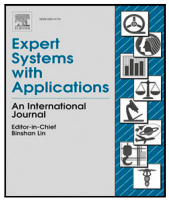
Expert Systems With Applications 248 (2024) 123469

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Evolving pathway activation from cancer gene expression data using nature-inspired ensemble optimization

Xubin Wang [a,b], Yunhe Wang [a,*], Zhiqiang Ma [c], Ka-Chun Wong [d], Xiangtao Li [b]

[a] School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
[b] School of Artificial Intelligence, Jilin University, Changchun, China
[c] School of Information Science and Technology, Northeast Normal University, Changchun, China
[d] Department of Computer Science, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Sampling
Feature selection
Ensemble learning
Ant colony optimization
Class-imbalanced learning

ABSTRACT

Class-imbalanced biological datasets pose significant challenges in machine learning and data analysis tasks. Prior methods to handle imbalance rely on data oversampling, which increases computational costs and overfitting. While feature selection and ensemble learning are promising techniques, current applications in imbalanced contexts are limited. To address these challenges, we present a novel framework called Hybrid Sampling Nature-Inspired Optimization Ensemble (HSNOE) to enhance the identification of hidden responders in imbalanced biological datasets. Our contributions are three-fold: 1) A hybrid undersampling and oversampling technique to mitigate class-imbalance; 2) Integrate an ant colony optimization-based feature selection that identifies informative feature subsets; 3) An ensemble classifier integrating diverse models trained on optimized features to improve performance. The experiments conducted on the five biological datasets demonstrate that HSNOE exhibits more stable comprehensive performance across six evaluation metrics compared to ten benchmark methods. We also conducted a biological analysis specifically on the Pan-cancer dataset. Moreover, the HSNOE method has been made publicly available on GitHub.[1]

1. Introduction

Class imbalance poses a pervasive challenge when applying machine learning techniques to analyze biological datasets. In domains such as cancer genomics and precision medicine, these datasets often contain a minority of samples belonging to clinically important subtypes, commonly referred to as "hidden responders". While recent work has shown promise in identifying hidden responders (Li, Li, Wang, Zhang, & Wong, 2021; Way et al., 2018), class imbalance remains a major barrier to advancing precision oncology applications (Prasad, 2016). Biological datasets typically exhibit highly skewed class distributions where rare responding samples are vastly underrepresented compared to non-responders. This class imbalance adversely affects the performance of current machine learning algorithms, impeding the identification of informative biomarkers and limiting prediction capabilities.

The imbalance bias is especially detrimental for biological tasks where detecting small responding subgroups is critical. Oversampling techniques that replicate minority samples are commonly used due to simplicity but prone to overfitting limited responding sample sizes (Rahman, Hassan, & Ahad, 2021). Models trained on replicated minorities often fail to generalize. Meanwhile, undersampling risks eliminating information providing insights into resistance or aiding biomarker discovery (Rahman et al., 2021). This lost information represents missed opportunities to characterize and predict differential responses. Additionally, single models intrinsically favor majority performance during training while lacking diversity to adequately model skewed distributions (Brown, Wyatt, Harris, & Yao, 2005). They become disproportionately biased towards dominant non-responding classes, obstructing detection of critical rare responders. Feature selection stability and reproducibility also diminishes on imbalanced data as biomarker identification heavily relies on predominant samples, potentially overlooking key features associated with minority response.

To overcome these limitations, ensemble approaches have emerged as a promising paradigm for learning from class imbalanced biomedical data (Gupta & Gupta, 2022; Rahman et al., 2021). Ensemble

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: wangxb19@mails.jlu.edu.cn (X. Wang), wangyh082@hebut.edu.cn (Y. Wang), mazq@nenu.edu.cn (Z. Ma), kc.w@cityu.edu.hk (K.-C. Wong), lixt314@jlu.edu.cn (X. Li).

[1] https://github.com/wangxb96/HSNOE

https://doi.org/10.1016/j.eswa.2024.123469
Received 2 January 2023; Received in revised form 9 February 2024; Accepted 10 February 2024
Available online 12 February 2024

methods integrate multiple diverse base learners, leveraging their complementary strengths to improve overall performance across skewed distributions (Zhang & Ma, 2012). Advanced sampling techniques can be incorporated to retain information from both majority and minority samples. Hybrid oversampling and undersampling procedures prevent overfitting while minimizing lost information. Diversified feature selection stabilizes identification of biomarkers linked to rare treatment response (Gao, Bian, Wang, Li, & Wang, 2022; Wang & Jia, 2022). Optimized ensemble architectures can account for class imbalance during training to enhance identification of critical responding subgroups for precision oncology. In summary, specialized ensemble frameworks customized for biological class imbalance hold strong potential for improving detection of the most clinically valuable responding minorities within heterogeneous tumors.

This paper proposes a hybrid sampling nature-inspired optimization ensemble (HSNOE) framework. Specifically, we integrate various methods that address class-imbalanced learning problems, such as sampling, feature selection, and ensemble learning, to improve the identification performance. To prevent a situation where the model cannot capture the characteristics of the class-imbalanced data, we use a hybrid sampling method that combines the neighborhood cleaning rule (NCL) (Laurikkala, 2001) and synthetic minority over-sampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to balance the original data. After this, we synergize a nature-inspired approach with the ensemble method in a cooperative manner to optimize the sample features and generate the diverse optimized ensemble. We conducted extensive experiments on five biological datasets to evaluate the performance of HSNOE compared to ten benchmark methods. The results demonstrate that HSNOE is a highly competitive method in handling class imbalance and outperforms the benchmark methods. Furthermore, we present a detailed biological analysis of HSNOE, specifically focusing on its application to a Pan-cancer dataset. This analysis provides valuable insights into the effectiveness of HSNOE in identifying hidden responders and uncovering related pathways within the context of cancer research. By employing HSNOE, we not only improve the identification of hidden responders, but also gain a deeper understanding of the underlying biological mechanisms and pathways associated with cancer.

The main contributions of the proposed HSNOE method can be summarized as follows:

1. The proposed hybrid sampling technique combines NCL undersampling and SMOTE oversampling to balance biological data and mitigate the impact of class imbalance.
2. The application of an evolutionary algorithm for nature-inspired feature selection helps identify the most informative subset of features from high-dimensional biological data. This approach reduces model complexity, enhances interpretability.
3. The proposed ensemble framework integrates multiple diverse classifiers trained on optimized feature sets to enhance prediction performance. This contributes to improved performance in detecting small hidden responder subgroups.

Overall, HSNOE provides a unique combination of techniques to address the challenges of class imbalance, feature selection, and ensemble learning in biological data analysis, with a focus on identifying hidden responders. Additionally, we have consolidated the main abbreviations used in this paper in Table 1. The remainder of this paper is organized as follows: Section 2 provides a detailed description of the proposed HSNOE method, including the hybrid sampling technique, nature-inspired feature selection, and nature-inspired ensemble learning. Section 3 presents the datasets, baselines, evaluation metrics and experimental setup used to assess HSNOE's performance. The obtained results and their comprehensive analysis are presented in Section 4, highlighting the superiority of HSNOE. We conduct a case study and biological analysis on a Pan-cancer data in Section 5. In Section 6, we discuss the findings, including the limitations, and future directions of our approach. Finally, Section 7 summarizes the main contributions and emphasizes the significance of HSNOE in identifying hidden responders.

Table 1
List of abbreviations.

Acronym	Explanation
ACO	Ant Colony Optimization
ANN	Artificial Neural Network
AUROC	Area Under the Receiver Operating Characteristic curve
AUPRC	Area Under the Precision-Recall curve
DT	Decision Tree
G-mean	Geometric Mean
HSNOE	Hybrid Sampling Nature-inspired Optimization Ensemble
KNN	K-nearest Neighbor
ML	Machine Learning
NB	Naïve Bayes
NCL	Neighborhood Cleaning Rule
PPI	Protein-protein Interaction Network
RF	Random Forest
SMOTE	Synthetic Minority Over-sampling TEchnique
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas

2. Methods

2.1. Methodology overview of HSNOE

In this study, we propose a novel method called Hybrid Sampling Nature-Inspired Optimization Ensemble (HSNOE) to tackle the challenges of class imbalance in biological data analysis. We evaluate HSNOE on biological datasets denoted as D, which consist of n samples. Each sample d_i is represented by a feature vector containing f features, along with an associated subtype label y. The HSNOE model consists of three main components: hybrid sampling, nature-inspired feature selection, and nature-inspired ensemble learning.

In the hybrid sampling phase, we initially split the original data D into a training set D_1 and a test set D_2 using a 9:1 ratio. To address class imbalance, we employ the neighborhood cleaning rule (NCL), a data cleaning technique, to eliminate noisy samples from the training set D_1. Additionally, we utilize the synthetic minority oversampling technique (SMOTE), an oversampling method, to augment the representation of minority class samples. The resulting dataset is denoted as the training set X. Next, we introduce a nature-inspired feature selection method that combines ant colony optimization (ACO) with artificial neural networks (ANNs) to identify informative features from the training set X. The ACO approach employs ANN as an evaluator for the feature subset, optimizing the selection process. In the nature-inspired ensemble learning phase, we construct a set of base classifiers B, consisting of K-nearest neighbor (KNN), decision tree (DT), discriminant analysis (DISCR), Naïve Bayes (NB), and artificial neural networks (ANN). To ensure diversity among the base classifiers, we employ a diverse subspace generation method, which generates different subspaces. Subsequently, ACO is utilized to optimize the ensemble from the base classifiers B, resulting in the creation of the final ensemble model Ψ.

Finally, we evaluate the performance of the test set D_2 by employing plurality voting based on the predictions of the ensemble model Ψ. This approach combines the predictions of multiple base classifiers to make a collective decision. The overall algorithm for HSNOE is summarized in Algorithm 1. The primary objectives of HSNOE are to enhance classification performance in biological data analysis and provide insights into the underlying biological mechanisms. To achieve these goals, HSNOE implements a multi-step framework, as depicted in Fig. 1. By integrating hybrid sampling, feature optimization, and ensemble modeling, HSNOE aims to address current challenges in class-imbalanced biological data analysis. The multi-pronged framework strives to improve rare class identification, biomarker discovery, and predictive capabilities on skewed real-world datasets.

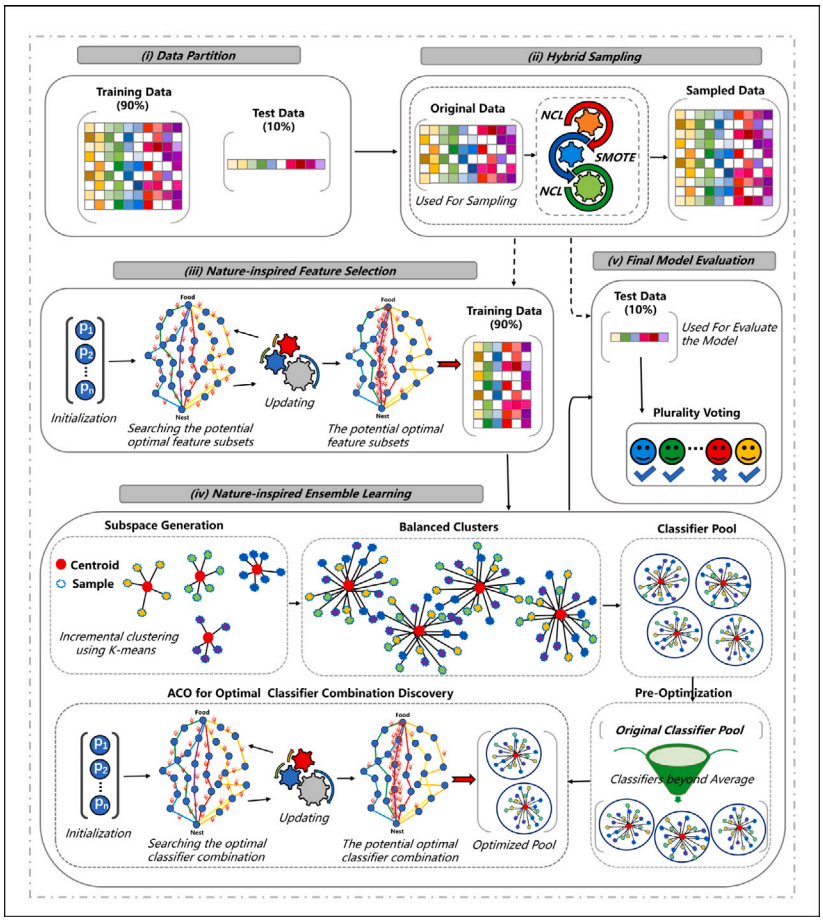


Fig. 1. The framework of the proposed HSNOE model. It consists of four main phases: In Phase 1, the original data is pre-processed through random splitting into training and test sets at a ratio of 9:1. Oversampling and undersampling techniques are then applied to balance the classes. Phase 2 employs an ACO-based nature-inspired feature selection method to identify optimal feature sets. Phase 3 trains multiple classification models on the selected feature subsets from the previous phase. An ACO-based nature-inspired ensemble learning approach is utilized to select optimal model sets. Finally, in Phase 4, a plurality voting scheme fuses the predictions from the various models selected in Phase 3 to determine the final class prediction.

2.2. Hybrid sampling

To address the issue of class imbalance in the dataset, we employed the neighbor cleaning rule (NCL) along with the synthetic minority oversampling technique (SMOTE) to clean and oversample the minority class data. First, we divided the original data D into a training set D_1 and a test set D_2 at a ratio of 9:1. Here, we use the training data D_1 as an example to show our hybrid sampling process. Initially, the dataset D_1 is partitioned into two subsets G and O, where G contains all samples belonging to the minority class, and the remaining samples constitute O. The ENN rule (Wilson & Martinez, 2000) is applied to O to identify and remove noisy samples, resulting in a new dataset X_1'. Subsequently, SMOTE is applied to oversample the minority class data by creating synthetic samples based on the K-nearest neighbors (K = 5) of each sample in the minority class. The number of synthetic samples generated for each minority class sample is determined by the sampling multiplicity N based on the imbalanced ratio.

Specifically, for each minority class sample x_i in X_1', its K-nearest minority class neighbors are identified, and several samples are randomly selected. For each selected sample x_j, a new sample x_{new} is generated using the following formula:

$$x_{new} = x_i + rand(0,1) * (x_j - x_i). \tag{1}$$

where $rand(0,1)$ is a random number between 0 and 1. This process results in an oversampled dataset X_2'. Despite applying NCL in the hybrid sampling step, some low-quality samples may still exist in the oversampled dataset. These samples can adversely affect classification performance. Therefore, the NCL method is employed again on X_2' to remove noisy samples, resulting in the processed dataset X.

The NCL method followed by SMOTE is used in this study to address class imbalance in the dataset and improve the performance of the classification model. The authors carefully describe the data preprocessing steps, which involves cleaning and oversampling the minority class data to generate a balanced dataset. In particular, we utilized the *imbalanced-learn package* (Lemaître, Nogueira, & Aridas, 2017) to generate Fig. 2 as an example, where the initial sample size was 80 and the class imbalance ratio was 9:1. We found that after using NCL, the number of noisy samples decreased. With the use of SMOTE, the number of minority class samples increased, thereby alleviating the class imbalance situation.

2.3. Nature-inspired feature selection

The training data is denoted as $X = \{(x_1, y_1), \ldots, (x_{n'}, y_{n'})\}$, where $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,f})$ represents the feature vector with f denoting the number of features. The label y belongs to the set $\{1, 2, \ldots, c\}$, where c represents the different subtypes. However, biological data often contain only a few highly related features, necessitating the use of feature selection methods for further analysis (Qu et al., 2021). In this study, we propose the incorporation of ant colony optimization (ACO) with artificial neural networks (ANN) for feature selection. The feature selection process consists of three critical components: population initialization, ACO search, and the objective function (see Fig. 3).

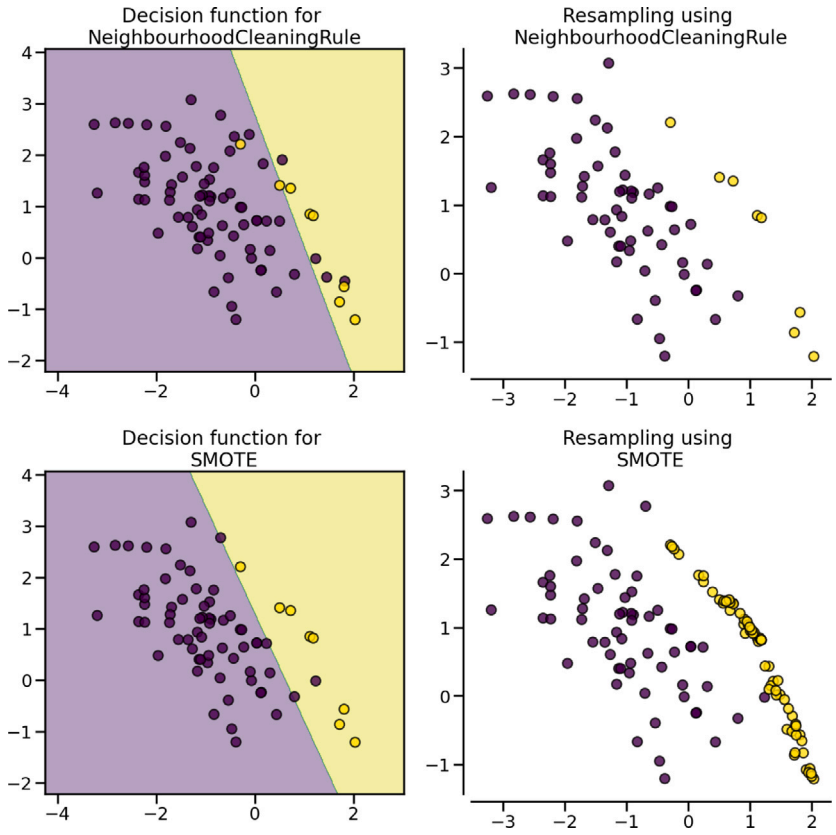


Fig. 2. An example of 80 samples with an imbalanced ratio of 9:1 was used on our hybrid sampling. This involved using neighborhood cleaning rule (NCL) to reduce noisy samples and SMOTE to increase the number of minority class samples, effectively alleviating the class imbalance situation.

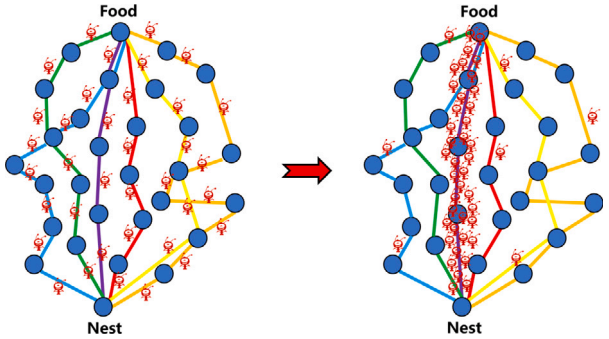


Fig. 3. The schematic diagram of ant colony optimization (ACO). It visually represents the iterative process where artificial ants construct solutions by probabilistically selecting edges based on pheromone trails and heuristic information, while updating the trails according to solution quality, ultimately guiding the search towards optimal solutions.

2.3.1. Population initialization

An ACO population Pop with $|Pop|$ individuals $P = \{p^1, p^2, \ldots, p^{|Pop|}\}$ is first initialized randomly with real numbers. For an individual p^k, it can be depicted as follows:

$$p^k = \{g_1, g_2, \ldots, g_f\}, \tag{2}$$

where g_f denotes the fth feature and f is the number of features. In general, the kth ant travels from feature i to feature j in a stochastic manner with the probability shown below:

$$p_{ij}^k = \begin{cases} \frac{(\tau_{ij}^{\alpha})(\eta_{ij}^{\beta})}{\sum_{l \in J_i^k}(\tau_{il}^{\alpha})(\eta_{il}^{\beta})}, & if\ j \in J_i^k \\ 0, & otherwise, \end{cases} \tag{3}$$

where τ_{ij} denotes the amount of virtual pheromone for transition from state i to j, η_{ij} denotes the heuristic desirability of its state transition at feature i to feature j, $\alpha > 0$ and $\beta > 0$ are two parameters that control the influence of τ_{ij} and η_{ij}, respectively, J_i^k denotes the neighbor features of feature i that allowed a visit by the ant k.

2.3.2. ACO search

The ACO population then searches for the optimal feature subsets in the feature space by pheromone updates. In each generation, the best individual is saved after the ACO search. Specifically, the pheromone updating rule is shown in the following equations (Aghdam, Ghasem-Aghaee, & Basiri, 2009):

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \sum_{k=1}^{m} \Delta\tau_{ij}^k + \Delta\tau_{ij}^g \tag{4}$$

$$\Delta\tau_{ij}^k = \begin{cases} \phi \cdot \gamma(S^k) + \frac{(1-\phi)\cdot(n-|S^k|)}{n}, & if\ i \in S^k, \\ 0 \qquad otherwise \end{cases} \tag{5}$$

where ρ denotes the pheromone evaporation coefficient, m denotes the number of ants, S^k denotes the feature subset found by ant k, $\gamma(S^k)$ denotes the measure of the classifier performance, and $|S^k|$ denotes the size of S^k, $\phi \in [0, 1]$ is the parameter that controls the relative weight.

2.3.3. Feature selection objective function

While classification accuracy has traditionally been the primary focus of model evaluation, the ability to discover meaningful predictive features is equally important. Biological datasets often present in high dimensions with ubiquitous redundant or irrelevant attributes. To address this, we optimize a multi-objective function that considers both predictive power and parsimony of the selected feature set.

Specifically, our objective aims to minimize two goals: maximizing classification performance via metrics like area under the ROC

Algorithm 1 HSNOE Algorithm

Input: Original Data: $D = \{(d_1, y_1), ..., (d_n, y_n)\}$ $d \in R^d$, $y \in \{1, 2, ..., c\}$, a set of base classifiers B, upper bounds of cluster K, population of ACO (Pop), evaluation times T, the feature selection function f_1, and the classifier optimization function f_2

Output: The selected features F' and the evaluation performance

D_1, D_2 ←Partition original data D into training and test sets at a 9:1 ratio
X ←Using hybrid sampling to process D_1
Initialize a population of |Pop| individuals
while ($t <= T$) **do**
  Pop ←f_1(Pop)
  p ←best individual in Pop
  Update t;
end while
F' ←informative features from p
X ←$X(F')$
D_2 ←$D_2(F')$
$X = X_1 \cup ... \cup X_5, X_i \cap X_j = \emptyset (i \neq j)$
for each X_i in X **do**
  $X^{-i} = X - X_i$
  for $k = 1 \rightarrow K$ **do**
    C^S ←partition X^{-i} into k clusters
    $S = S + 1$
    C^S ←balance clusters C^S
  end for
  CP ←Train classifiers on C^S by B
  for each cp_i in CP **do**
    AUROC(i) ←calculate each cp_i's AUROC of X_i
  end for
  CP ←(cp_i if AUROC(cp_i) > mean(AUROC))
  Initialize a population of |Pop| individuals
  while ($t <= T$) **do**
    Pop ←f_2(Pop)
    p ←best individual in Pop
    Update t;
  end while
  ψ ←select classifiers in CP from p
  Optimized model Ψ ←ψ
end for
Evaluation performance←classify samples of D_2 by Ψ
Return Informative features F', evaluation performance

curve, while simultaneously minimizing the number of retained features. This balanced approach guides the algorithm to identify the most informative features, yielding a optimized representation of the underlying biological signal. By coupling predictive ability with dimensionality reduction, our framework is designed not only to achieve robust classification, but also to provide contextually relevant feature selection. Intuitively, the subset of features most strongly linked to targets of interest are preferentially preserved. Therefore, the proposed multi-factorial objective encourages discovery of the most determining attributes for downstream biological insights and therapeutic development. Specifically, the objective function f_1 is depicted as follows:

$$f_1 = \zeta * \sigma + (1 - \zeta) * \frac{fnum}{f}, \tag{6}$$

where ζ is the parameter that controls the importance of two goals, σ denotes the area above the ROC curve (1 - AUROC), f is the number of features and $fnum$ denotes the number of selected features in the evolution.

2.4. Nature-inspired diverse ensemble learning

In this section, a nature-inspired diverse ensemble learning method is proposed to enhance the identification ability of the HSNOE and includes three important components: diverse classifier pool generation, classifier pool optimization, and ensemble prediction.

2.4.1. Diverse classifier pool generation

In the training phase, we use fivefold cross-validation to evaluate the performance. Specifically, the training set X can be depicted as $X = X_1 \cup \cdots \cup X_5, X_i \cap X_j = \emptyset (i \neq j, i, j \in [1, 5])$. For each input biological data $X_i = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, where $y \in \{1, 2, \ldots, c\}$, c is the class of data and m is the number of input samples. K-means (MacKay & Mac Kay, 2003) is utilized to perform stepwise clustering of data X_i, with the parameter k increasing from 1 to t. t denotes the number of classes in data X_i. Specifically, in each K-means clustering, the clusters are obtained by minimizing the following function:

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{v \in S_i} \|v - \mu_i\|^2, \tag{7}$$

where v denotes the feature vector and μ_i denotes the centroid of cluster S_i. After that, a random subspace containing all the clusters is generated. However, these data clusters in the subspace are class pure. Therefore, it is necessary to balance the clusters to achieve an unbiased result.

Assuming there are s clusters in the subspace, the relation between these clusters can be depicted as $C^1 \cap C^2 \cap ... \cap C^s = \emptyset$ and $C^1 \cup C^2 \cup \cdots \cup C^s = X_i$. Here, each cluster represents a unique class. Then, we balance the cluster by adding samples from the other classes. Specifically, if C^i has l samples and centroid c^i, we first calculate the normalized Euclidean distance from the samples of each other class to the centroid c^i of C^i and then add the l nearest samples from each class to the cluster C^i. The same operation is repeated until all clusters in the subspace are balanced.

Then, five structurally different classifiers, including K-nearest neighbor (KNN), decision tree (DT), discriminant analysis (DISCR), naïve Bayes (NB), and artificial neural networks (ANN), are trained by these clusters in the subspace. After that, all trained classifiers are placed into the classifier pool, CP. Furthermore, a nature-inspired classifier pool optimization method is proposed to optimize the classifier pool CP.

2.4.2. Classifier pool optimization

In the classifier pool optimization phase, we preoptimize the classifier pool. Specifically, those classifiers with below-average performance are removed from the classifier pool CP. Then, an ACO-based nature-inspired method is employed to optimize the preoptimized classifier pool. In detail, an ACO population is first initialized, and each individual in the population can be denoted as $p = \{cp_1, cp_2, \ldots, cp_r\}$. Here, cp_r denotes the classifier in CP, and r is the number of classifiers in CP. After that, the ACO population starts searching for the optimal solution. Specifically, the ant colony search and pheromone update approach follows the steps in nature-inspired feature selection. For the objective function f_2, two goals are maintained to achieve optimization and balance, which can be depicted as follows:

$$f_2 = \zeta * \sigma + (1 - \zeta) * \frac{|\psi|}{r}, \tag{8}$$

where ζ is the same setting as in the nature-inspired feature selection to control the importance of the two goals, σ denotes the area above the ROC curve (1 - AUROC), $|\psi|$ denotes the number of selected classifiers ψ and r is the number of classifiers in CP.

2.4.3. Ensemble prediction

Ensemble techniques that aggregate predictions from multiple models are widely used to boost performance. Plurality voting is a simple yet powerful fusion approach wherein the class receiving the most votes across the ensemble is selected as the overall prediction. In our method, we first employ this process to construct a meta-learner (Ψ) during training. Specifically, classifiers (ψ) in the candidate pool are evaluated on validation data, after which the top performers are retained to compose Ψ. At testing, Ψ is directly applied to new samples. Each constituent classifier independently generates predictions, which are then combined using plurality voting. This vote-based fusion has key advantages. First, classification errors from individual models tend to cancel out, resulting in a more robust collective decision. Additionally, Ψ can be deployed as a single optimized unit for inference, avoiding retraining costs.

2.5. Time complexity analysis

In this section, we will analyze the time complexity of the proposed algorithm, considering three main parts: hybrid sampling, nature-inspired feature selection, and nature-inspired ensemble learning. The detailed analysis is as follows:

1. Hybrid Sampling:
   - NCL costs $O(n^2)$, where n is the number of samples in the training dataset.
   - SMOTE costs $O(n \log n)$, where n is the number of samples.
2. Feature Selection:
   - The time complexity of feature selection is $O(N \times n \times T)$, where N is the size of the population in the ACO algorithm, n is the number of samples in the training dataset, and T is the number of iterations.
3. Ensemble Learning:
   - K-means clustering has a time complexity of $O(K \times n \times I)$, where K is the number of clusters, n is the number of input samples, and I is the number of iterations needed for convergence.
   - The number of clusters increases from 1 to t. Thus, a total of $t \times (t+1)/2$ clusters are generated. Therefore, the time complexity for clustering is $O((t \times (t+1)/2) \times n \times I)$.
   - The cluster balancing stage costs $O(l \times v)$, where l is the number of samples in the original cluster, and v is the number of total classes after cluster balancing.
   - The balanced clusters are then trained by the base classifiers in B. The time complexity for training the base classifiers is $O((t \times (t+1)/2) \times n^2)$ in the worst-case scenario.
   - For optimization, the pre-optimization step costs $O(n)$, and the nature-inspired classifier optimization costs $O(N \times n \times T)$, where N is the population size, n is the number of input samples, and T is the number of iteration times.

In summary, the total time complexity can be analyzed as follows: $O(n^2) + O(n \log n) + O(N \times n \times T) + O((t \times (t+1)/2) \times n \times I) + O(l \times v) + O((t \times (t+1)/2) \times n^2) + O(N \times n \times T) = O(n^3)$.

3. Implementation

3.1. Data collection

In this study, we conducted experiments using five class-imbalanced biological datasets with imbalanced ratios ranging from 8.6 to 28. The ecoli, sick_euthyroid, yeast_ml8, arrhythmia, and yeast_me2 datasets were sourced from Ref. Ding (2011). Detailed information about each dataset can be found in Table 2.

Table 2
Five imbalanced biological datasets used in the study.

Dataset	Samples	Features	IR
ecoli	336	7	8.6
sick_euthyroid	3,163	42	9.8
yeast_ml8	2,417	103	13
arrhythmia	452	278	17
yeast_me2	1,484	8	28

Moreover, we also used a cancer gene expression dataset called Pan-cancer (Li et al., 2021) for further evaluation. The Pan-cancer gene expression profiles were sourced from The Cancer Genome Atlas (TCGA) PanCanAtlas project. This large-scale effort integrated multi-center mutation calls, Illumina RNAseq data, and GISTIC2.0 copy number thresholds across 16 prevalent cancer types, totaling 4759 patient samples assayed for 20,486 genetic features. To discern significantly cancer-associated genes, the binary label y was designated as 1 if a gene exhibited strong associations with mutations according to integrative analyses, or 0 otherwise.

3.2. Metrics

In this study, six metrics including accuracy, AUROC (Area Under the Receiver Operating Characteristic curve), AUPRC (Area Under the Precision-Recall curve), F1-score, G-mean and their average were used to evaluate the performance of the algorithm on imbalanced datasets.

Accuracy: The proportion of correct predictions made by the model over the total number of predictions made.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

AUROC: It measures the ability of the model to distinguish between positive and negative samples and is commonly used in binary classification problems. The area under the ROC curve (AUROC) is obtained from the correlation between the true-positive rate (TPR) and false-positive rate (FPR) as follows:

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

$$FPR = \frac{FP}{TN + FP}, \tag{11}$$

AUPRC: It measures the trade-off between precision and recall and is useful in situations where the dataset is highly imbalanced. The area under the Precision-Recall curve (AUPRC) is obtained from the precision against the recall as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN}, \tag{13}$$

F1-score: It is a commonly used metric that balances both precision and recall, and is useful in imbalanced datasets, which can be defined as follows:

$$F1\text{-}score = \frac{2 * (precision * recall)}{precision + recall} \tag{14}$$

G-mean: It is a metric that takes into account both the true positive rate and the true negative rate and is useful in situations where the dataset is highly imbalanced.

$$G\text{-}mean = \sqrt{TPR * TNR} \tag{15}$$

$$TNR = \frac{TN}{TN + FP} \tag{16}$$

where TNR is the true negative rate (specificity).

Average: Mathematically, the average is calculated as the sum of all the individual metric values divided by the total number of metrics. In the context of the study, the average can be defined as:

$$Average = (Accuracy + AU\,ROC + AU\,PRC + F1\text{-}score + G\text{-}mean)/5 \quad (17)$$

These metrics are important in evaluating the performance of algorithms on imbalanced datasets, where the classes are not equally represented. By using a combination of these metrics, we can obtain a more comprehensive assessment of the algorithm's performance and its ability to accurately identify and classify cancer samples.

3.3. Baselines

In our experiments, we employed ten baseline methods to evaluate the performance of our proposed approach. These methods are widely used in the field of machine learning and have been extensively studied in various applications.

The first set of baseline methods includes K-nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), and Random Forest (RF). KNN is a non-parametric algorithm that classifies instances based on their proximity to neighboring instances. SVM, on the other hand, is a supervised learning algorithm that constructs hyperplanes to separate different classes in a high-dimensional space. DT is a hierarchical model that recursively partitions the feature space based on decision rules. NB is a probabilistic classifier that assumes independence among features. RF is an ensemble learning method that combines multiple decision trees to make predictions.

The second set of baseline methods comprises Multi-class Support Vector Machine (MSVM), Ensemble Tree (ET), Artificial neural network (ANN), Generalized Regression Neural Network (GRNN), and Probabilistic Neural Network (PNN). MSVM is an extension of SVM designed for multi-class classification problems. ET is an ensemble method that combines predictions from multiple decision trees. ANN is a computational model inspired by the structure and function of biological neural networks, capable of learning complex relationships between inputs and outputs. GRNN is a type of neural network that uses radial basis functions for regression tasks. PNN is a feedforward neural network with a specific architecture that captures the probability distribution of the training data.

These baseline methods provide a comprehensive comparison framework for evaluating the effectiveness and efficiency of our proposed approach in handling the class-imbalanced biological datasets. By leveraging the strengths and characteristics of these methods, we can assess the performance of our approach and gain insights into its strengths and limitations in comparison to established techniques.

3.4. Parameter settings

A rigorous 5-fold cross-validation strategy was implemented for all model training and optimization procedures in this study. The base classifiers in HSNOE were configured with specific parameters as follows: KNN utilized 5 neighbors, Discriminant analysis employed the diagonal linear function, Naïve Bayes adopted a kernel distribution. The remaining parameters were kept at their default settings. Both the nature-inspired feature selection and ensemble construction stages in HSNOE utilized the same configuration for the ACO algorithm. A population size of 100 individuals was evolved over 50 iterations. The selection threshold was set at 0.5. The balance coefficient (ζ) governing the multi-objective ACO optimization was maintained at 0.9 for both stages. To enhance the stability of the results, each experiment was independently repeated 10 times, and the average performance was reported as the final outcome. This repetition helps mitigate the impact of random variations and provides a more reliable evaluation of the algorithm's performance.

Table 3
Performance comparison of classification algorithms on ecoli dataset.

Methods	Accuracy	AUPRC	AUROC	F1-score	G-mean	Average
KNN	0.9244	0.5000	0.7634	0.9582	0.9582	0.8208
NB	0.8958	0.5000	0.5000	0.9451	0.9465	0.7575
MSVM	0.9244	0.5000	0.7634	0.9582	0.9582	0.8208
SVM	0.8952	0.5000	0.5249	0.9444	0.9455	0.7620
DT	0.9113	0.5000	0.7359	0.9508	0.9509	0.8098
RF	0.9298	0.5000	0.7437	0.9615	0.9616	0.8193
ET	0.9345	0.5000	0.7665	0.9640	0.9641	0.8258
GRNN	0.9091	0.9091	0.5000	–	0.0000	0.5795
ANN	0.9545	0.9744	0.8625	0.7476	0.8513	0.8781
PNN	0.9394	0.9815	0.8917	0.7208	0.8835	0.8834
HSNOE	**0.9932**	**0.9901**	**0.9797**	**0.9846**	**0.9900**	**0.9875**

4. Results and analysis

Table 3 compares the performance of 11 different classification algorithms (KNN, NB, MSVM, SVM, DT, RF, ET, GRNN, ANN, PNN, HSNOE) on the ecoli dataset based on 6 evaluation metrics — Accuracy, AUROC, AUPRC, F1-score, G-mean and their Average. We can see that the traditional algorithms like KNN, NB, MSVM and SVM perform reasonably well with accuracy around 0.9 but have lower scores on other metrics. Decision tree based algorithms like DT, RF and ET improve the average score to around 0.8 but their AUROC/AUPRC is only 0.5. Neural network models ANN and PNN achieve good scores overall (>0.85) indicating their stronger representational ability compared to other algorithms. The proposed HSNOE methodology outperforms all other algorithms significantly across all evaluation metrics. It achieves near perfect scores of >0.99 for Accuracy, AUROC, AUPRC and F1 while the second best scores are 0.95–0.98. This clearly demonstrates the effectiveness of HSNOE in handling class imbalance present in the ecoli dataset through its hybrid approach of sampling, feature selection and ensemble modeling. The results validate that it can improve rare class identification and provide more robust and balanced classification compared to other state-of-the-art algorithms.

Fig. 4 presents a performance comparison of various classification algorithms on the sick_euthyroid dataset. From the figure we can see that traditional algorithms such as KNN, NB, MSVM, and SVM achieve relatively high F1 and G-mean scores of around 0.95. However, their performance on other metrics, such as AUROC and AUPRC, is lackluster. These algorithms may struggle to handle the class imbalance in the dataset, leading to suboptimal results. Decision tree-based algorithms, namely DT, RF, and ET, stand out in terms of performance. Among them, RF achieves the highest average score of 0.8734 and the best AUROC of 0.9180. These results indicate that decision tree algorithms are robust and perform well even in the presence of class imbalance in the sick_euthyroid dataset. Neural network models, including GRNN, ANN, and PNN, obtain high AUPRC scores (>0.9), suggesting their potential to capture complex patterns in the data. However, their performance on other metrics is mediocre and highly variable, indicating a tendency towards overfitting. Further tuning and regularization techniques may be needed to improve the generalization ability of these models. The proposed HSNOE algorithm achieves competitive scores, outperforming neural network models in most metrics. However, it does not surpass the top-performing tree algorithms. This observation may be attributed to the sick_euthyroid dataset exhibiting less severe class imbalance compared to other datasets where HSNOE excelled. Nonetheless, HSNOE demonstrates balanced performance, validating its flexibility and effectiveness in handling imbalanced datasets.

Fig. 5 compares 11 algorithms on the yeast_ml8 imbalanced dataset involving subcellular location prediction. Traditional algorithms KNN and SVM-based models achieve good balanced accuracy (>0.92), indicating applicability to such classification problems. However, NB underperforms due to intrinsic assumptions, highlighting importance of model selection. Decision trees show decent performance but are surpassed by kernel methods, demonstrating potential limitations. Neural

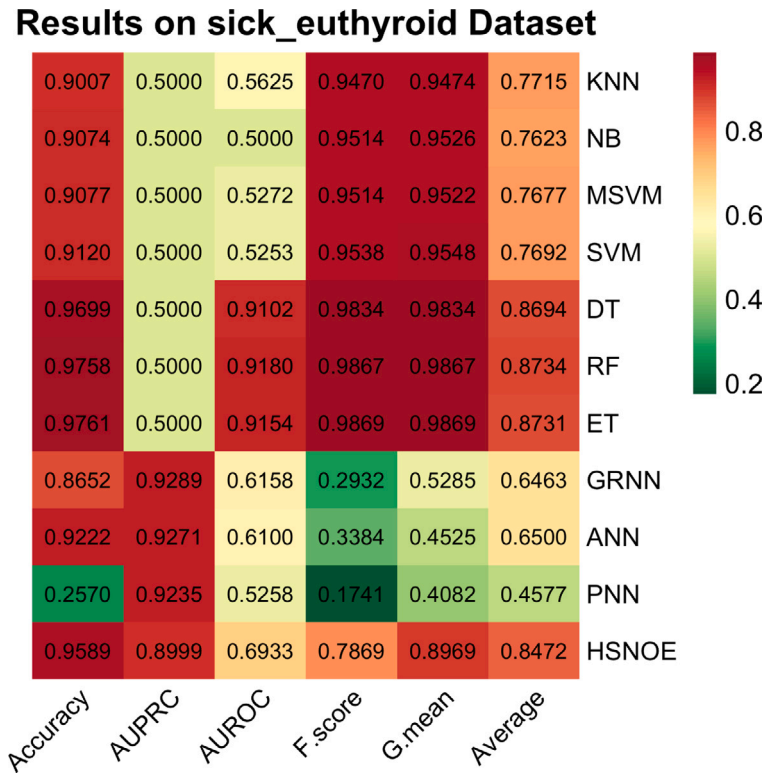
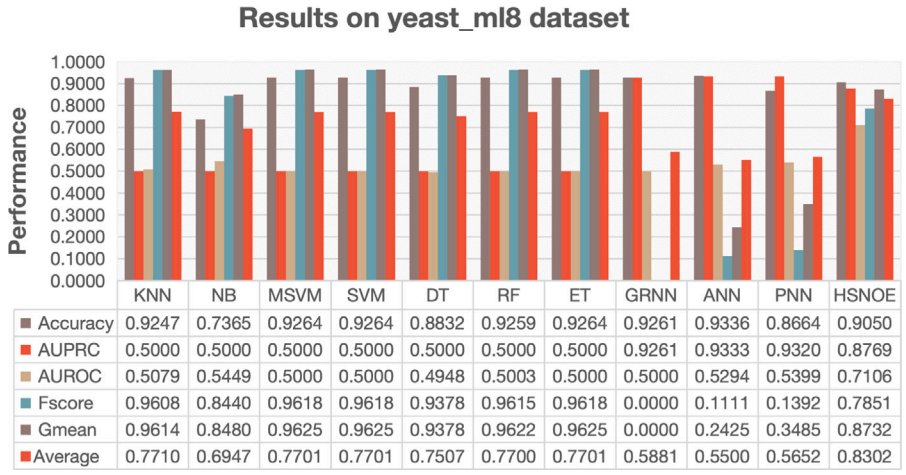


Fig. 4. Performance comparison of KNN, NB, MSVM, SVM, DT, RF, ET, GRNN, ANN, PNN, and HSNOE on the sick_euthyroid dataset based on metrics including Accuracy, AUROC, AUPRC, F1-score, G-mean, and their Average. In the heatmap, the redder the color, the better the performance, while the greener the color, the poorer the performance.

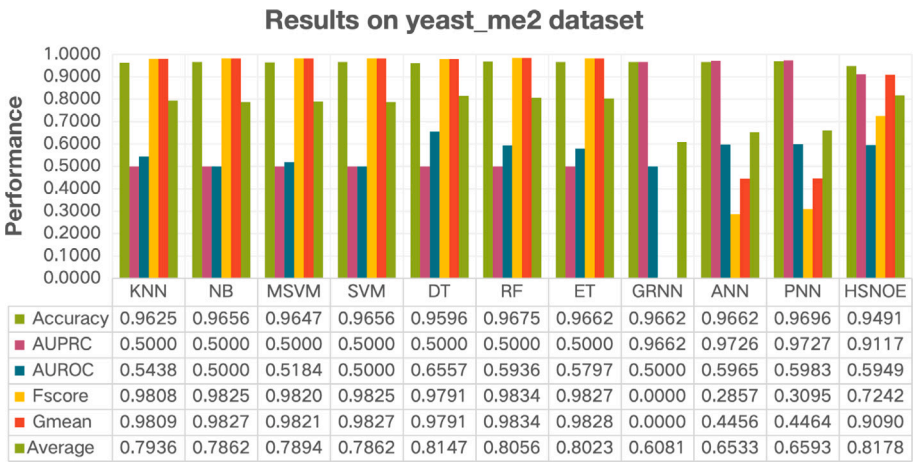


	KNN	NB	MSVM	SVM	DT	RF	ET	GRNN	ANN	PNN	HSNOE
Accuracy	0.9247	0.7365	0.9264	0.9264	0.8832	0.9259	0.9264	0.9261	0.9336	0.8664	0.9050
AUPRC	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.9261	0.9333	0.9320	0.8769
AUROC	0.5079	0.5449	0.5000	0.5000	0.4948	0.5003	0.5000	0.5000	0.5294	0.5399	0.7106
Fscore	0.9608	0.8440	0.9618	0.9618	0.9378	0.9615	0.9618	0.0000	0.1111	0.1392	0.7851
Gmean	0.9614	0.8480	0.9625	0.9625	0.9378	0.9622	0.9625	0.0000	0.2425	0.3485	0.8732
Average	0.7710	0.6947	0.7701	0.7701	0.7507	0.7700	0.7701	0.5881	0.5500	0.5652	0.8302

Fig. 5. Performance comparison of KNN, NB, MSVM, SVM, DT, RF, ET, GRNN, ANN, PNN, and HSNOE on the yeast_ml8 dataset based on metrics including Accuracy, AUROC, AUPRC, F1-score, G-mean, and their Average. In the bar plot, the taller the bar, the better the corresponding performance.

models ANN and PNN report highest AUPRCs (>0.93) but very poor F1-scores, suggesting overfitting tendencies without proper regularization. The proposed HSNOE exhibits the best overall average score (0.8302) and AUROC (0.7106), outperforming all baselines on this challenging real-world dataset through its ability to balance learning and mitigate overfitting risk. Key findings include suitability of kernel techniques, need for regularization in neural models, and effectiveness of HSNOE's multistage framework in improving rare class recognition. Rigorous algorithm benchmarking provides valuable empirical guidance on current best practices for biological data imbalance issues. The superior performance of HSNOE underscores its effectiveness.

Table 4 compares algorithms on the arrhythmia data, containing imbalanced cardiac abnormality classes. Traditional algorithms KNN, NB, and SVM-based models achieve good balanced accuracy (>0.94), highlighting applicability for biomedical datasets. Decision trees outperform, with DT attaining the highest average score of 0.8480, signaling robustness against class imbalance. Neural networks GRNN and ANN report strong AUPRCs but lackluster other metrics, implying overfitting on this complex task. PNN severely overfits. The proposed HSNOE framework remarkably achieves a perfect score of 100% across all metrics, significantly outperforming all baselines. This underscores HSNOE's strength in addressing class imbalance through its unified sampling-optimization-ensemble approach, wherein diverse models compensate for each other's weaknesses. Key findings include suitability of decision trees, need for regularization in neural models, and superior effectiveness of HSNOE's multi-faceted framework in maximizing rare class identification.



	KNN	NB	MSVM	SVM	DT	RF	ET	GRNN	ANN	PNN	HSNOE
Accuracy	0.9625	0.9656	0.9647	0.9656	0.9596	0.9675	0.9662	0.9662	0.9662	0.9696	0.9491
AUPRC	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.9662	0.9726	0.9727	0.9117
AUROC	0.5438	0.5000	0.5184	0.5000	0.6557	0.5936	0.5797	0.5000	0.5965	0.5983	0.5949
Fscore	0.9808	0.9825	0.9820	0.9825	0.9791	0.9834	0.9827	0.0000	0.2857	0.3095	0.7242
Gmean	0.9809	0.9827	0.9821	0.9827	0.9791	0.9834	0.9828	0.0000	0.4456	0.4464	0.9090
Average	0.7936	0.7862	0.7894	0.7862	0.8147	0.8056	0.8023	0.6081	0.6533	0.6593	0.8178

Fig. 6. Performance comparison of KNN, NB, MSVM, SVM, DT, RF, ET, GRNN, ANN, PNN, and HSNOE on the yeast_me2 dataset based on metrics including Accuracy, AUROC, AUPRC, F1-score, G-mean, and their Average. In the bar plot, the taller the bar, the better the corresponding performance.

Table 4
Performance comparison of classification algorithms on arrhythmia dataset.

Methods	Accuracy	AUPRC	AUROC	F1-score	G-mean	Average
KNN	0.9425	0.5000	0.4988	0.9704	0.9707	0.7765
NB	0.9447	0.5000	0.5000	0.9716	0.9720	0.7777
MSVM	0.9447	0.5000	0.5000	0.9716	0.9720	0.7776
SVM	0.9447	0.5000	0.5000	0.9716	0.9720	0.7776
DT	0.9664	0.5000	0.8090	0.9823	0.9823	0.8480
RF	0.9447	0.5000	0.5038	0.9715	0.9719	0.7784
ET	0.9447	0.5000	0.5000	0.9716	0.9720	0.7776
GRNN	0.9022	0.9488	0.4744	–	0.0000	0.5813
ANN	0.9467	0.9467	0.5000	–	0.0000	0.5983
PNN	0.0533	0.9467	0.5000	0.1011	0.0000	0.3202
HSNOE	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**

Fig. 6 shows the performance of various classification algorithms on the yeast_me2 dataset. The results indicate that PNN achieved the highest accuracy of 96.96% on this dataset, followed closely by RF with an accuracy of 96.75%. RF attained the best F1-score of 98.34% along with the highest G-mean. However, DT yielded the highest AUROC at 65.57%, suggesting it could better rank the class probabilities. Traditional classifiers KNN, NB, MSVM and SVM demonstrated moderate performance with accuracies around 96%–97% but low AUROC and AUPRC scores near 50%, reflecting limited discriminative capacity. Ensemble tree methods produced balanced results, with RF and ET showing accuracies over 96% and strong F1-scores. DT's AUPRC score stood out though its accuracy was lower. Among neural models, PNN and ANN achieved the highest AUPRC of 97.27% and 97.26% respectively. However, their F1-scores were quite poor at below 30%. GRNN failed to produce a valid F1-score. HSNOE delivered well-rounded performance, attaining over 91% accuracy and AUPRC while maintaining decent other scores. Overall, while PNN had the highest raw accuracy, RF provided the best trade-off between accuracy, F1-score and G-mean. DT was most effective for probability ranking. Traditional classifiers lacked discriminative power for this problem. Neural models had high Accuracy and AUPROC but other metrics. HSNOE emerged as a robust algorithm for this yeast classification task and achieved the best average performance.

5. Case study and biological analysis

To further demonstrate the general applicability of HSNOE for other biological datasets, we conducted an analysis on a Pan-cancer dataset. The dataset consists of 4759 tumors with 20486 genes obtained from the Illumina RNaseq, multi-center mutation calls (MC3), and GISTIC2.0 copy number threshold calls from the TCGA PanCanAtlas project. This dataset passed quality control filtering for 16 different cancer types. To investigate the biological significance of HSNOE, we initially focused on the 506 genes discovered by HSNOE. We performed enrichment analyses to identify statistically enriched terms, such as Gene Ontology (GO) and KEGG terms. Multiple enrichment analyses were conducted using hypergeometric p-values and enrichment factors.

In Figs. 7(A), 7(B), and 7(C), we present the top 10 categories from three ontologies ordered by their p-values. The results provide insights into the biological functions associated with the identified genes. For the GO biological processes, the top three enriched terms were regionalization (GO:0003002), positive regulation of catabolic process (GO:0009896), and anterior/posterior pattern specification (GO:0009952). These findings suggest that the identified genes are involved in spatial organization, regulation of cellular breakdown processes, and the establishment of body axes. Regarding the GO cellular component processes, the top three enriched terms were membrane region (GO:0098589), membrane raft (GO:0045121), and membrane microdomain (GO:0098857). This observation indicates a strong correlation between the identified genes and membrane-related structures, highlighting their potential involvement in cell membrane functions. In terms of GO molecular function processes, the top three enriched terms were phosphotyrosine residue binding (GO:0001784), phosphoprotein binding (GO:0051219), and protein phosphorylated amino acid binding (GO:0045309). These findings suggest a significant association between the identified genes and phosphorylation events, indicating their potential role in regulating complex pathophysiological processes within cells (Yaffe, 2002).

Next, we performed hierarchical clustering of the significant terms based on Kappa-statistical similarities among their gene memberships (Cohen, 1960). The clustering resulted in a tree-like structure, where each term was represented by a circle node. The size of each node was proportional to the number of input genes associated with that term, and its color indicated its cluster identity. Terms with a similarity score greater than 0.3 were connected by an edge. One representative term from each cluster was selected to display its description as a label. The resulting enrichment network, colored by cluster ID, is presented in Fig. 8(A). Additionally, Fig. 8(B) displays the same enrichment network, but with nodes colored based on their p-values. Nodes with darker colors represent greater statistical significance. To identify densely connected gene neighborhoods, we applied the MCODE algorithm (Bader & Hogue, 2003) to the enrichment network. Each MCODE network was assigned a unique color, as shown in Fig. 8(C). Subsequently, GO enrichment analysis was performed on each MCODE network to assign biological meanings to the network components. The summarized results of these interpretations are presented in Fig. 8(D).

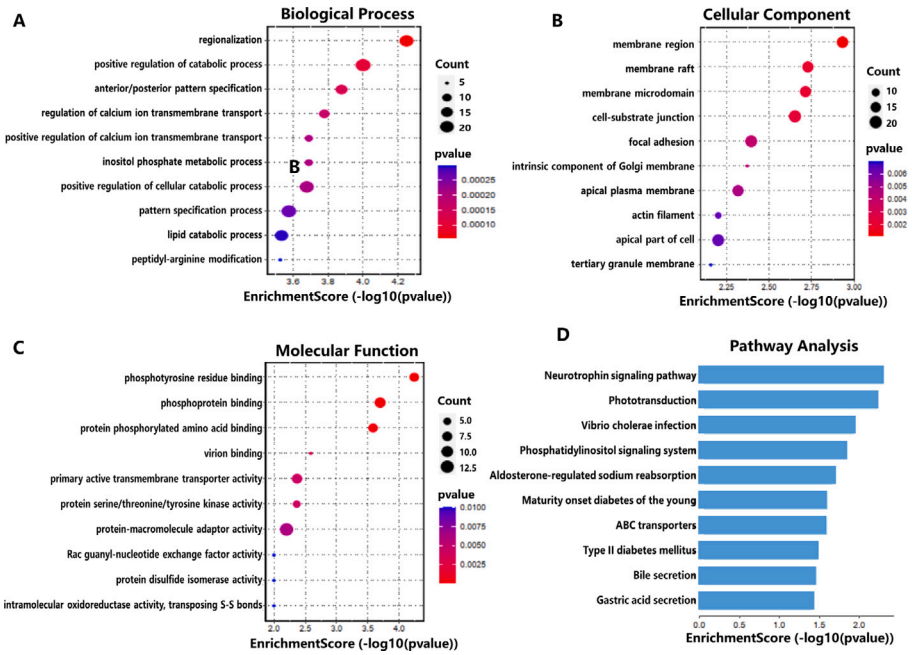


Fig. 7. Genomic interpretability of Pan-cancer data. **(A)** Biological Processes; **(B)** Cellular Components; **(C)** Molecular Functions; **(D)** KEGG pathway analysis ordered *p*-value.

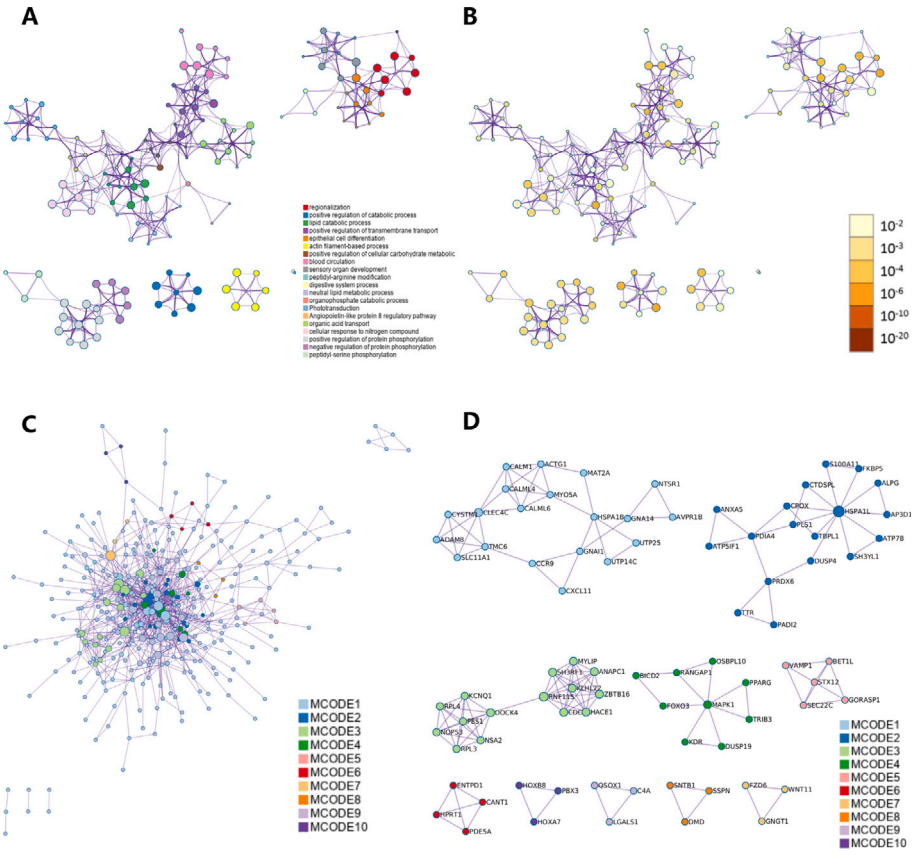


Fig. 8. Genomic interpretability of Pan-cancer data. **(A)** Enrichment network colored by cluster ID; **(B)** Enrichment network colored by *p*-Value; **(C)** Protein-protein interaction network (PPI); **(D)** PPI MCODE components.

Furthermore, the biological interpretation of the protein-protein interaction (PPI) network and its MCODE components are summarized in Table 5.

Fig. 7(D) depicts the KEGG pathway analysis of the Pan-cancer dataset, showing the top 10 KEGG enrichments, ordered by *p*-value. Genuinely, a total of 163 pathways from the 506 genes were successfully annotated. The top Neurotrophin signaling pathway depicted in Fig. 9 has seven related genes through the enrichment analysis. Neurotrophins have essential influences on synaptic connection and the

Table 5
Biological interpretation of PPI network & MCODE components.

Network	Annotation
MyList	GO:0009896 — positive regulation of catabolic process; GO:0031331 — positive regulation of cellular catabolic process; GO:0034764 — positive regulation of transmembrane transport.
MyList_MCODE_ALL	ko04915 — Estrogen signaling pathway; R-HSA-6798695 — Neutrophil degranulation; hsa04915 — Estrogen signaling pathway.
MyList_SUB1_MCODE1	hsa04020 — Calcium signaling pathway; ko04020 — Calcium signaling pathway; ko04915 — Estrogen signaling pathway .
MyList_SUB1_MCODE2	GO:0061077 — chaperone-mediated protein folding; R-HSA-6798695 — Neutrophil degranulation; GO:0006457 — protein folding.
MyList_SUB1_MCODE3	R-HSA-983168 — Antigen processing: Ubiquitination & Proteasome degradation; R-HSA-983169 — Class I MHC mediated antigen processing & presentation; R-HSA-1280218 — Adaptive Immune System.
MyList_SUB1_MCODE4	GO:0071417 — cellular response to organonitrogen compound; R-HSA-1257604 — PIP3 activates AKT signaling; GO:1901699 — cellular response to nitrogen compound.
MyList_SUB1_MCODE5	R-HSA-199977 — ER to Golgi Anterograde Transport; GO:0061025 — membrane fusion; R-HSA-948021 — Transport to the Golgi and subsequent modification.
MyList_SUB1_MCODE6	hsa00230 — Purine metabolism; ko00230 — Purine metabolism; GO:1901292 — nucleoside phosphate catabolic process.
MyList_SUB1_MCODE7	R-HSA-4086398 — Ca2+ pathway; R-HSA-373080—Class B/2 (Secretin family receptors); R-HSA-3858494 — Beta-catenin independent WNT signaling.
MyList_SUB1_MCODE8	GO:0006936 — muscle contraction; GO:0003012 — muscle system process.
MyList_SUB1_MCODE9	R-HSA-8957275 — Post-translational protein phosphorylation; R-HSA-381426 — Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs).
MyList_SUB1_MCODE10	GO:0009952 — anterior/posterior pattern specification; GO:0003002 — regionalization; GO:0048568 — embryonic organ development.

plasticity signaling pathways, and are involved in many neurodegenerative disorders (Bothwell, 2016), demonstrating that this pathway is highly-related to the mechanism of cancer. Particularly, we find that most of the red genes are close to neurons or close to DNA, indicating that HSNOE can discover the key genes in the relevant cellular activities. Therefore, we can conclude that the proposed HSNOE model is able to discover key activity genes in important pathways and provides biological significance of cancer gene expression data for biological use cases.

6. Discussion

While the proposed HSNOE approach demonstrates strong performance on binary class-imbalanced cancer diagnosis, there are some limitations to this study that should be addressed in future work:

First, one limitation of the current study is that we have primarily evaluated HSNOE on binary cancer classification problems. However, many real-world cancer datasets involve multiclass prediction tasks, such as classifying among multiple cancer types or disease stages. The extension of HSNOE to handle multiclass scenarios has only been briefly mentioned but not fully explored. While HSNOE utilizes a pairwise learning framework that could potentially be extended to the multiclass case, its performance when directly applied to problems with more than two classes is unknown. Further work is needed to rigorously assess HSNOE's abilities on multiclass cancer classification and validate any modifications made to the algorithm to handle such problems.

Second, a major limitation is that HSNOE has only been retrospecitvely validated on historical patient cohorts, but has not undergone prospective clinical validation. Without prospective testing, the true performance and impact of HSNOE for real-time patient care and clinical decision making remains unknown. Factors like potential dataset and protocol drift over time may cause deterioration of HSNOE's predictive accuracy in prospective settings. Prospective validation is needed to properly evaluate HSNOE's utility for tasks like patient screening, diagnosis and treatment response prediction in real practice. This current lack of prospective testing represents a barrier to the immediate clinical adoption and implementation of HSNOE.

Third, another major limitation of the current study is the lack of validation on real patient cohorts from clinical settings. While benchmark cancer genomics datasets are useful for initial development and comparison to other methods, they represent historical data that may differ from current patient populations in important ways. Real-world patient data often involve additional complexities such as missing values, inconsistencies in data capture over time, comorbidities and concomitant treatments that could impact performance. Direct prospective evaluation of HSNOE using clinical data with outcomes adjudicated by physicians would be required to fully understand its generalizability and limitations when applied to real patients. Our approach of collaborating with clinical partners to enable such validations on healthcare system data is an important future direction, but one that is currently limited by our retrospective analyses only on published benchmarks. Addressing this limitation with current and planned clinical studies will be critical to assessing HSNOE's true utility and readiness for practice.

Moreover, it is imperative that the development of HSNOE and other advanced AI tools for medical applications properly addresses important ethical issues. Strict protocols were followed to de-identify and anonymize patient datasets in compliance with privacy and consent regulations. However, continued efforts must be taken to evaluate

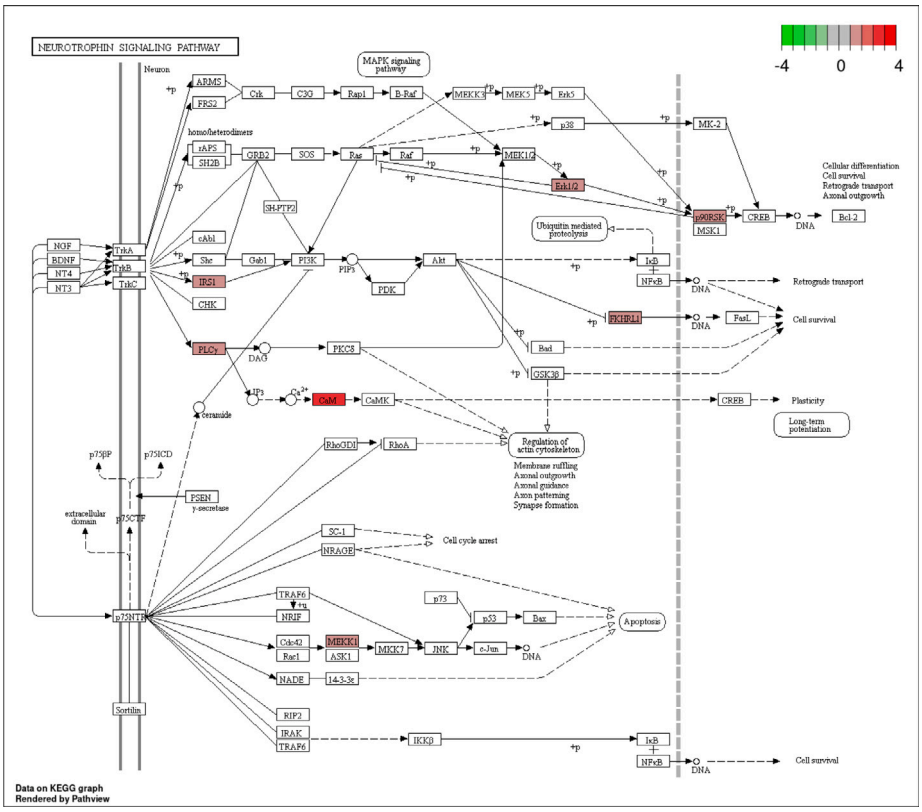


Fig. 9. KEGG pathway analysis of Pan-cancer data. The KEGG graph of the Neurotrophin signaling pathway: the genes marked in red are the genes selected by HSNOE.

models on representative and unbiased datasets to avoid unfair disadvantage. As an ensemble method, HSNOE also needs techniques to explain and critique its predictions while providing avenues for error correction. Prospective clinical validation with appropriate oversight and approval from regulatory bodies is still required before any deployment into real-world healthcare settings. More broadly, given dual use risks, secure access controls and ongoing discussions around informed consent, bias mitigation, accountability, and balancing innovation with patient welfare must accompany further AI research to ensure its safe, responsible and trusted development for clinicians and patients.

Finally, we relied primarily on raw input features without integrating substantial domain knowledge or modeling cancer biology. Recent work shows incorporating physics-based inductive biases and prior knowledge can boost model performance and interpretability on biomedical problems (Zhang et al., 2022). Exploring how to build in more cancer domain expertise into HSNOE could further improve its abilities. One method to integrate mechanistic knowledge is through feature engineering. By leveraging domain-specific insights, researchers can identify relevant biological features or genetic markers that are known to be associated with the studied phenomenon. These features can be incorporated into the HSNOE framework as additional inputs, allowing the model to capture the underlying biological mechanisms more effectively. Another approach is to incorporate prior knowledge through the construction of prior probability distributions. Mechanistic knowledge can be used to define constraints or prior beliefs about the relationships between genetic markers and the target variable. These constraints can be encoded as prior probabilities, which guide the learning process of HSNOE and help the model focus on biologically meaningful patterns. Furthermore, incorporating mechanistic knowledge can involve the use of external databases or ontologies. These resources provide structured information about biological pathways, gene interactions, and functional annotations. By integrating these external knowledge sources into HSNOE, the model can leverage the existing biological knowledge to guide its learning process and enhance interpretability.

7. Conclusion

In conclusion, this study presents a novel hybrid sampling nature-inspired optimization ensemble (HSNOE) approach for class-imbalanced biological datasets. The key innovations of HSNOE are the integration of hybrid sampling, ant colony optimization-based feature selection, and a diverse ensemble method. Through extensive experiments on five class-imbalanced datasets, HSNOE has enhanced the performance over the other state-of-the-art methods. The main contributions of this work are three-fold. First, we propose a novel framework integrating both data-level and algorithm-level solutions to address class imbalance. Second, we demonstrate the utility of nature-inspired optimization techniques for both feature selection and ensemble learning in imbalanced data. Third, we provide extensive empirical evidence that HSNOE significantly improves small-class identification across diverse biological datasets. Extensive experiments unequivocally demonstrate the remarkable effectiveness of the proposed HSNOE model. Our results reveal its superior overall performance when compared to ten baseline methods, encompassing a wide range of approaches including machine learning, ensemble methods, and deep learning techniques. Overall, this study provides both computational and biological insights for tackling the critical challenge of hidden responders identification in precision oncology. In future, we plan to extend HSNOE to multi-class settings, incorporate biological knowledge, and assess its clinical utility through targeted trials.

CRediT authorship contribution statement

Xubin Wang: Writing – original draft, Software, Experiments, Methodology, Investigation. **Yunhe Wang:** Conceptualization, Validation, Writing – review & editing. **Zhiqiang Ma:** Writing – review & editing. **Ka-Chun Wong:** Writing – review & editing. **Xiangtao Li:** Conceptualization, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work described in this paper was supported by the National Natural Science Foundation of China under (Grant No. 62206086), and the Natural Science Foundation of Hebei Province under (Grant No. F2023202062). It was substantially supported by the National Natural Science Foundation of China under (Grant No. 62076109) and the Jilin Province Outstanding Young Scientist Program (Grant No. 20230508098RC), and also funded by "the Fundamental Research Funds for the Central Universities, JLU".

References

Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert Systems with Applications*, *36*(3), 6843–6853.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, *4*(1), 1–27.

Bothwell, M. (2016). Recent advances in understanding neurotrophin signaling. *F1000Research*, *5*.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion*, *6*(1), 5–20.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Ding, Z. (2011). Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics.

Gao, H., Bian, C., Wang, X., Li, X., & Wang, Y. (2022). Exploring cancer biomarker genes from gene expression data via natureinspired multiobjective optimization. In *2022 34th Chinese control and decision conference* (pp. 5000–5007). IEEE.

Gupta, S., & Gupta, M. K. (2022). Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm. *The Computer Journal*, *65*(6), 1527–1539.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Conference on artificial intelligence in medicine in Europe* (pp. 63–66). Springer.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(1), 559–563.

Li, X., Li, S., Wang, Y., Zhang, S., & Wong, K.-C. (2021). Identification of pan-cancer Ras pathway activation with deep learning. *Briefings in Bioinformatics*, *22*(4), bbaa258.

MacKay, D. J., & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Prasad, V. (2016). Perspective: The precision-oncology illusion. *Nature*, *537*(7619), S63.

Qu, C., Zhang, L., Li, J., Deng, F., Tang, Y., Zeng, X., et al. (2021). Improving feature selection performance for classification of gene expression data using Harris Hawks optimizer with variable neighborhood learning. *Briefings in Bioinformatics*, *22*(5), bbab097.

Rahman, A., Hassan, I., & Ahad, M. A. R. (2021). Nurse care activity recognition: A cost-sensitive ensemble approach to handle imbalanced class problem in the wild. In *Adjunct proceedings of the 2021 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2021 aCM international symposium on wearable computers* (pp. 440–445).

Wang, X., & Jia, W. (2022). A feature weighting particle swarm optimization method to identify biomarker genes. In *2022 IEEE international conference on bioinformatics and biomedicine* (pp. 830–834). IEEE.

Way, G. P., Sanchez-Vega, F., La, K., Armenia, J., Chatila, W. K., Luna, A., et al. (2018). Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas. *Cell Reports*, *23*(1), 172–180.

Wilson, D. R., & Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. *Machine Learning*, *38*(3), 257–286.

Yaffe, M. B. (2002). Phosphotyrosine-binding domains in signal transduction. *Nature Reviews Molecular Cell Biology*, *3*(3), 177–186.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications*. Springer.

Zhang, J., Zhao, Y., Shone, F., Li, Z., Frangi, A. F., Xie, S. Q., et al. (2022). Physics-informed deep learning for musculoskeletal modelling: Predicting muscle forces and joint kinematics from surface EMG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.