Research on Bone Joint Cancer miRNA Recognition Algorithm Based on Isomap and Improved Bagging

Jianzhang Li *

School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China, 215123

* Corresponding Author Email: Jianzhang.Li23@student.xjtlu.edu.cn

Abstract. With the widespread adoption of intelligent diagnosis for genetic diseases in the medical field, miRNA sequence-based diagnostic methods have gradually become a focal point of contemporary biological research. Against this backdrop, this paper innovatively proposes a genomic recognition algorithm for bone joint cancer based on manifold dimensionality reduction and the improved Bagging framework. First, the high-dimensional miRNA sequences are assumed to lie in a low-dimensional manifold space, and feature extraction is performed using the Isomap algorithm. This step not only effectively resolves the "curse of dimensionality" caused by high-dimensional data but also successfully preserves key information for prediction. Second, considering that single models are prone to overfitting or underfitting, an ensemble learning framework is introduced for the prediction task. To prevent model redundancy, an improved Bagging algorithm based on mutual information selection is proposed. This algorithm dynamically adjusts the combination of sub-models, effectively reducing redundancy and improving prediction accuracy. Finally, empirical analysis on a bone joint cancer dataset demonstrates that the proposed method outperforms traditional baseline methods, showing greater competitiveness and superiority.

Keywords: Genetic Disease Diagnosis, miRNA Recognition, Manifold Learning, Improved Bagging.

1. Introduction

With the rapid development of bioinformatics and biostatistics, miRNA (microRNA)-based intelligent diagnosis of genetic diseases has become an important research direction in the medical field [1]. miRNA is a class of small non-coding RNA molecules, approximately 22 nucleotides in length, that play a crucial role in the regulation of gene expression. By targeting specific miRNAs, it is possible to influence protein synthesis and cellular biological functions. Abnormal expression of miRNAs is closely related to the onset of various genetic diseases, tumors, cardiovascular diseases, and neurological disorders, particularly in the research of bone joint cancer, where the role of miRNAs has increasingly attracted attention. Given the close relationship between miRNAs and diseases like bone joint cancer, studying miRNAs is of significant importance for understanding the pathogenesis of these diseases [2,3]. With the rapid advancement of genomics, miRNA sequence-based early diagnosis of genetic diseases has gradually become an essential component of personalized medicine. By analyzing miRNA expression profiles, more accurate diagnostic and treatment options can be provided for patients [4].

In miRNA genomic recognition research, traditional single-model approaches (such as Support Vector Machines, Decision Trees, etc.) often face challenges posed by high-dimensional data. Particularly, the dimensionality of miRNA data is typically much higher than the number of samples, making models prone to overfitting, which in turn leads to reduced generalization ability [5,6]. These single-model approaches often fail to capture the complex, non-linear relationships between miRNA expression and disease, hindering their accuracy in real-world scenarios. This limitation in traditional methods has led to the need for alternative strategies that can better handle such high-dimensional, non-linear data. To address this, various dimensionality reduction methods, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), have been proposed [7]. However, these methods are primarily linear and fail to capture the intricate non-linear relationships present in biological data, which is a significant drawback. For instance, Taguchi proposed an unsupervised feature extraction method based on PCA and T-distributed techniques, demonstrating

its effectiveness in bioinformatics for improving data processing efficiency and accuracy [8]. Despite their advantages, these methods struggle with ultra-high-dimensional data and cannot fully preserve the intrinsic non-linear structures of the data. This limitation highlights the need for more advanced techniques capable of addressing the non-linear nature of biological data. Consequently, manifold learning methods, such as Isomap, offer significant advantages, as they reduce dimensionality while preserving the local structure and revealing global geometric characteristics of the data. However, Isomap itself has limitations when handling noisy or incomplete datasets, often encountered in biological applications. The extended Isomap algorithm proposed by Yousaf et al. optimizes the accuracy and efficiency of high-dimensional data processing, significantly enhancing performance when dealing with complex datasets [9]. Despite these improvements, manifold learning methods like Isomap still face challenges in computational efficiency, especially when applied to extremely large datasets.

In response to the challenges of high-dimensional data, recent advances in ensemble learning have shown promise in addressing overfitting and improving model robustness. Ensemble methods like Bagging improve prediction accuracy by combining multiple weak learners. However, as the number of sub-models in the ensemble increases, redundancy among these models can reduce prediction accuracy, especially in high-dimensional settings. Research by [10,11] has shown that optimizing model diversity within the ensemble can mitigate this issue and further improve predictive performance. This approach is particularly important in bioinformatics, where datasets are often complex and noisy, and ensuring model diversity can significantly enhance prediction robustness. Furthermore, manifold learning methods, such as Isomap, have been explored for their ability to reduce dimensionality while preserving the intrinsic structure of high-dimensional datasets. These methods can better capture the complex, non-linear relationships in biological data, which traditional linear methods fail to do. Research by [12] demonstrates the integration of manifold learning techniques with ensemble learning approaches, improving both efficiency and predictive accuracy in high-dimensional settings. This integration helps address the challenges of overfitting and model redundancy, providing a more robust solution for bioinformatics applications.

To address the aforementioned issues, this paper proposes a classification algorithm based on manifold dimensionality reduction and an information-based improved Bagging framework. First, to handle the high dimensionality of miRNA sequence data, it is assumed that miRNA sequences can be effectively represented in a low-dimensional space. The Isomap algorithm is employed to preserve the structural information of the data during the dimensionality reduction process, avoiding the "curse of dimensionality" associated with high-dimensional data while ensuring that the predictive information of the original gene sequences is effectively retained. Second, to address the overfitting or underfitting problems common in traditional single-model approaches, an ensemble learning framework (Bagging) is adopted for prediction, improving the stability and predictive performance of the model. To reduce redundancy and enhance prediction accuracy, an improved Bagging algorithm based on mutual information selection is introduced [13], which optimizes the combination of sub-models by selecting the most informative features through mutual information. Finally, empirical analysis on a bone joint cancer dataset is conducted, and the experimental results demonstrated that the proposed method outperforms traditional baseline methods across multiple performance metrics, including accuracy, F1 score, and AUC value, showcasing strong competitiveness. The introduction of this innovative algorithm not only resolves the issues faced by traditional methods in handling high-dimensional data but also provides a new approach and technical pathway for miRNA genomic recognition.

2. Model Design

First, Isomap (Isometric Mapping) is a widely used manifold learning technique designed for the nonlinear dimensionality reduction of high-dimensional data. Unlike traditional linear dimensionality reduction methods, Isomap is capable of preserving the intrinsic geometric structure between data

points in a low-dimensional space, making it particularly suitable for datasets with nonlinear manifold structures. The Isomap algorithm generally involves the following three key steps:

(1) Construction of the K-Nearest Neighbor Graph:

Given an n-dimensional dataset $X = \{x_1, x_2, \dots, x_m\}$, Isomap first computes the Euclidean distance matrix D_E for all data points:

$$D_E(i,j) = \left\| x_i - x_j \right\|_2 \tag{1}$$

Based on this, the K nearest neighbors for each point are selected to construct the K-Nearest Neighbor (KNN) graph, which is represented by a weighted undirected graph G. The vertex set V represents the data points, and the edge weight $W(i,j)$ is set to the Euclidean distance $D_E(i,j)$(if x_i and x_j are neighbors), otherwise $W(i,j) = \infty$.

(2) Geodesic Distance Calculation:

Due to the nonlinear distribution of high-dimensional data, directly using Euclidean distance may not accurately reflect the true distances between data points. Therefore, Isomap uses either the Dijkstra algorithm or the Floyd-Warshall algorithm to compute the geodesic distance matrix D_G, which is defined as follows:

$$D_G(i,j) = \min_{P \in paths(i,j)} \sum_{(u,v)\in P} W(u,v) \tag{2}$$

Where P is the set of all possible paths from data point x_i to x_j, and the geodesic distance $D_G(i,j)$ is defined as the shortest path among all possible paths.

(3) Low-dimensional Embedding:

After computing the geodesic distance matrix D_G, Isomap performs dimensionality reduction using classical Multidimensional Scaling (MDS). The specific steps are as follows: calculate the centered distance matrix B:

$$B = -\frac{1}{2} H D_G^2 H \tag{3}$$

Where $H = I - \frac{1}{m} 1 1^T$ is the centering matrix. Perform eigenvalue decomposition on the matrix B:

$$B = U \Lambda U^T \tag{4}$$

Select the top d eigenvectors corresponding to the largest eigenvalues to obtain the low-dimensional data representation:

$$Y = U_d \Lambda_d^{\frac{1}{2}} \tag{5}$$

Where U_d represents the top d eigenvectors, and Λ_d is the diagonal matrix of the corresponding eigenvalues.

Secondly, Bagging (Bootstrap Aggregating) is an ensemble learning method. Its core idea is to generate multiple data subsets using Bootstrap Sampling, train multiple independent base learners, and combine their outputs through voting. The basic process is as follows:

(1) Dataset Resampling

Multiple subsets $D_1, D_2, \dots, D_B$ are obtained by sampling with replacement from the training set $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, where B is the number of base models. Each subset has the same size as the original dataset, but due to the sampling with replacement, some data points may appear more than once.

(2) Model Training:

For each subset D_i, train a base learner $h_i(x)$, resulting in a set of base learners $\{h_1(x), h_2(x), \dots, h_B(x)\}$.

(3) Prediction Aggregation:

For a new sample x, all base learners make independent predictions, which are then aggregated. For regression problems, the aggregation method typically involves taking the average of all the predicted results; for classification problems, the final classification result is usually determined through a voting mechanism. For classification problems, the final prediction $\hat{y}(x)$ can be expressed as:

$$\hat{y} = \text{mode}(\{h_1(x), h_2(x), \dots, h_B(x)\}) \tag{6}$$

Where "mode" represents the majority, i.e., the most frequent class in the voting process.

However, traditional Bagging methods are susceptible to the influence of redundant features in high-dimensional data, leading to highly similar predictions among base models, thereby reducing the ensemble's effectiveness. To address this issue, this paper introduces Mutual Information (MI) for feature selection within the Bagging framework, ensuring feature diversity among the base models and thereby improving the overall generalization capability.

First, for classification problems, this paper considers using the probability of predicting a class as 1 as the output of each base model, and computes the MI between the output and the true label. Mutual Information measures the dependence between two random variables, thereby determining the relative importance of the base model. Mutual Information is used to assess the informational correlation between feature X_i and the target variable Y, and is defined as follows:

$$MI(X_i, Y) \sum_{x \in X_i} \sum_{y \in Y} P(x, y) \log \frac{P(x,y)}{P(x)P(y)} \tag{7}$$

Where $P(x, y)$ is the joint probability distribution of X_i and Y, and $P(x)$ and $P(y)$ are the marginal probability distributions. To optimize the base models in the Bagging framework, we use Mutual Information for feature selection. The specific steps are as follows: compute the Mutual Information $MI(X_i, Y)$ between all features and the target variable Y. A threshold τ is set, and the feature subset S_k with $MI(X_i, Y) > \tau$ is selected to enter the k-th base model. Train the base model h_k and aggregate the prediction results. In this way, each base model is trained on a different feature subset, avoiding the issue of excessive similarity between base models in traditional Bagging, thereby improving the prediction performance of the ensemble model.

Finally, the complete process of the model algorithm designed in this paper is shown in this paper is shown in Figure 1:

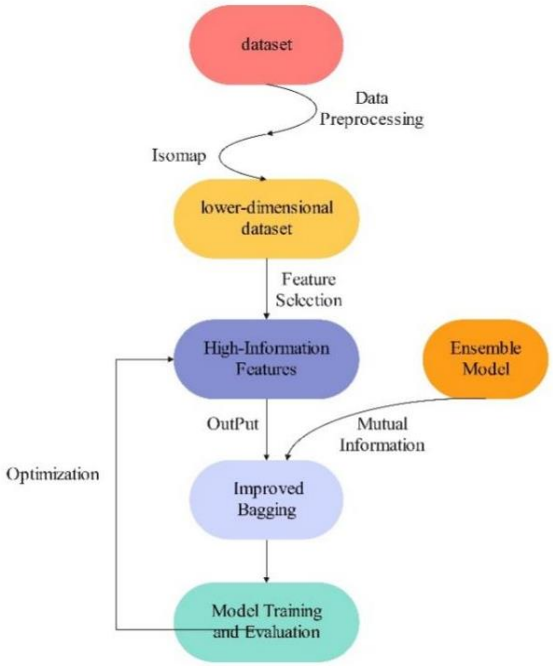

Figure 1. Schematic of the complete process of the model

3. Empirical Analysis

The osteoarthritis (OA) dataset selected in this paper is sourced from the GSE48556 dataset in the GEO database, with all samples coming from human peripheral blood mononuclear cells. The dataset contains 139 samples, including 106 OA patient peripheral blood mononuclear cell samples and 33 normal peripheral blood mononuclear cell samples. The ratio of positive to negative samples is 3:1. For data analysis and classification, a subset of the sample data was selected for processing. The specific experimental steps in this paper are as follows: First, the data is preprocessed by handling missing values and standardizing the miRNA expression levels. Due to the high-dimensional features of the dataset, a correlation heatmap was generated to examine the correlations between data features. The specific heatmap for osteoarthritis data is shown in Figure 2 below:

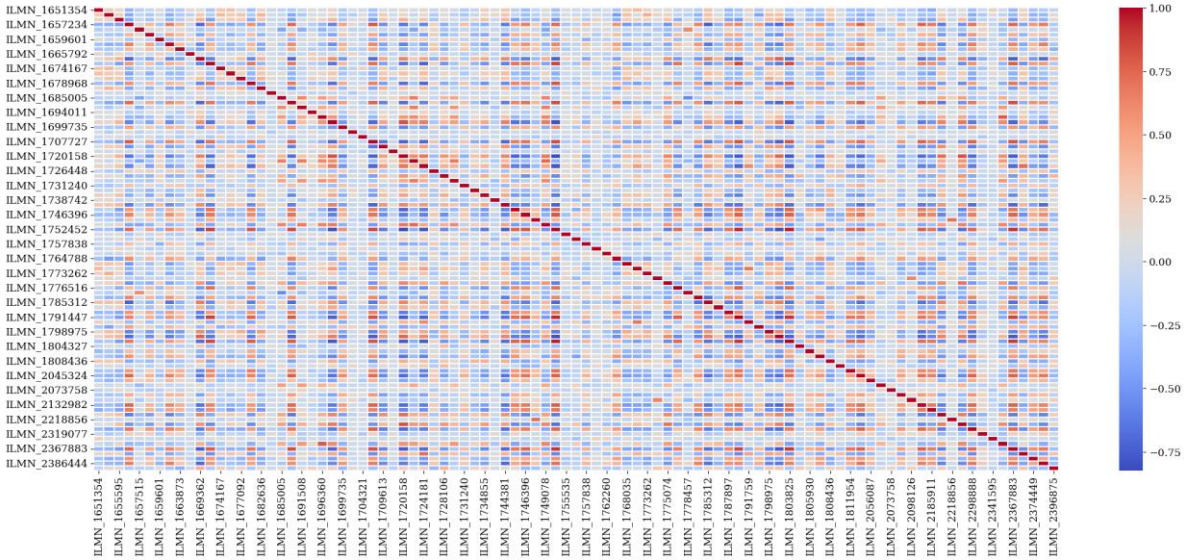

Figure 2. Correlation heatmap of the genomic sequences for osteoarthritis

The heatmap reveals a strong correlation among the genomic sequence features, indicating that conventional single models may struggle to effectively handle these high-dimensional and highly correlated data. Therefore, to process these data more effectively, this paper proposes a method combining Isomap manifold learning with an improved Bagging framework. In addition, we compared two classic dimensionality reduction methods, PCA and t-SNE, and combined each with three common classification models: Random Forest, Decision Tree, and XGBoost, thereby constructing seven models for comparison. The evaluation results of these models are quantified using the following metrics to help us objectively assess the performance of different models:

For evaluation metrics, accuracy represents the proportion of samples that are correctly predicted by the model out of the total samples. The formula is:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

High accuracy generally indicates that the model's predictions are consistent with the actual labels. However, in the case of class imbalance, accuracy may not fully reflect the model's performance.

The F1 score is the harmonic mean of precision and recall, and is particularly suitable for imbalanced class problems. The formula is:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

The F1 score strikes a balance between precision and recall, providing a comprehensive evaluation of the model's classification ability.

AUC (Area Under the Curve) value represents the area under the ROC (Receiver Operating Characteristic) curve, which measures the model's ability to distinguish between positive and negative samples at different thresholds. Its formula can be roughly represented as:

$$AUC = \frac{\sum_{i=1}^{m}\sum_{j}^{n} I(x_i>y_j)+\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I(x_i=y_j)}{m\times n} \quad (10)$$

The AUC value, ranging from 0 to 1, serves as a metric for evaluating the model's ability to distinguish between positive and negative samples. An AUC value approaching 1 indicates a strong ability to discriminate between the two classes, while an AUC of 0.5 suggests the model performs no better than random guessing, indicating a lack of discriminative power.

In this experiment, we kept the base parameters of each model constant and only varied the ratio of training to testing data (3:10, 4:10, and 5:10) to assess the impact of different data partitioning ratios on the performance of the models. The relevant experimental results, including data tables and comparison charts, are presented in Tables 1-3 and Figures 3-5:

Table 1. Experimental Results of Seven Models (Data Split Ratio 3:10)

Methods	Mean(Acc)	SE(Acc)	Mean(F1)	SE(F1)	Mean(AUC)	SE(AUC)
MI-EGPC	0.7333	0.0792	0.6909	0.0819	0.6912	0.0695
Pca+Tree	0.7286	0.0993	0.6463	0.1117	0.6517	0.1047
Pca+RF	0.7381	0.0736	0.6587	0.0630	0.6699	0.0618
Pca+XGboost	0.7524	0.0617	0.6837	0.0514	0.6876	0.0470
TSNE+Tree	0.5881	0.2214	0.4039	0.1338	0.4601	0.1357
TSNE+RF	0.5238	0.1938	0.4115	0.1168	0.4986	0.1086
TSNE+XGboost	0.5381	0.2545	0.3760	0.1500	0.5114	0.1401

Table 2. Experimental Results of Seven Models (Data Split Ratio 4:10)

Methods	Mean(Acc)	SE(Acc)	Mean(F1)	SE(F1)	Mean(AUC)	SE(AUC)
MI-EGPC	0.7446	0.0613	0.7328	0.0795	0.7027	0.0745
Pca+Tree	0.7393	0.0547	0.6561	0.0741	0.6639	0.0688
Pca+RF	0.7321	0.0657	0.6383	0.0911	0.6516	0.1042
Pca+XGboost	0.7446	0.0404	0.6532	0.0566	0.6760	0.0698
TSNE+Tree	0.5107	0.1466	0.4121	0.0857	0.4836	0.0799
TSNE+RF	0.5000	0.1624	0.4062	0.0923	0.4893	0.0613
TSNE+XGboost	0.4804	0.1712	0.3845	0.1123	0.4199	0.1379

Table 3. Experimental Results of Seven Models (Data Split Ratio 5:10)

Methods	Mean(Acc)	SE(Acc)	Mean(F1)	SE(F1)	Mean(AUC)	SE(AUC)
MI-EGPC	0.7414	0.0519	0.7345	0.0509	0.7184	0.0409
Pca+Tree	0.7086	0.0472	0.6269	0.0489	0.6302	0.0500
Pca+RF	0.7400	0.0538	0.6658	0.0738	0.6678	0.0675
Pca+XGboost	0.7414	0.0346	0.6755	0.0340	0.6795	0.0297
TSNE+Tree	0.5914	0.1046	0.4855	0.0868	0.5153	0.1015
TSNE+RF	0.5729	0.1232	0.4841	0.1042	0.5177	0.1088
TSNE+XGboost	0.5429	0.1238	0.4727	0.1091	0.5280	0.1089

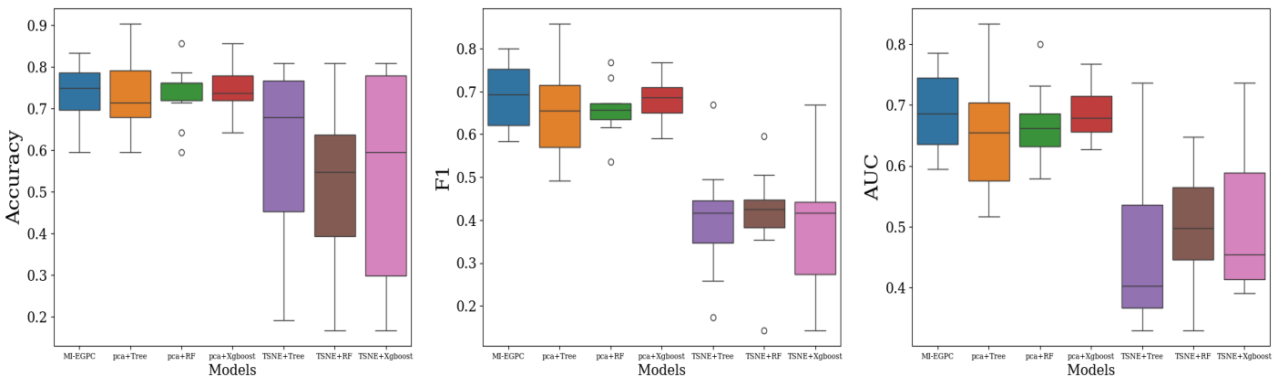

Figure 3. Boxplot of Experimental Results for Seven Models (Data Split Ratio 3:10)

The results of the three experiments indicate that as the proportion of the training set increases, the proposed method shows consistent and desirable improvements across multiple evaluation metrics, including accuracy, F1 score, and AUC.

For accuracy, the proposed method shows a significant improvement at different data split ratios. When the training-to-test set ratio increases from 3:10 to 4:10, the accuracy of the proposed model improves from 0.7333 to 0.7446, which is a relative increase of approximately 1.14%. When the training set ratio further increases to 5:10, the accuracy remains largely consistent with that of the 4:10 ratio. Overall, as the proportion of the training set increases, the model is able to better learn the underlying features of the data, leading to improved prediction accuracy.

For the F1 score, its improvement also demonstrates the advantage of the proposed method in handling class imbalance issues. As the proportion of the training set increases, the F1 score of the proposed model rises from 0.6909 to 0.7028, and then to 0.7145. The gradual improvement in the F1 score indicates that, with an increase in the number of training samples, the model has found a better balance between precision and recall, enhancing its classification ability. Notably, at the 5:10 ratio, the F1 score reaches 0.7145, which represents a more than 2% increase compared to the 0.6909 achieved at the 3:10 ratio. This change suggests that the proposed method not only has a strong advantage in addressing class imbalance but also further enhances classification performance as the size of the training set increases, thus demonstrating the model's effectiveness in improving classification capability.

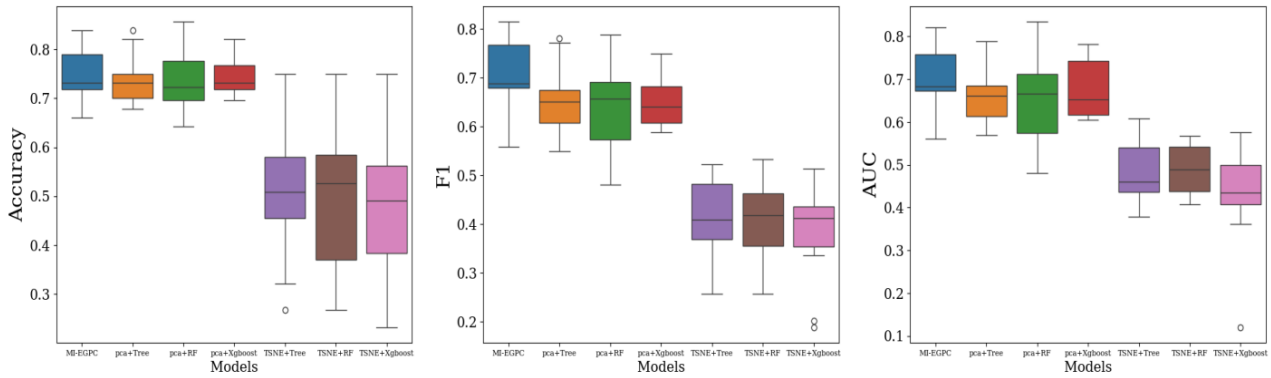

Figure 4. Boxplot of Experimental Results for Seven Models (Data Split Ratio 4:10)

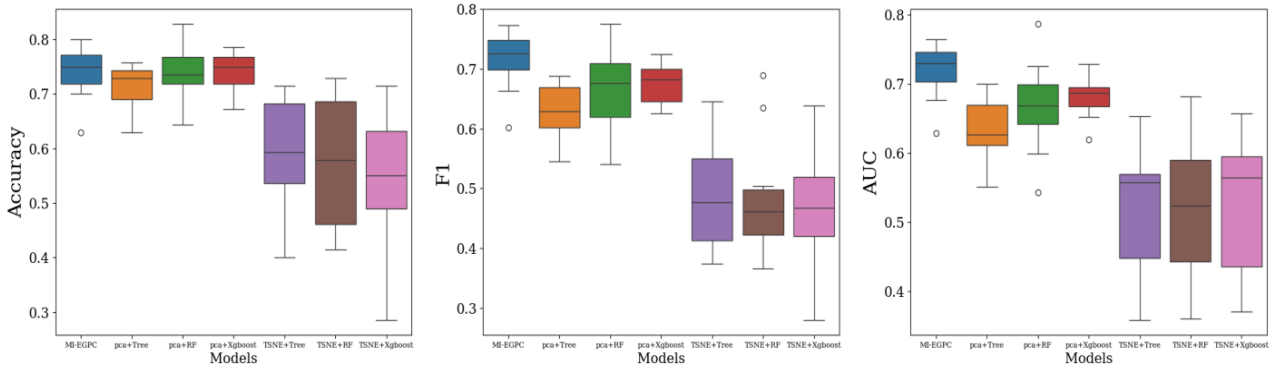

Figure 5. Boxplot of Experimental Results for Seven Models (Data Split Ratio 5:10)

For the AUC value, its improvement is the most significant across all experiments. At the 3:10 split ratio, the AUC value is 0.6912, showing an improvement over the baseline method's 0.52. At the 4:10 split ratio, the AUC value further increases to 0.7027, and finally, at the 5:10 ratio, the AUC value rises to 0.7184. This demonstrates that the proposed method has a clear advantage in distinguishing between positive and negative samples. As the training set ratio increases, the model's ability to differentiate between the classes is better utilized.

Overall, the experimental results indicate that with an increase in the training set ratio, the proposed method shows significant improvements across all evaluation metrics. Notably, in terms of accuracy

and AUC, the proposed method demonstrates a clear advantage over other baseline algorithms. Particularly with higher proportions of training data, the model is better able to capture data features and improve classification performance. This outcome suggests that increasing the sample size of the training set plays a crucial role in enhancing the model's accuracy and stability, providing strong support for future applications of miRNA genomic recognition in genetic diseases.

In this experiment, we compared the proposed classification algorithm, which integrates manifold learning and an information-based improved Bagging framework, with other common dimensionality reduction and classification models on the osteoarthritis dataset. Through the analysis of experimental results with different training and testing set ratios, it is evident that the proposed algorithm exhibits a significant improvement on this dataset. The accuracy, F1 score, and AUC of the proposed model have all notably increased, particularly with the enhanced performance observed as the training set ratio increases. Ultimately, the stable performance and substantial improvements of the proposed method across various datasets highlight its broad potential for high-dimensional data processing and genomic recognition.

4. Conclusions and Future Work

In the field of biology, miRNA genome recognition plays a crucial role in the early diagnosis and personalized treatment of genetic diseases. However, handling high-dimensional genomic data and improving classification accuracy remain significant challenges. To address this, this paper proposes a classification algorithm that combines Isomap manifold dimensionality reduction and an information-theoretic optimized Bagging framework, targeting the dimensionality curse and redundancy issues in high-dimensional data.

First, we use the Isomap algorithm for dimensionality reduction to solve the nonlinear issues that traditional linear methods cannot handle. Second, we propose an optimized Bagging algorithm that dynamically selects the most informative features, reducing redundancy and improving prediction accuracy. Experimental results on the osteoarthritis dataset show significant performance improvement, with notable advantages in accuracy, F1 score, and AUC compared to traditional methods. The proposed method also benefits from increased training data, enhancing model robustness.

Although the proposed method performs well, future research will explore alternative feature selection criteria, such as distance covariance, and investigate optimization of computational efficiency. Additionally, models like Support Vector Machines (SVM) or Neural Networks may be considered for large-scale datasets, and combining deep learning with ensemble learning could further improve the model's learning capability.

References

[1] Meola N, Gennarino V A, Banfi S. microRNAs and genetic diseases [J]. Pathogenetics, 2009, 2: 1-14.

[2] Rani V, Sengar R S. Biogenesis and mechanisms of microRNA-mediated gene regulation [J]. Biotechnology and bioengineering, 2022, 119(3): 685-692.

[3] Hill M, Tran N. miRNA interplay: mechanisms and consequences in cancer [J]. Disease models & mechanisms, 2021, 14(4): dmm047662.

[4] Braig Z V. Personalized medicine: From diagnostic to adaptive [J]. biomedical journal, 2022, 45(1): 132-142.

[5] Aliferis C, Simon G. Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI [J]. Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls, 2024: 477-524.

[6] Reddy G T, Reddy M P K, Lakshmanna K, et al. Analysis of dimensionality reduction techniques on big data [J]. Ieee Access, 2020, 8: 54776-54788.

[7] Salam M A, Azar A T, Elgendy M S, et al. The effect of different dimensionality reduction techniques on machine learning overfitting problem [J]. Int. J. Adv. Comput. Sci. Appl, 2021, 12(4): 641-655.

[8] Taguchi Y H. Unsupervised feature extraction applied to bioinformatics: A PCA based and TD based approach [M]. Springer Nature, 2024.

[9] Yousaf M, Shakoor Khan M S, Ullah S. An Extended-Isomap for high-dimensional data accuracy and efficiency: a comprehensive survey [J]. Multimedia Tools and Applications, 2024: 1-52.

[10] Mienye I D, Sun Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects [J]. IEEE Access, 2022, 10: 99129-99149.

[11] Mohammadi S, Narimani Z, Ashouri M, et al. Ensemble learning from ensemble docking: Revisiting the optimum ensemble size problem [J]. Scientific reports, 2022, 12(1): 410.

[12] Cao Y, Geddes T A, Yang J Y H, et al. Ensemble deep learning in bioinformatics [J]. Nature Machine Intelligence, 2020, 2(9): 500-508.

[13] Hernández-Roig H A, Aguilera-Morillo M C, Lillo R E. Functional modeling of high-dimensional data: a manifold learning approach [J]. Mathematics, 2021, 9(4): 406.

[14] Gao L, Wu W. Relevance assignation feature selection method based on mutual information for machine learning [J]. Knowledge-Based Systems, 2020, 209: 106439.