model, we have

$$CS = TP - \frac{\sum_i TP'_i}{100}, \quad (8)$$

$$CR = \frac{\sum_i (TP'_i - (TP - x))}{100} = \frac{\sum_i TP'_i}{100} - 80\%TP. \quad (9)$$

A possible approach to integrate the three objectives is by using a weighted sum of the three components:

$$\alpha F_1 + \beta CS + \gamma CR \quad (10)$$

where α , β , γ are positive scaling factors, which will be determined empirically based on specific needs such as higher CS or higher CR. By default, α is set to 2 to amplify the

weight of the $F1'$ score, ensuring that classifiers with better $F1'$ performance are given more importance. β is set to -1 , and γ is set to 0.1 to mitigate the impact of negative values.

2.5. Constructing an Integrative Pipeline

We have developed a two-level stacking pipeline [36] to incorporate the five classifiers, namely, LR, SVM, RF, LGBM, and XGB, via a meta-learner. The pipeline utilizes each of the classifiers to generate one prediction using the selected feature genes. The meta-learner then integrates the five predictions to produce the final classification, where the weights of the five classifiers are determined to optimize the classification result for each given problem, measured using $F1'$, CS, and CR (see Section 2.3). Specifically, the meta-learner is implemented using a gradient boosting algorithm [37], which effectively combines the weighted predictions by minimizing a loss function tailored to the specific classification metrics. The first-level weight calculation process is as follows:

$$\text{Score}_i = F1'_i + CS_i + CR_i \quad (11)$$

where $F1'_i$, CS_i , and CR_i are scores by the i^{th} classifier. The weight of the i^{th} classifier, as determined by the second-level learner (meta-learner), is defined as follows:

$$w_i = \frac{\exp(\text{Score}_i)}{\sum_{j=1}^5 \exp(\text{Score}_j)}. \quad (12)$$

2.6. Comparative Evaluation of Predictive Performance, Interpretability, and Robustness of Models

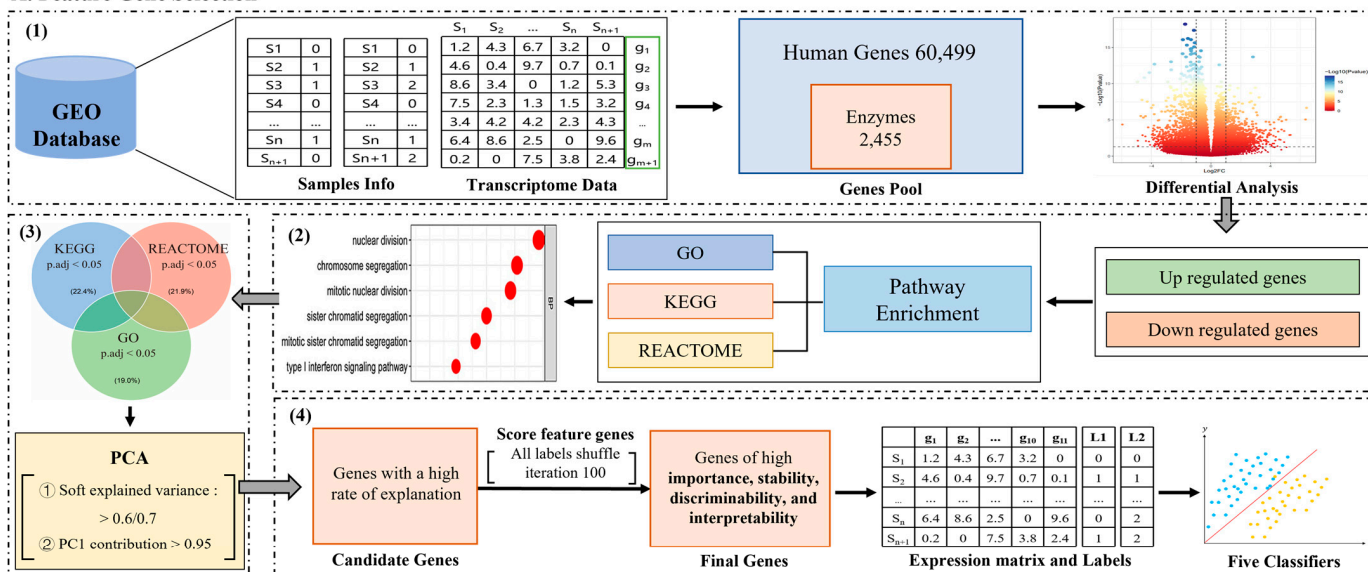
To compare the quality of feature selection by our procedure and that by a state-of-the-art method, we have employed Recursive Feature Elimination (RFE) [38], a widely used method for feature selection.

We have compared our classification framework with both the machine learning and deep learning models using two classification problems (see Section 2.1), in the following categories: prediction accuracy, interpretability, model stability, and robustness. Here, for the deep learning models, we utilized a CNN-based architecture—LeNet [39]—and a deep neural network (DNN) model [40].

To evaluate the interpretability of a classification model, we compared it against two classes of widely used models: intrinsic interpretable white-box models [41] and post hoc surrogate models [42]. For intrinsic interpretability, focusing on the structural simplicity of a classification method, we selected two state-of-the-art white-box models: the Explainable Boosting Machine (EBM) [43] and RuleFit [44]. For post hoc interpretability, focusing on the ability to construct an interpretable approximation rather than to reconstruct the actual logical process for computing the results, we employed the game theory-based SHAP [45] technique. The evaluation criteria for interpretability included the contributions to classification results by feature genes and their functional annotation.

Figure 1 summarizes the procedure for feature and model selection. The test data, code, and the stacking pipeline are available for download at https://drive.google.com/drive/folders/17CWxRNx1Cdm17Iib_d8_lyg12VrRohTx (accessed on 20 January 2025).

A. Feature Gene Selection



B. Model Selection and the Stacked Voting Classifier

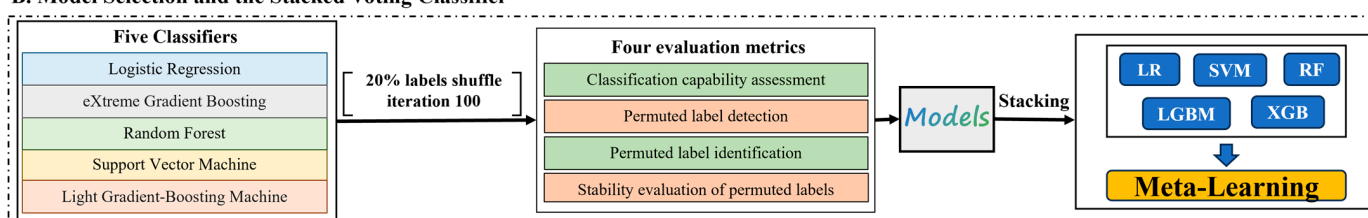


Figure 1. Construction of a feature selection and model selection framework. **(A)** This step includes four stages: (1) Identification of differentially expressed genes. (2) Identification of representative metabolic pathways in samples with distinct labels. (3) Identification of key genes within the representative metabolic pathways. (4) Identification of genes that are insensitive to adversarial samples. **(B)** Evaluation of model suitability for data based on accuracy, classification stability, and robustness. Subsequently, computation of the weights of primary classifiers based on scores derived from model selection, followed by the construction of a meta-classifier.

3. Results

We assessed our classification framework using two classification problems, namely, the binary and ternary classification problems defined in Section 2.1, and compared its performance with the five state-of-the-art methods specified in Section 2.6.

3.1. Feature Gene Selection

To assess the importance scores of candidate feature genes, we evaluated their sensitivities to adversarial samples with each of the six classifiers, ours plus the five classifiers mentioned in Section 2.2. For each gene, we performed random permutations on the labels on a subset S' of the whole sample set S 100 times and measured the changes in the gene's importance scores [29]. A gene was considered *sensitive* to the adversarial samples if its importance scores in at least 95% of permuted sets were higher than its score in the un-permuted set, which was removed from further analyses.

Our past experience has shown that among all genes, both protein and RNA genes, enzyme genes tend to have the highest discerning power, which makes sense, as the behaviors of a disease are the direct acts of enzymes, rather than signaling or regulatory genes [46]. Hence, we used enzyme genes as the feature genes for classification problems. As the first step of feature selection based on *differentially expressed genes* (DEGs), we identified 215 enzyme genes among 3368 DEGs in the binary classification problem and

435 enzyme genes among 6657 DEGs in the tertiary dataset (Figure 2A). Furthermore, biological pathway enrichment analyses of the DEGs revealed that, for the binary classification problem, metastasized and primary cancers enriched 1079 and 278 functional pathways, respectively, and for the ternary dataset, pathway enrichment analyses showed that 294, 1182, and 1192 pathways are enriched by DEGs in cancer samples with metastases to bone, liver, and lung, vs. controls, respectively (Figure 2B).

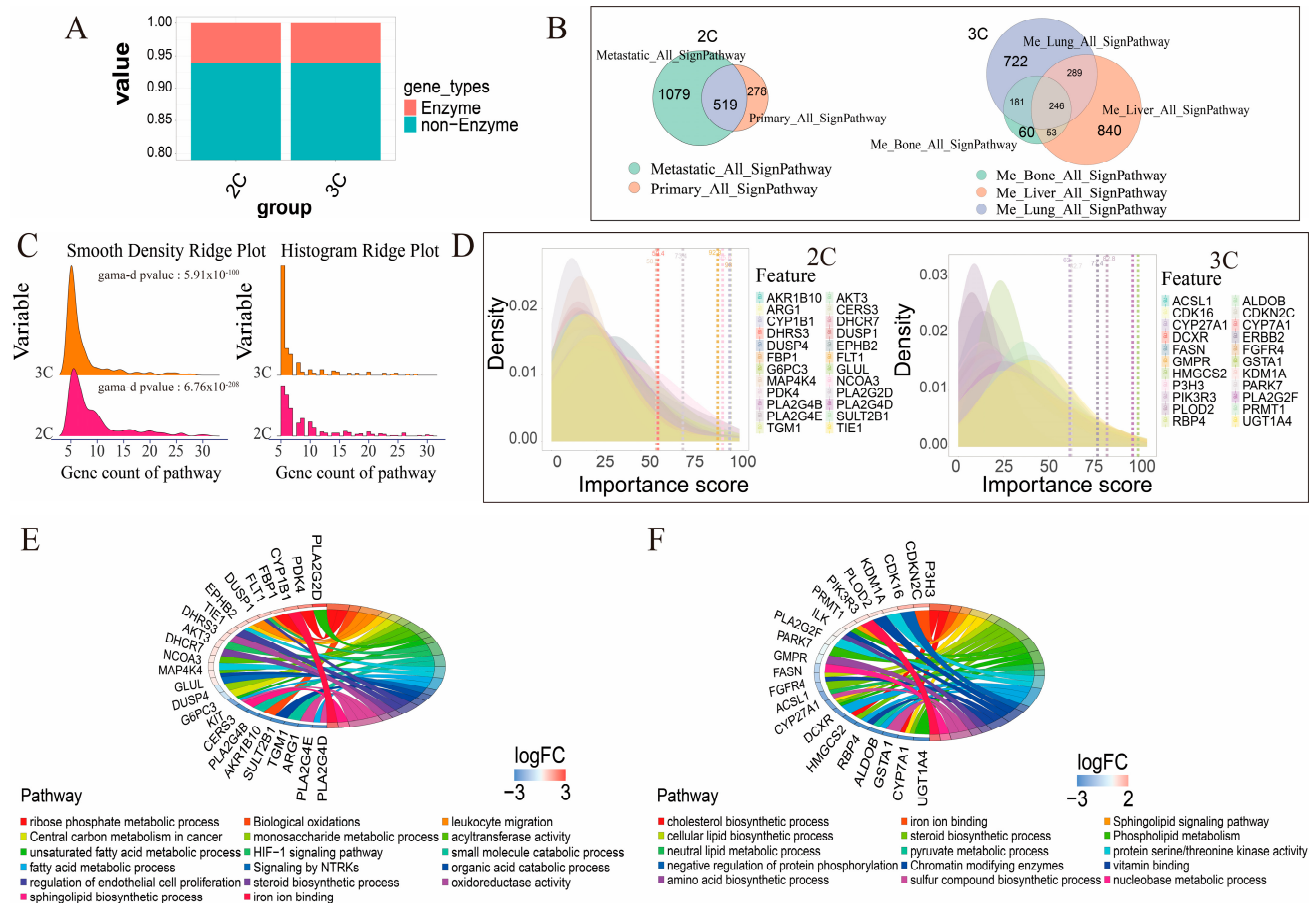


Figure 2. Feature evaluation. (A) Statistical overview of DEGs and enzyme genes in the binary and ternary datasets. (B) Occurrence statistics of the enriched pathways across sample groups with distinct labels in the binary and three-class datasets. (C) Gene counts in pathways meeting thresholding criteria in the binary and ternary datasets. Kolmogorov–Smirnov (KS) test for γ distribution is provided. (D) Distribution of importance scores over adversarial samples for the final selected features in the binary and ternary datasets. (E) Relationships between the final selected feature genes and cellular functions in the binary classification dataset. (F) Relationships between the final selected feature genes and cellular functions in the ternary classification dataset. Note: “2C” represents the binary classification dataset, and “3C” represents the ternary classification dataset.

For the classification problem, we selected among the enriched pathways that meet the selection criteria defined in Section 2.3. We noted that the number of genes per pathway across all the selected pathways follows a left-skewed γ distribution ($\alpha < 1$) (Figure 2C), based on which we defined the selection criteria equation $V' = 0.7 \times r^\alpha$, $\alpha = 0.5$. Here, V' represents the filtering threshold for the first principal component (PC1).

For the selected feature genes based on the selected pathways (Section 2.4), we calculated each gene’s importance score over the adversarial samples. As shown in Figure 2D, genes meeting our criteria consistently exhibit lower importance scores over the adversarial samples than over the samples with the correct labels. Applying a majority-vote rule, we selected 25 enzyme genes for the binary classification problem and 23 for the ternary prob-

lem (Figure S1). Table S1 lists the ranking of the importance scores of the distinguishing enzyme genes for the two problems.

As illustrated in Figure 2E and Table S2, the relationships between the feature genes and biological pathways reveal distinct differences between metastatic and non-metastatic cancers. Notably, genes such as AKT3 [47], involved in the hypoxia-stress-induced HIF-1 signaling pathway, CYP1B1 [48], which plays a role in oxidative stress and iron accumulation, and PLA2G2D [49], which is implicated in fatty acid synthesis, were key discriminants. These findings align with evidence showing that primary cancer sites are often characterized by hypoxic and highly oxidative environments [50–53]. Specifically, iron accumulation is associated with intracellular alkalosis, while fatty acids act to mitigate this alkalosis [54,55]. Furthermore, 80% (19/25) of the selected enzymes produce H^+ in the reactions they catalyze, which is consistent with the upregulation of acidifying metabolic reprogramming in cancer [54,55].

In the ternary classification task, biological pathways differentiating the three metastasis sites—liver, lung, and bone—highlighted processes such as cholesterol synthesis, lipid metabolism, oxidoreductase activity, and pyruvate metabolism (see Table S2, Figure 2F). These findings are consistent with known characteristics of the three organs. For example, immune response-related pathways (ILK [56]), high-energy phosphate transfer reactions (PLA2G2F in liver [57]), oxidation–reduction processes (GSTA1 in lung [58]), the regulation of nerve cell differentiation (RBP4 in lung), and erythrocyte differentiation and melanosome organization in bone metastases (PIK3R3 [59]) were particularly prominent. Additionally, cholesterol accumulation, seen in the involvement of CYP7A1 [60], was common across all metastatic sites [61,62].

To further validate the effectiveness of our feature selection approach, we compared the performance of the five models trained using the selected enzyme genes against models trained on the full feature set (60,499 genes). In the binary classification problem, as shown in Figure 3A,B, LR, SVM, and RF achieved 3–4% higher F1 scores with our selected features, while LGBM and XGB exhibited a 5% decrease. Similarly, in the ternary problem (Figure 3C,D), all models except SVM showed performance improvements of 2–12% with our selected features.

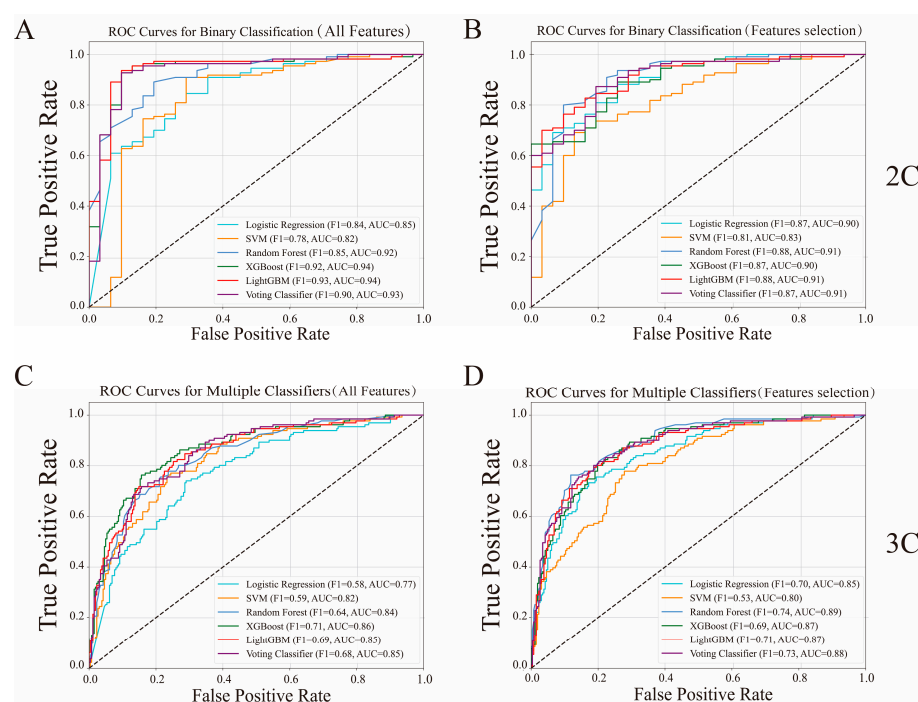


Figure 3. Comparison of prediction performance for five classifiers and the stacking voting classifier using all genes vs. selected feature genes in two problems. (A) Prediction performance in the binary

classification using the full gene set (60,499 genes). (B) Prediction performance in the binary classification using our 25 enzyme genes. (C) Prediction performance in the ternary problem using the full gene set (60,499 genes). (D) Prediction performance in the ternary problem using our 23 enzyme genes. Note: “2C” represents the binary classification dataset, and “3C” represents the ternary classification dataset.

3.2. Model Assessment

We then used the procedure given in Section 2.4 for model selection from the five candidate classifiers. Figure 4 shows the performance in terms of performance accuracy, stability, and robustness in the two classification problems of the five classifiers, which are detailed in Table 1. Overall, in the binary problem, RF and the stacking voting classifier exhibit the best predictive performance, while in the ternary problem, the stacking voting classifier demonstrates the highest predictive performance (Figure 4A,B).

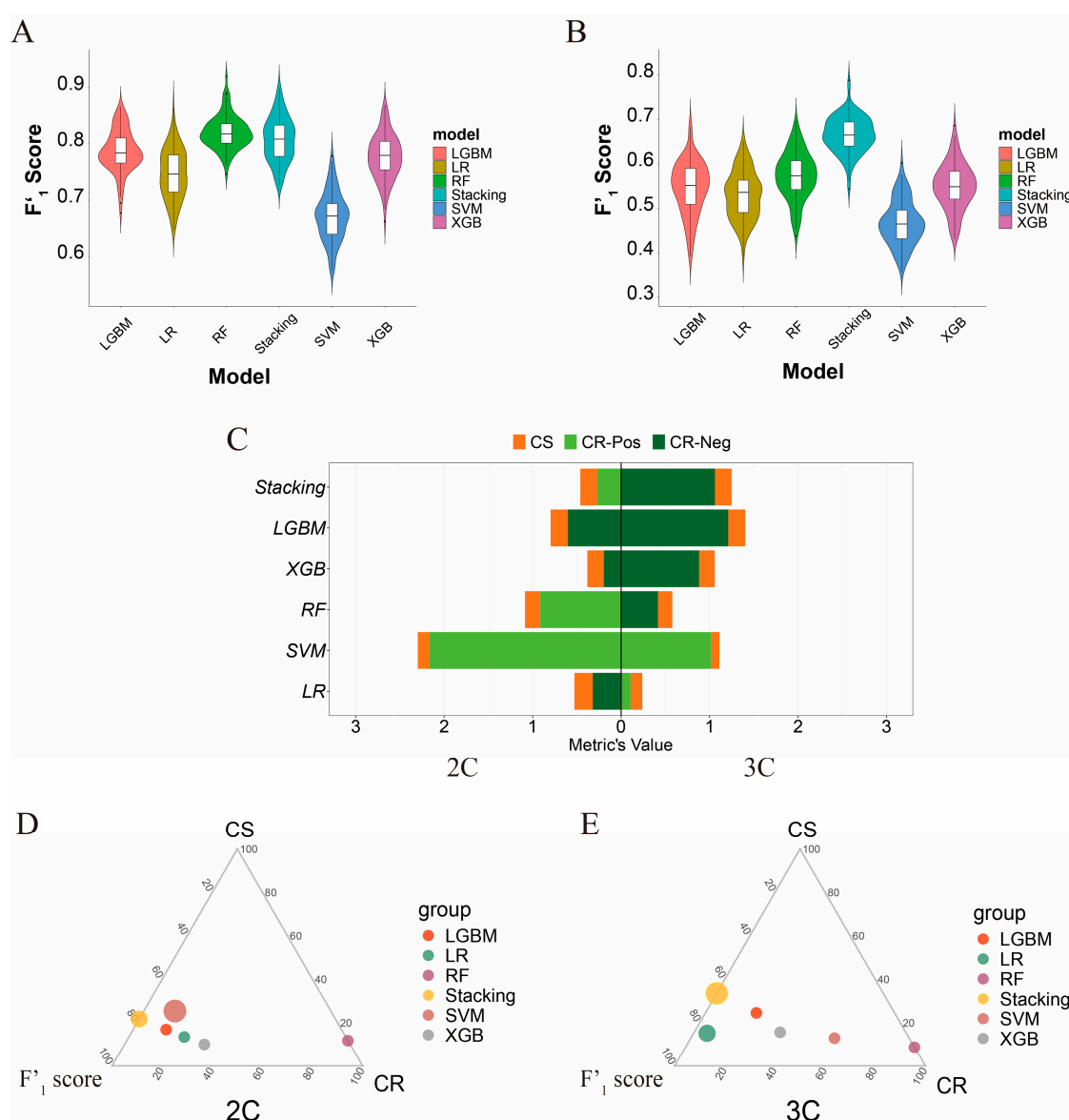


Figure 4. Model selection. (A) F_1 score evaluation of six models in the binary problem. The square box represents the variance of the F_1 score over 100 iterations. (B) F_1 score evaluation of six models in the ternary problem. (C) Classification stability and robustness of the six models. (D) Comprehensive evaluation of F_1 score, CS, and CR metrics of the six models in the binary problem. (E) Comprehensive evaluation of F_1 score, CS, and CR metrics of the six models in the ternary problem.

Table 1. Performance measures for model selection, in terms of performance accuracy, classification stability, and robustness.

Model	Labels	F ₁ Score	CS	CR	Final Score
LR	2C	0.7492	0.2073	−0.3200	1.2591
SVM	2C	0.6739	0.1327	2.1650	1.4316
RF	2C	0.8221	0.1710	0.9150	1.5647
XGB	2C	0.7892	0.1869	−0.1950	1.3720
LGBM	2C	0.7791	0.1969	−0.6000	1.3013
Stacking	2C	0.8097	0.1904	0.2700	1.4560
LR	3C	0.5262	0.1325	0.1067	0.9306
SVM	3C	0.4641	0.0948	1.0167	0.9351
RF	3C	0.5692	0.1617	−0.4167	0.9350
XGB	3C	0.5452	0.1786	−0.8800	0.8238
LGBM	3C	0.5412	0.1928	−1.2100	0.7686
Stacking	3C	0.6659	0.1902	−1.0600	1.0356

Figure 4C and Table 1 present the classification stability of the five classifiers in the two problems. In both the binary and ternary problems, SVM has the best performance. In terms of classification robustness, SVM also exhibits the highest performance in both the binary and multiclass datasets.

Overall, when the F₁' scores for the five classifiers on the unmodified label samples are high, SVM is the preferred model. However, when the F₁' scores are suboptimal, the stacking voting classifier emerges as the superior choice (Figure 4D,E).

3.3. Performance Analysis of a Stacking-Based Voting Meta-Classifer

Due to imbalance issues across sample groups with different labels in both problems (the binary problem: 368:102; the ternary problem: 207:77:52), the F1 score, which is more suitable for imbalanced data, is reported for the test set (without adversarial examples). In both the binary and ternary problems, our classifier ranks in the top two performers, (Figure 5A,B). The prediction accuracy in the test set shows a similar trend to the F1 score (Figure 5C,D).

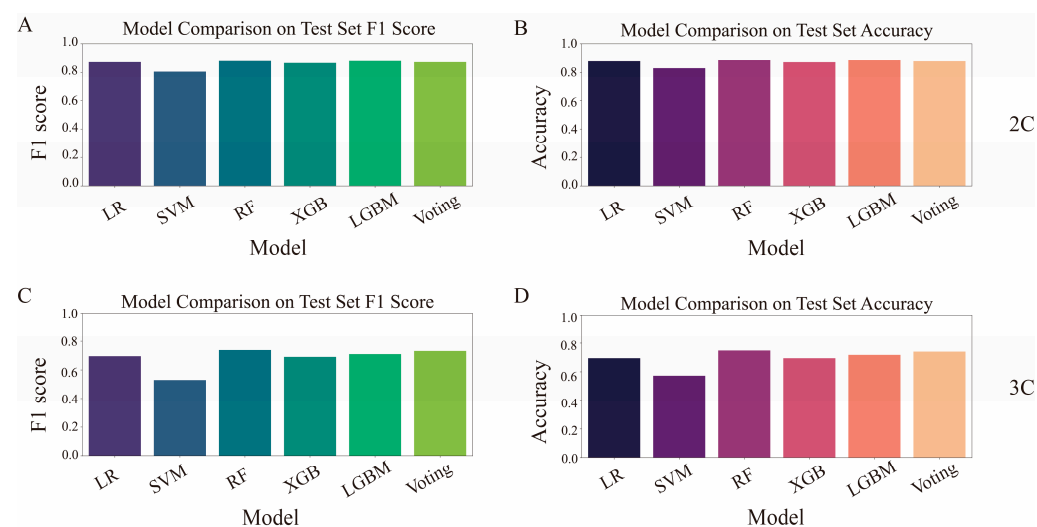


Figure 5. F1-score and accuracy statistics of the stacking voting classifier in the test set by our framework. (A) F1-score statistics of our classifier in the binary problem. (B) Accuracy statistics of our classifier in the binary problem. (C) F1-score statistics of our classifier in the ternary problem. (D) Accuracy statistics of our classifier in the ternary problem.

3.4. Comparison Between General-Purpose Feature Selection and Our Feature Selection

We compared our feature selection procedure with a widely used general-purpose feature selection procedure, the RFE approach, aiming to gain insights about the level of improvement that domain-specific knowledge can provide for general-purpose feature selection.

We compared the classification results obtained using RFE with those using our feature selection method. As shown in Figure 6A,B and Table S3, across all five models, our feature selection method outperforms RFE in terms of predictive performance, except for XGB in the binary classification. In the ternary problem, our feature selection method also outperforms RFE in terms of predictive performance, except for SVM.

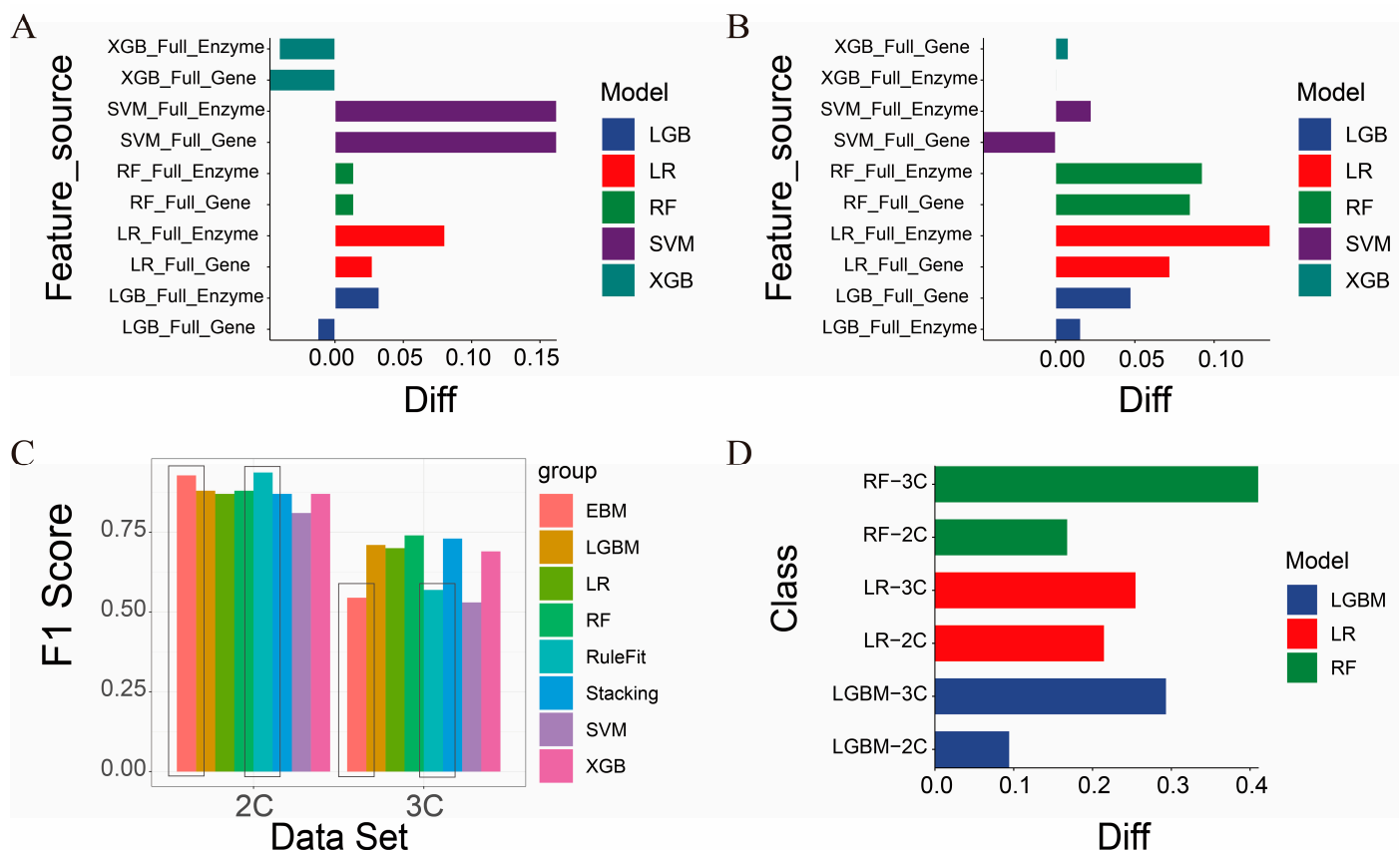


Figure 6. Comparison of prediction performance between our model, RFE, interpretable white-box models, and SHAP. **(A)** Difference in F1 scores between features selected by our framework and the top features selected by RFE in the binary classification. Positive values indicate higher F1 scores achieved by our framework. **(B)** Difference in F1 scores between features selected by our framework and the top features selected by RFE in the ternary classification. Positive values indicate higher F1 scores achieved by our framework. **(C)** F1-score comparison between features selected by our framework and the top features selected by white-box models (EBM and RuleFit) in both binary and ternary classification problems. **(D)** Difference in F1 scores between features selected by our framework and the top features selected by SHAP in the binary and ternary classification problems. Positive values indicate higher F1 scores achieved by our framework.

Similarly, among the 25 selected feature genes for the binary problem and the 23 selected genes for the ternary problem, we observe that the majority were regulatory non-enzyme protein-coding genes and non-protein-coding genes, irrespective of the classifier used (Table S4). Furthermore, within these genes, only the top genes selected by LGBM's RFE demonstrated excellent resistance to adversarial perturbations in other mod-

els, whereas the genes selected by the other models exhibited poor resistance to adversarial samples (Table S5).

3.5. Comparison of Model Interpretability

3.5.1. Comparison with the White-Box Models

White-box models derive their interpretability from both the predictive performance of the model and the ranking of feature importance; generally, better prediction performance and higher-ranked feature importance correspond to more understandable feature explanations [63]. Accordingly, we used EBM and RuleFit to select the same number of feature genes as our framework from the 60,499 genes in the dataset: 25 genes for the binary classification dataset and 23 genes for the ternary classification dataset. As shown in Figure 6C, for the simpler binary classification task, EBM and RuleFit exhibit the best performance (Table S6), whereas they clearly underperform in the more complex ternary classification task compared to the stacking voting classifier. Furthermore, when adversarial samples are introduced, the $F1'$ score of the stacking voting classifier significantly outperforms both EBM and RuleFit in both binary and ternary settings. After considering both CS and CR, the stacking voting classifier remains the top performer (Table 2).

Table 2. Comparison of the stacking voting classifier, interpretable white-box models, SHAP, and neural network models in terms of performance accuracy, classification stability, and robustness.

Method	Labels	F1_Score	CS	CR	Final Score
Stacking	2C	0.8097	0.1904	0.2700	1.4560
EBM	2C	0.6869	0.18017	0.8699	1.2806
RuleFit	2C	0.6414	0.2215	−2.555	0.8058
LeNet	2C	0.6286	0.1145	2.633	1.4060
DNN	2C	0.6684	−0.0018	2.246	1.5632
Stacking	3C	0.6659	0.1902	−1.0600	1.0356
EBM	3C	0.5366	0.16255	−0.32	0.8787
RuleFit	3C	0.5470	0.16258	−0.4633	0.8671
LeNet	3C	0.2344	0.00061	1.9013	0.6583
DNN	3C	0.2653	−0.0006	1.914	0.7226

Regarding feature gene interpretability, EBM provides more reasonable selections than RuleFit, as approximately 30% of the genes identified by RuleFit lack functional annotations (Table S7). In contrast, while both EBM and RuleFit predominantly select non-enzyme genes (95%), these genes, such as SMIM2 antisense RNA 1, are not notably linked to tumor hallmarks, unlike the feature genes identified by our framework (Table S7).

3.5.2. Comparison with SHAP Models

Due to the substantial memory and time requirements for training SHAP with SVM and XGB models—exceeding 1 TB of memory—we limited our SHAP analysis to LR, LGBM, and RF results. The interpretability of SHAP models similarly relies on feature importance [64]. Consistent with the RFE results, when training models using an equivalent number of our selected features, we found that for both binary and ternary classification problems, the F1 scores based on our feature selection surpassed those obtained using SHAP by at least 10% (Figure 6D, Table S8). Similarly, to analyze the feature interpretability provided by SHAP, we examined the top features it ranked as highly important.

We observed that 60% of these features were non-coding protein genes or pseudogenes (Tables S9 and S10).

3.5.3. Comparison with Neural Network Models

To assess the performance of neural network models within our framework, we first evaluated the LeNet and DNN models using a complete gene set comprising 60,499 genes. As shown in Table S8, the prediction accuracies of LeNet and DNN for the binary classification task were 0.7882 and 0.815, respectively. For the multiclass classification task, the accuracies of LeNet and DNN were 0.29881 and 0.5291, respectively (Table S11).

Further evaluation of the impact of adversarial samples on the neural network models revealed that, as shown in Table 2, the DNN model outperformed all other models in the binary classification dataset. However, in the multiclass dataset, the performance of both neural networks significantly declined, with F1 scores of 0.2344 and 0.2653, respectively.

Additionally, we have provided a summary flowchart of the analysis framework and evaluation process in Figure 7.

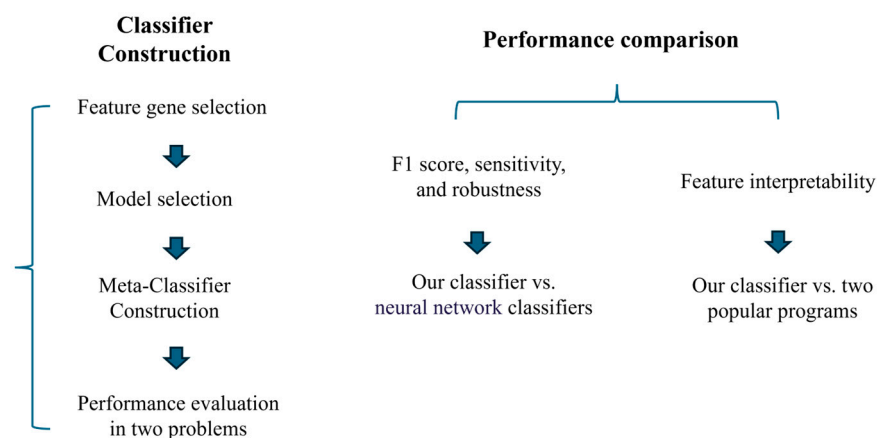


Figure 7. Overview of our algorithmic framework.

4. Discussion

This work introduces a machine learning framework designed specifically for classification tasks using transcriptomic data. The goal is to assist researchers in effectively combining biological insights with machine learning models that are best suited to the specific nature of the classification problem. The biological knowledge is incorporated through pathway-based analyses [65]. For selecting the optimal model, two essential pieces of information are needed: the most relevant features and models that exhibit the necessary characteristics for superior performance. This method not only achieves high prediction accuracy but also highlights features with significant interpretability. Moreover, by employing five base learners within a stacked ensemble framework, guided by comprehensive model selection criteria, we developed a meta-model that guarantees robust performance, even in the presence of adversarial perturbations.

Additionally, we compared our framework with current popular interpretable white-box models, such as EBM [43] and RuleFit [66], as well as post hoc interpretability proxy models like SHAP [45]. Our framework demonstrated superior interpretability by selecting biologically meaningful and robust features that are directly linked to relevant metabolic pathways [67]. We have demonstrated that our overall classification framework achieved higher prediction accuracies compared to state-of-the-art methods, highlighting the effectiveness of our integrated approach in enhancing both the explainability and performance of classification models in transcriptomic data analysis [68].

Furthermore, we explored the performance of neural network models within our framework. Neural networks exhibited notable advantages in terms of robustness to adversarial samples, as evidenced by their high CR values in both the binary and ternary classification tasks. This robustness suggests that neural networks can maintain performance integrity even when faced with perturbed or noisy data [69]. However, their prediction accuracy in the ternary classification problem was significantly hampered, primarily due to the limited size of the training dataset. The scarcity of samples restricts the neural networks' ability to generalize effectively across multiple classes, resulting in suboptimal performance compared to more traditional machine learning models [70]. This trade-off between robustness and predictive accuracy underscores the challenges of applying deep learning techniques to small-sample omics data, emphasizing the need for larger datasets or enhanced training strategies to fully leverage the strengths of neural networks in such contexts [71].

It is important to mention that our current study is somewhat exploratory in nature. As we continue to address a broader range of biological classification problems, we plan to expand our analyses systematically by incorporating additional models, potentially grouped into categories such as tree-based and SVM-like techniques, among others. This will allow us to deepen our understanding of which types of classification problems are best suited to specific classes of methods [72]. Moreover, future work will consider a broader set of performance metrics, including the effects of noise and missing data on classification tasks, as well as identifying the most effective techniques for addressing these challenges [73].

In summary, our research introduces a comprehensive and flexible framework for feature and model selection, enabling the identification of the most suitable models based on the specific traits of the data. This work contributes to improving classification accuracy, stability, and robustness, particularly in bioinformatics and other fields that deal with large, complex datasets. Future developments will refine and extend this framework, making it applicable to a wider variety of datasets and classification tasks. Our methodology provides a practical approach for data analysts, especially those who may not be deeply familiar with the mathematical foundations of machine learning algorithms, allowing them to select the optimal model for their challenges and maximize the potential of machine learning in data-heavy environments.

5. Conclusions

We have designed a framework for feature and model selection tailored to omics data classification tasks, alongside a meta-classifier learning approach that ensures both high classification performance and the interpretability of the results. This framework is implemented through a multi-step selection process, involving gene selection, model identification, and the aggregation of meta-classifiers. The interpretability is achieved by focusing on pathways that are well-characterized, have strong discriminative power, and play a key role in explaining the primary variance of gene expression across a wide range of samples.

We have shown the efficacy of this framework in addressing two distinct classification challenges: one involving binary classification and the other a three-class classification task. We believe this framework has the potential to provide a versatile solution for both feature and model selection in a wide range of omics classification applications.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes16030297/s1>, Figure S1 and Tables S1–S11.

Author Contributions: Conceptualization by Z.H., X.M., Y.W., and Y.X.; methodology by Z.H., X.M., and Y.X.; software by Z.H., Y.C., and X.M.; formal analysis by Z.H., Y.C., and X.M.; investigation by Z.H. and X.M.; data curation by Z.H., X.M., B.S., G.X., and S.Q.; original draft preparation by Z.H.; review and editing by Y.W. and Y.X.; visualization by Z.H. and Q.C.; supervision by Y.W. and Y.X.; project administration by Y.X.; funding acquisition by Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number T2350010, and the APC was funded by the Key University Laboratory of Metabolism and Health of Guangdong, Southern University of Science and Technology, Shenzhen 518055, China, grant number 2022KSYS007. The Development Project of Jilin Province of China (Grant No. 20220508125RC) also supported this research. Additional support was provided by the project “The Role of Sialic Acid in Tumor Cell Migration”, funded by the National Natural Science Fund of China Research Fund for International Scientists, grant number W2431059.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the reported results can be found at https://drive.google.com/drive/folders/1ik6-qABgVwk_InoXubN9ltpXYVfLDrGG (accessed on 19 January 2025).

Acknowledgments: The senior author thanks Southern University of Science and Technology for its start-up fund and for the support provided by the Center for Computational Science and Engineering at Southern University of Science and Technology. Additionally, we extend our gratitude to Xiaojuan Wu for her assistance in data collection and result visualization.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of this manuscript; or in the decision to publish the results.

References

1. Valous, N.A.; Popp, F.; Zörnig, I.; Jäger, D.; Charoentong, P. Graph machine learning for integrated multi-omics analysis. *Br. J. Cancer* **2024**, *131*, 205–211. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Shen, C. Fast and Scalable Multi-Kernel Encoder Classifier. In Proceedings of the Future Technologies Conference 2024, London, UK, 14–15 November 2024; Volume 1156, pp. 161–177. [\[CrossRef\]](#)
3. Tolles, J.; Meurer, W.J. Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA* **2016**, *316*, 533–534. [\[CrossRef\]](#)
4. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [\[CrossRef\]](#)
5. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf (accessed on 4 December 2017).
6. Wolpert, D.H. What is important about the No Free Lunch theorems? *arXiv* **2020**, arXiv:2007.10928. [\[CrossRef\]](#)
7. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608. [\[CrossRef\]](#)
8. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [\[CrossRef\]](#)
9. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '15, Sydney, Australia, 10–13 August 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1721–1730. [\[CrossRef\]](#)
10. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
11. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *arXiv* **2021**, arXiv:2103.11251. [\[CrossRef\]](#)

12. Cedergreen, N.; Pedersen, K.E.; Fredensborg, B.L. Quantifying synergistic interactions: A meta-analysis of joint effects of chemical and parasitic stressors. *Sci. Rep.* **2023**, *13*, 13641. [[CrossRef](#)]
13. Guo, Y.; Hu, H.; Chen, W.; Yin, H.; Wu, J.; Hsieh, C.-Y.; He, Q.; Cao, J. SynergyX: A multi-modality mutual attention network for interpretable drug synergy prediction. *Brief. Bioinform.* **2024**, *25*, bbae015. [[CrossRef](#)] [[PubMed](#)]
14. He, S.; Chen, H.; Zhu, Z.; Ward, D.G.; Cooper, H.J.; Viant, M.R.; Heath, J.K.; Yao, X. Robust twin boosting for feature selection from high-dimensional omics data with label noise. *Inf. Sci.* **2015**, *291*, 1–18. [[CrossRef](#)]
15. Xiao, G.; Guan, R.; Cao, Y.; Huang, Z.; Xu, Y. KISL: Knowledge-injected semi-supervised learning for biological co-expression network modules. *Front. Genet.* **2023**, *14*, 1151962. [[CrossRef](#)] [[PubMed](#)]
16. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2017**, arXiv:1607.02533. [[CrossRef](#)]
17. Sheng, J.; Lam, S.; Zhang, J.; Zhang, Y.; Cai, J. Multi-omics fusion with soft labeling for enhanced prediction of distant metastasis in nasopharyngeal carcinoma patients after radiotherapy. *Comput. Biol. Med.* **2024**, *168*, 107684. [[CrossRef](#)] [[PubMed](#)]
18. Pérez-González, A.P.; García-Kroepfly, A.L.; Pérez-Fuentes, K.A.; García-Reyes, R.I.; Solis-Roldan, F.F.; Alba-González, J.A.; Hernández-Lemus, E.; de Anda-Jáuregui, G. The ROSMAP project: Aging and neurodegenerative diseases through omic sciences. *Front. Neuroinform.* **2024**, *18*, 1443865. [[CrossRef](#)] [[PubMed](#)]
19. Huang, Z.; Chen, Q.; Mu, X.; An, Z.; Xu, Y. Elucidating the Functional Roles of Long Non-Coding RNAs in Alzheimer's Disease. *Int. J. Mol. Sci.* **2024**, *25*, 9211. [[CrossRef](#)]
20. Mounir, M.; Lucchetta, M.; Silva, T.C.; Olsen, C.; Bontempi, G.; Chen, X.; Noushmehr, H.; Colaprico, A.; Papaleo, E. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.* **2019**, *15*, e1006701. [[CrossRef](#)] [[PubMed](#)]
21. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68–A77. [[CrossRef](#)]
22. Zhan, P.L.; Canavan, M.E.; Ermer, T.; Pichert, M.D.; Li, A.X.; Maduka, R.C.; Kaminski, M.F.; Boffa, D.J. Nonregional Lymph Nodes as the Only Metastatic Site in Stage IV Esophageal Cancer. *JTO Clin. Res. Rep.* **2022**, *3*, 100426. [[CrossRef](#)] [[PubMed](#)]
23. Park, S.-Y.; Nam, J.-S. The force awakens: Metastatic dormant cancer cells. *Exp. Mol. Med.* **2020**, *52*, 569–581. [[CrossRef](#)] [[PubMed](#)]
24. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B.L.; Barrell, D.; Zadissa, A.; Searle, S.; et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **2012**, *22*, 1760–1774. [[CrossRef](#)]
25. Trupp, M.; Altman, T.; Fulcher, C.A.; Caspi, R.; Krummenacker, M.; Paley, S.; Karp, P.D. Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. *Genome Biol.* **2010**, *11* (Suppl. S1), O12. [[CrossRef](#)]
26. Crammer, K.; Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.* **2002**, *2*, 265–292.
27. Ganaie, M.A.; Tanveer, M.; Suganthan, P.N.; Snasel, V. Oblique and rotation double random forest. *Neural Netw.* **2022**, *153*, 496–517. [[CrossRef](#)]
28. Qahtan, A. Machine Learning—Evaluation (Cross-validation, Metrics, Importance Scores...). In *Clinical Applications of Artificial Intelligence in Real-World Data*; Asselbergs, F.W., Denaxas, S., Oberski, D.L., Moore, J.H., Eds.; Springer International Publishing: Cham, Switzerland, 2023; pp. 175–187. [[CrossRef](#)]
29. Riyahi, S.; Dev, H.; Behzadi, A.; Kim, J.; Attari, H.; Raza, S.I.; Margolis, D.J.; Jonisch, A.; Megahed, A.; Bamashmos, A.; et al. Pulmonary Embolism in Hospitalized Patients with COVID-19: A Multicenter Study. *Radiology* **2021**, *301*, E426. [[CrossRef](#)]
30. Yu, T. AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments. *PLoS Comput. Biol.* **2022**, *18*, e1009826. [[CrossRef](#)]
31. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
32. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [[CrossRef](#)] [[PubMed](#)]
33. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [[CrossRef](#)] [[PubMed](#)]
34. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2021**, *2*, 100141. [[CrossRef](#)] [[PubMed](#)]
35. Hanel, R.; Corominas-Murtra, B.; Liu, B.; Thurner, S. Fitting Power-laws in empirical data with estimators that work for all exponents. *PLoS ONE* **2017**, *12*, e0170920. [[CrossRef](#)]
36. Xie, K.; Hou, Y.; Zhou, X. Deep centroid: A general deep cascade classifier for biomedical omics data classification. *Bioinformatics* **2024**, *40*, btac039. [[CrossRef](#)] [[PubMed](#)]
37. Wu, Y.; Chow, K.-H.; Wei, W.; Liu, L. Exploring Model Learning Heterogeneity for Boosting Ensemble Robustness. *arXiv* **2023**, arXiv:2310.02237. [[CrossRef](#)]
38. Huang, X.; Zhang, L.; Wang, B.; Li, F.; Zhang, Z. Feature clustering based support vector machine recursive feature elimination for gene selection. *Appl. Intell.* **2018**, *48*, 594–607. [[CrossRef](#)]

39. Islam, M.R.; Matin, A. Detection of COVID 19 from CT Image by The Novel LeNet-5 CNN Architecture. In Proceedings of the 2020 23rd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 19–21 December 2020; pp. 1–5. [\[CrossRef\]](#)
40. Hu, Q.; Guo, Y.; Xie, X.; Cordy, M.; Ma, L.; Papadakis, M.; Le Traon, Y. Test Optimization in DNN Testing: A Survey. *ACM Trans. Softw. Eng. Methodol.* **2024**, *33*, 1–42. [\[CrossRef\]](#)
41. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [\[CrossRef\]](#)
42. Biswas, A.; Liu, S.; Garg, S.; Morshed, M.G.; Vakili, H.; Ghosh, A.W.; Balachandran, P.V. Integrating adaptive learning with post hoc model explanation and symbolic regression to build interpretable surrogate models. *MRS Commun.* **2024**, *14*, 983–989. [\[CrossRef\]](#)
43. Knezevic, N.N.; Manchikanti, L.; Hirsch, J.A. Principles of Evidence-Based Medicine. In *Essentials of Interventional Techniques in Managing Chronic Pain*; Singh, V., Falco, F.J.E., Kaye, A.D., Soin, A., Hirsch, J.A., Eds.; Springer International Publishing: Cham, Switzerland, 2024; pp. 101–118. [\[CrossRef\]](#)
44. Yang, Z.; Zhang, A.; Sudjianto, A. GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recognit.* **2021**, *120*, 108192. [\[CrossRef\]](#)
45. Marcílio, W.E.; Eler, D.M. From explanations to feature selection: Assessing SHAP values as feature selection mechanism. In Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Porto de Galinhas, Brazil, 7–10 November 2020; pp. 340–347. [\[CrossRef\]](#)
46. van der Knaap, J.A.; Verrijzer, C.P. Undercover: Gene control by metabolites and metabolic enzymes. *Genes Dev.* **2016**, *30*, 2345–2369. [\[CrossRef\]](#) [\[PubMed\]](#)
47. Zhang, Z.; Yao, L.; Yang, J.; Wang, Z.; Du, G. PI3K/Akt and HIF-1 signaling pathway in hypoxia-ischemia. *Mol. Med. Rep.* **2018**, *18*, 3547–3554. [\[CrossRef\]](#)
48. Song, Y.-S.; Annalora, A.J.; Marcus, C.B.; Jefcoate, C.R.; Sorenson, C.M.; Sheibani, N. Cytochrome P450 1B1: A Key Regulator of Ocular Iron Homeostasis and Oxidative Stress. *Cells* **2022**, *11*, 2930. [\[CrossRef\]](#)
49. Miki, Y.; Kidoguchi, Y.; Sato, M.; Taketomi, Y.; Taya, C.; Muramatsu, K.; Gelb, M.H.; Yamamoto, K.; Murakami, M. Dual Roles of Group IID Phospholipase A2 in Inflammation and Cancer. *J. Biol. Chem.* **2016**, *291*, 15588–15601. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Tan, R.; Zhou, Y.; An, Z.; Xu, Y. Cancer Is a Survival Process under Persistent Microenvironmental and Cellular Stresses. *Genom. Proteom. Bioinform.* **2022**, *21*, 1260–1265. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Sun, H.; Zhang, C.; Cao, S.; Sheng, T.; Dong, N.; Xu, Y. Fenton reactions drive nucleotide and ATP syntheses in cancer. *J. Mol. Cell Biol.* **2018**, *10*, 448–459. [\[CrossRef\]](#)
52. Röhrig, F.; Schulze, A. The multifaceted roles of fatty acid synthesis in cancer. *Nat. Rev. Cancer* **2016**, *16*, 732–749. [\[CrossRef\]](#)
53. Li, J.; Lim, J.Y.S.; Eu, J.Q.; Chan, A.K.M.H.; Goh, B.C.; Wang, L.; Wong, A.L.-A. Reactive Oxygen Species Modulation in the Current Landscape of Anticancer Therapies. *Antioxid. Redox Signal.* **2024**, *41*, 322–341. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Sun, H.; Zhou, Y.; Skaro, M.F.; Wu, Y.; Qu, Z.; Mao, F.; Zhao, S.; Xu, Y. Metabolic Reprogramming in Cancer Is Induced to Increase Proton Production. *Cancer Res.* **2020**, *80*, 1143–1155. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Zhou, Y.; Chang, W.; Lu, X.; Wang, J.; Zhang, C.; Xu, Y. Acid–base Homeostasis and Implications to the Phenotypic Behaviors of Cancer. *Genom. Proteom. Bioinform.* **2023**, *21*, 1133–1148. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Górka, A.; Mazur, A.J. Integrin-linked kinase (ILK): The known vs. the unknown and perspectives. *Cell. Mol. Life Sci.* **2022**, *79*, 100. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Feng, S.; Zhou, M.; Huang, Z.; Xiao, X.; Zhong, B. A colorectal liver metastasis prediction model based on the combination of lipoprotein-associated phospholipase A2 and serum biomarker levels. *Clin. Chim. Acta* **2025**, *568*, 120143. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Wang, W.; Liu, F.; Wang, C.; Wang, C.; Tang, Y.; Jiang, Z. Glutathione S-transferase A1 mediates nicotine-induced lung cancer cell metastasis by promoting epithelial-mesenchymal transition. *Exp. Ther. Med.* **2017**, *14*, 1783–1788. [\[CrossRef\]](#) [\[PubMed\]](#)
59. He, W.; Cao, X.; Rong, K.; Chen, X.; Han, S.; Qin, A. Combination of AZD3463 and DZNep Prevents Bone Metastasis of Breast Cancer by Suppressing Akt Signaling. *Front. Pharmacol.* **2021**, *12*, 652071. [\[CrossRef\]](#) [\[PubMed\]](#)
60. Xia, W.; Wang, H.; Zhou, X.; Wang, Y.; Xue, L.; Cao, B.; Song, J. The role of cholesterol metabolism in tumor therapy, from bench to bed. *Front. Pharmacol.* **2023**, *14*, 928821. [\[CrossRef\]](#)
61. Liu, W.; Chakraborty, B.; Safi, R.; Kazmin, D.; Chang, C.; McDonnell, D.P. Dysregulated cholesterol homeostasis results in resistance to ferroptosis increasing tumorigenicity and metastasis in cancer. *Nat. Commun.* **2021**, *12*, 5103. [\[CrossRef\]](#)
62. Giacomini, I.; Gianfanti, F.; Desbats, M.A.; Orso, G.; Berretta, M.; Prayer-Galetti, T.; Ragazzi, E.; Cocetta, V. Cholesterol Metabolic Reprogramming in Cancer and Its Pharmacological Modulation as Therapeutic Strategy. *Front. Oncol.* **2021**, *11*, 682911. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses from a Practical Point of View. *IEEE Access* **2019**, *7*, 154096–154113. [\[CrossRef\]](#)

64. Lee, Y.-G.; Oh, J.-Y.; Kim, D.; Kim, G. SHAP Value-Based Feature Importance Analysis for Short-Term Load Forecasting. *J. Electr. Eng. Technol.* **2023**, *18*, 579–588. [[CrossRef](#)]
65. Bull, C.; Byrne, R.M.; Fisher, N.C.; Corry, S.M.; Amirkhah, R.; Edwards, J.; Hillson, L.V.S.; Lawler, M.; Ryan, A.E.; Lamrock, F.; et al. Dual gene set enrichment analysis (dualGSEA); an R function that enables more robust biological discovery and pre-clinical model alignment from transcriptomics data. *Sci. Rep.* **2024**, *14*, 30202. [[CrossRef](#)] [[PubMed](#)]
66. Verdasco, M.P.; García-Cuesta, E. An Interpretable Rule Creation Method for Black-Box Models based on Surrogate Trees—Srules. *arXiv* **2024**, arXiv:2407.20070. [[CrossRef](#)]
67. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You? In ”: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD ’16, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1135–1144.
68. Crawford, J.; Greene, C.S. Incorporating biological structure into machine learning models in biomedicine. *Curr. Opin. Biotechnol.* **2020**, *63*, 126–134. [[CrossRef](#)]
69. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2014**, arXiv:1312.6199. [[CrossRef](#)]
70. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
71. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
72. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, USA, 2012.
73. Sauber-Cole, R.; Khoshgoftaar, T.M. The use of generative adversarial networks to alleviate class imbalance in tabular data: A survey. *J. Big Data* **2022**, *9*, 98. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.