American Journal of Biochemistry and Biotechnology

Research Article

Random Forest-Based Genome Analysis for Disease Association and SNP Marker Identification

Xiangdong Liu

Department of Nursing, Zhoukou Polytechnic, Wenchang Street, Chuan Hui District, Zhoukou, Henan, China

Article history
Received: 07-11-2024
Revised: 07-04-2024
Accepted: 15-04-2025

Abstract: At present, genomic data analysis has the problems of insufficient model interpretation and limited computational efficiency in disease association research. As a powerful machine-learning algorithm, Random Forest has demonstrated excellent ability in processing complex and multi-dimensional data sets and has gradually become a popular tool in the field of bioinformatics. The purpose of this paper is to explore how RF can be applied to genome data mining and reveal the potential relationship between gene variation and disease. The study collected whole-genome sequencing data from 10,000 patients and labeled whether they had a specific type of cardiovascular disease. RF was used to construct a prediction model, and the SNP loci closely related to the occurrence of diseases were identified. The results show that compared with traditional statistical methods, the AUC value of the RF model reaches 0.92, the accuracy rate is as high as 89%, and the sensitivity and specificity are 87 and 90%, respectively. This indicates that RF can effectively identify key genetic markers, providing a valuable list of candidate genes for subsequent studies. In order to further verify the effectiveness of the RF model, this study selected the top 50 high-risk SNPs for functional annotation analysis. These variants were found to be mainly focused on genes known to be involved in lipid metabolism, inflammatory response, and immunomodulatory pathways. Rs6511723 is located in the APOE-C1/C4/C2 region, which is involved in cholesterol transport; rs1121980, on the other hand, is close to the FTO gene, which can affect weight and obesity risk by encoding a fatty acid oxidase. The ROC curve of MHILDA based on 5x validation is close to the true prediction interval, and its average AUC is 90.45%, which fully reflects its stable performance, and combined with experimental data, it can be determined that the MHILDA model shows excellent performance on the reference dataset and can accurately predict the potential LDA.

Keywords: Random Forest, Genome, Data Analysis, Disease Association

Introduction

With the rapid advancements of genomics, scientists now have the unprecedented ability to deeply explore the most subtle mysteries of the human body. Since the successful completion of the Human Genome Project in 2003, life science research has entered a brand-new era. People have gradually realized that small variations in DNA sequences may be the key to unlocking the mechanism of complex diseases (You *et al.*, 2023). In order to fully explore this potential, it is particularly important and urgent to develop efficient computational tools to process and analyze massive amounts of genetic information (Ou *et al.*, 2024).

In this context, Machine Learning (ML) has rapidly emerged as an interdisciplinary subject, which makes algorithms automatically learn laws and patterns and make predictions or decisions with the help of statistical and computer science principles (Gualdi *et al.*, 2024; Seyedmirzaei *et al.*, 2024). Among many ML algorithms, Random Forest (RF) has attracted much attention because of its advantages of flexibility, easy understanding and interpretation (Wang *et al.*, 2024). The RF-integrated learning method was proposed by Leo Breiman in 2001. It constructs multiple independent decision trees to form a forest and finally synthesizes the voting of all trees to determine the output result (Zhao *et al.*, 2025). RF can effectively cope with the challenges brought by noisy data and highly nonlinear relationships, especially when the number of features in large-scale data sets is much larger than the number of samples and is very suitable for high-throughput genomic data

SCIENCE Publications

analysis tasks (Gómez-Méndez and Joly, 2024; Rabiei *et al.*, 2024).

Based on the above characteristics, more and more researchers have begun to try to apply RF to genomics research and have achieved outstanding results in various complex diseases such as cancer, diabetes and mental disorders. For example, in a transcriptomics study of brain tissue samples from patients with Alzheimer's disease, RF successfully identified a set of core gene modules closely related to the neurodegeneration process, providing important clues for the design of new drug targets. The other is a screening project for hypertension susceptibility genes. Through RF analysis, a number of previously undiscovered candidate SNP sites were found and they showed good predictive performance.

Although the current research on the correlation between genomic data analysis and diseases has made some progress, it still faces two major challenges: First, the high computational cost caused by model complexity limits the rapid analysis of large-scale genomic data; Second, the prediction results of the model are not explanatory and it is difficult to intuitively reveal the internal relationship between key genes and diseases. To solve these problems, this study aims to develop an efficient and interpretable Random Forest algorithm in order to improve computational efficiency and enhance the interpretability of results while ensuring prediction accuracy.

In the field of genomics research, traditional genomic data analysis methods are affected by the "dimensionality disaster" of high-dimensional data when screening key gene markers related to diseases, which has limited accuracy and it is difficult to mine complex nonlinear association patterns between genes and diseases and at the same time, the ability to deal with data noise and missing values is insufficient. In this study, the random forest algorithm was used to accurately screen key gene markers by virtue of its powerful feature selection ability in processing high-dimensional data. The characteristics of constructing the comprehensive results of multiple decision trees are used to reveal complex correlation patterns. Relying on strong robustness to data noise, random sampling and feature selection to alleviate the impact of missing values, so as to effectively overcome the limitations of traditional methods. However, the current random forest algorithm has low computational efficiency and poor model interpretability in the face of large-scale genomic datasets, which is the problem and existing research gap that needs to be solved in this study.

Compared with traditional studies that only focus on the association between a single SNP locus and disease, random forest-based genomic data analysis has full advantages in the field of cardiovascular disease research. Cardiovascular disease is a complex multifactorial disease that involves the interaction of genetic and environmental factors. Random forests use ensemble learning strategies to construct multiple decision trees and summarize the results, which can effectively cope with the high-dimensional characteristics of genomic data, accurately analyze the comprehensive impact of multiple SNP site combinations on cardiovascular disease risk, such as identifying SNP combinations associated with vascular smooth muscle cell proliferation and migration, which is essential for atherosclerosis, a key pathological basis of cardiovascular disease and automatically assess the importance of SNP features to avoid missing key information. In terms of data mining depth, it is good at capturing the complex interaction patterns between multiple SNP loci, revealing the deep gene-gene and gene-environment interactions in cardiovascular diseases, such as some seemingly unrelated SNP synergies, which can affect the inflammatory response of the cardiovascular system by regulating the expression of inflammation-related genes, change the risk of disease and can efficiently process massive SNP data to mine more hidden signals. From the perspective of research, this analysis method can carry out dynamic risk assessment of cardiovascular diseases based on individual SNP information, lifestyle, clinical indicators, etc., predict the likelihood of onset at different time nodes and integrate genomic data of various cardiovascular diseases such as coronary heart disease, myocardial infarction, arrhythmia to carry out cross-disease analysis, identify shared SNP features or pathways, such as finding SNP characteristics that are closely related to cardiometabolic pathways in coronary heart disease and myocardial infarction, in order to deeply understand the common pathogenesis of cardiovascular diseases. It provides a new and comprehensive theoretical basis for accurate classification, early diagnosis and personalized treatment and strongly promotes the transformation of cardiovascular disease genomics research from single disease association analysis to comprehensive mechanism analysis and clinical application.

Related Theories and Technologies

Fundamentals of Genomic Data Analysis

In the field of genomic data analysis and disease association research based on random forest, the Random Forest (RF) algorithm shows significant innovations and expansions in methods. In terms of the innovation of combining with biological pathways, traditional genomic analysis usually only focuses on the association between a Single Nucleotide Polymorphism (SNP) and a disease. However, this study ingeniously integrates the results of SNP importance assessment with biological pathways. Specifically, after accurately calculating the importance score of each SNP using the RF algorithm, with the help of rich biological knowledge, the relevant SNPs are accurately mapped into specific biological pathways. Taking the research on cardiovascular diseases as an example, researchers will comprehensively analyze the importance of SNPs in pathways closely related to lipid metabolism, inflammatory response, etc. This integration method enables the resea (rch to no longer be trapped by isolated genetic variations, but to deeply understand the synergistic effect of genetic factors on functional pathways in the process of disease occurrence and development from the macroscopic perspective of systems biology. At the same time, through the in - depth combination of RF and biological pathways, potential biological mechanisms that have never been discovered before can be uncovered. Some genetic variations that seem insignificant in single - SNP analysis, once placed in a specific biological pathway and combined with the comprehensive analysis of RF, are very likely to reveal key regulatory roles in the entire pathway, thus helping to uncover the complex molecular network behind the disease and opening up new targets and ideas for disease prevention, diagnosis and treatment.

In terms of supplementing and improving traditional Genome-Wide Association Studies (GWAS), traditional GWAS mainly relies on univariate analysis to detect common SNPs significantly associated with diseases. However, it has many limitations, such as difficulty in effectively detecting low - frequency variations and being powerless when considering gene-gene and gene-environment interactions. In contrast, the RF algorithm can efficiently process high-dimensional data and naturally take into account the interactions between variables during model construction. When analyzing genomic data, RF can simultaneously incorporate a large number of SNPs and other potential influencing factors, such as environmental factors, into the analysis scope, effectively capturing the genetic effects that GWAS is likely to miss. Moreover, the RF algorithm constructs multiple decision trees and makes comprehensive predictions. Compared with the univariate analysis of GWAS, it can create a more accurate disease prediction model. For example, in the study of complex diseases such as diabetes, the RF model can organically integrate multiple risk SNPs discovered by GWAS, as well as clinical characteristics and lifestyle factors and then construct a more comprehensive prediction model, greatly improving the accuracy of predicting the risk of diabetes onset and providing a more powerful tool for early intervention and precise prevention of diseases. In addition, the results of GWAS often only present the SNP loci associated with diseases, but lack in - depth and thorough explanations of how these loci function in biological processes. The RF algorithm, by combining with biological pathways, can closely link SNPs with specific biological functions, making the research results more biologically interpretable. For instance, the RF analysis results can clearly and explicitly point out which biological pathways play key roles in the occurrence of diseases and what the relationships are between different SNPs in these pathways, greatly helping researchers to more thoroughly understand the genetic basis and pathogenesis of diseases.

In genomics, how to mine the disease code hidden in the vast genetic information has become a major challenge for researchers (Pinheiro *et al.*, 2024). Random Forest (RF), as a powerful machine-learning algorithm, has shown extraordinary application value in the field of genomic data analysis with its unique data processing capabilities and model construction efficiency.

RF algorithm combines the intuition of decision trees and the stability of Bagging and deeply analyzes genomic data by constructing multiple independent decision tree models (Gómez-Méndez and Joly, 2023; Usha Ruby *et al.*, 2023). Among them, each decision tree is constructed based on different data subsets and feature subsets, which reduces the possible bias caused by a single model by enhancing the diversity of models (Zhang *et al.*, 2024). When dealing with high-dimensional genomic data, RF can effectively avoid overfitting while maintaining sensitive capture of important variables and can maintain good prediction performance even when the number of variables far exceeds the sample size.

In the analysis of genomic data, the feature selection function of RF can adaptively evaluate the importance of each feature and isolate the key genetic variation regions closely related to the occurrence and development of diseases. By calculating indicators such as the Gini index or information gain of each feature, RF can quickly locate those genes that really drive the disease process, providing powerful clues for subsequent biological experiments (Gronowski *et al.*, 2024; Cai *et al.*, 2024). Modern genome research often involves thousands or even millions of genetic markers, which puts forward extremely high requirements for data processing. The parallel processing characteristics of RF enable it to run efficiently in distributed computing environments and easily cope with such large-scale data sets (Lemane *et al.*, 2024). In addition, the built-in Out-Of-Bag (OOB) error estimation mechanism of the algorithm allows the model performance to be evaluated during the model construction stage without the need for additional cross-validation steps, which greatly saves time and computational resources.

In this study, the random forest algorithm is used as the core analysis tool to carry out in-depth analysis of massive and complex genomic data, with the goal of accurately and deeply revealing the intrinsic association between genomic data and diseases. At the beginning of the study, large-scale genomic datasets were collected and organized, random forest algorithms were used to build models and key gene loci potentially related to diseases were mined through data feature screening and importance evaluation. In order to further enhance the reliability and universality of the research conclusions, external datasets covering genomic data from different regions, ethnicities and sample sizes were deliberately introduced from authoritative public databases and other published high-quality studies. After integrating internal and external datasets, the constructed random forest model was verified and optimized for multiple times through cross-comparison and the consistency and difference of data features were analyzed from different dimensions, so as to further consolidate the research results. At the same time, the functional annotation work was carried out in depth for the top Single Nucleotide Polymorphisms (SNPs) screened by the model. Firstly, bioinformatics tools were used to predict gene function and pathway enrichment of top SNPs to determine the biological processes and signaling pathways that they potentially affect. Subsequently, a series of experiments were carefully designed, such as cell function experiments, molecular biology experiments, etc., to rigorously verify the results of functional annotation. In cell function experiments, genes carrying top SNPs are knocked out or overexpressed by gene editing technology and changes in cell phenotype, including cell proliferation, apoptosis and migration, are observed. In molecular biology experiments, real-time quantitative PCR, western blotting and other technologies are used to detect the expression level of related genes and the synthesis and modification of proteins, so as to comprehensively and accurately interpret the mechanism of genomic data in the process of disease occurrence and development and provide a solid theoretical basis and experimental support for early diagnosis, precision treatment and prevention of diseases.

Although RF is often considered a black-box model, it does not affect the in-depth understanding of the internal operation logic of the model by drawing feature importance charts and local interpretation methods (such as SHAP) in genomics research (Luo *et al.*, 2024).

In the field of genomic data analysis and disease association research, Random Forest (RF) has significant advantages over traditional statistical methods such as logistic regression and chi-square test. Logistic regression assumes that there is a linear relationship between variables, which makes it difficult to capture complex nonlinear associations in genomic data and the chi-square test is not capable of dealing with multivariate interactions. Quantitative evidence shows that in genomic data analysis, the accuracy of RF models is 85%, while that of traditional logistic regression models is only 60% and the chi-square test is even lower. In terms of disease risk prediction, the area under the receiver operating characteristic curve (AUC) of RF was significantly higher than that of traditional methods. In view of the "black box" feature of RF, the relative importance of each gene feature to disease prediction can be visualized by using the feature importance ranking to identify key genes and the SHapley Additive exPlanations (SHAP) value can be used to assign a contribution value to the model output from a game theory perspective to explain the model decision-making process, so that researchers can not only use the powerful analytical capabilities of RF, but also deeply understand the biological mechanism behind the model. It provides strong support for early diagnosis and precise treatment of diseases.

Random Forest Algorithm Principle

In view of the significant impact of confounding factors such as demographics (e.g., age, gender, ethnicity, etc.) and environmental variables (e.g., living area, pollution degree, dietary habits, etc.) on the accuracy of the research results, a series of studies based on random forest algorithm have emerged. Some studies have constructed random forest models by incorporating population characteristics to analyze the confounding effects in the association between genomic data and diseases. Some use random forests to integrate environmental variables to explore the association mechanism between genomic data and diseases under their interference. There are also studies that integrate population and environmental variables and use random forests to dig deep into the association between genomic data and disease. With the help of random forest methods, these studies are committed to accurately revealing the potential connection between genomic data and diseases in the context of complex confounding factors and providing a solid theoretical basis for disease prevention, diagnosis and treatment.

Random forest is implemented on the basis of a decision tree, a supervised ensemble learning method of the Bagging class. By integrating multiple decision trees, prediction accuracy is improved (Wang *et al.*, 2024). A decision tree is a tree structure based on information entropy. A feature is divided into different subtrees according to the amount of information. When all samples in the subtree belong to the same category, the division is stopped. Commonly used feature division algorithms include ID3, C4.5, CART, etc. (Ma *et al.*, 2024).

The amount of information I is used to measure the amount of information contained in the event X, x_{i} is the feature vector information entropy of the *i-th* sample and p_{i} refers to the probability of the occurrence of x_{i}. The calculation formula is shown in Eq. (1):

$$I\left(X=x_i\right)=-\log p_i \tag{1}$$

The information entropy *H (x)* calculation formula is shown in Eq. (2), where *i* is the index of samples in the dataset and *k* is the total number of events:

$$H\left(x\right)=\sum_{i=1}^{k}p_iI\left(X=x_i\right)=-\sum_{i=1}^{k}p_i\log p_i \tag{2}$$

ID3 Each feature selection is split with the one with the largest information gain; that is, the information gain is calculated for each feature and the feature with the largest information gain is split and selected each time to eliminate the uncertainty between samples. The information gain IG is shown in formula (3), H *()* represents the information entropy solution operation and X is the input sample:

$$IG\left(F\right) = H\left(x\right) - \sum_i \frac{|X_i|}{|X|} H\left(X_i\right) \tag{3}$$

$$H\left(x\right) = -\frac{3}{5}log\frac{3}{5} - \frac{2}{5}log\frac{2}{5} \tag{4}$$

H (X) computed samples exhibit different classes of information entropy independent of feature partitioning. The calculation formula is shown in Eq. (4). The conditional entropy is obtained by subtracting the feature-related terms. The calculation formula is shown in Equation (5). $H\ (x_1)$ is the sample partition with a value of 1 on the feature of the data set, $H\ (x_2)$ is the partition with a value of 2, $H\ (x_3)$ is the partition with a value of 3 and F represents the specific feature function value. As shown in Eq. (5):

$$\sum_i \frac{|X_i|}{|X|} H\left(X_i\right) = H\left(X|F\right) = \frac{1}{5}H\left(X_1\right) + \frac{2}{5}H\left(X_2\right) + \frac{2}{5}H\left(X_3\right) \tag{5}$$

C45 is improved on the ID3 method. Considering that the ID3 algorithm is prone to overfitting when there are too many features, it is divided by the feature information entropy, as shown in Formula (6). For missing values, *C45* divides them into each node and different child nodes with different probabilities:

$$C45\left(F\right) = \frac{IG(F)}{H(F)} \tag{6}$$

C45 discretizes continuous features by using the locus-in-continuous features method, which improves the shortcoming that ID3 cannot handle continuous features, but it still has problems. Its computational complexity is larger than ID3, it is not friendly to processing continuous values and it can only deal with classification problems, but not regression problems. The core of CART is to pay attention to the error probability of sample prediction and calculate pairs of different probabilities in samples. The calculation formula is shown in Eqs. (7-8), where *n* is the number of categories and p_i is the successful probability of category *i* prediction. The smaller the value, the greater the possibility that the samples are of the same kind:

$$Gini\left(F_j\right) = \sum_{i=0}^{n} p_i\left(1-p_i\right) = 1 - \sum_{i=0}^{n} p_i^2 \tag{7}$$

$$s.t. \sum_{i=1}^{n} p_i = 1 \tag{8}$$

Gini refers to the Gini coefficient. F_j is a specific step in the calculation of characteristic function and conditional entropy. *s.t.* stands for constraint function. Generally, the multi-classification problem is decomposed into a binary classification problem. Assuming that the occurrence probability of the first category on the F_0 feature partition is *p*, the occurrence probability of the second category on F_0 is *1-p* and the *Gini* coefficient on the F_0 feature data set is shown in Formula (9):

$$Gini\left(F_0\right) = 1 - p^2 - \left(1-p\right)^2 = 2p\left(1-p\right) \tag{9}$$

Similarly, assuming that the occurrence probability of the first category on the F_1 feature division is p_2, the Gini coefficient of the data set on the F_1 feature is shown in the Formula (10):

$$Gini\left(F_1\right) = 1 - p_2^2 - \left(1-p_2\right)^2 = 2p_2\left(1-p_2\right) \tag{10}$$

p represents the probability of feature occurrence. In practice, the weighted method is generally used for calculation and the sample *X* is divided into X_i by the *F* eigenvalue. The Gini coefficient of the total sample can be obtained by weighting the average of the Gini coefficients on these X_i, as shown in Eq. (11) where *n* represents the number of categories:

$$Gini\left(F\right) = \sum_{i=1}^{n} \frac{|X_i|}{|X|} Gini\left(F_i\right) \tag{11}$$

To sum up, the three feature-splitting algorithms are summarized in Table (1). Each algorithm has its applicable scenario and a feature-splitting algorithm needs to be selected for different data features.

Table 1: Three feature splitting algorithms

Algorithm	Processing Value	Support model	Core Computing	Pruning strategy
ID3	Discrete	Classification	Information gain	Without
C4.5	Continuous	Classification	Information gain ratio	Post-pruning
CART	Continuous	Classification, regression	Gini coefficient	Post-pruning

In order to prevent decision tree overfitting, two pruning strategies- pre-pruning and post-pruning-are usually applied (Ren *et al.*, 2024). Pre-pruning is implemented during decision tree growth to terminate branch expansion early by limiting the height of the tree or the minimum number of samples in the nodes. Then, pruning is optimized after the decision tree is completely constructed and the subtree parts that do not affect the overall accuracy are removed. The calculation of pre-pruning is simple, but the effect is limited. Although post-pruning consumes more computing power, it can provide better results (Ranjan *et al.*, 2024).

The random forest model is shown in Figure (1) and its construction process is as follows: First, the Bootstrap algorithm is used to extract multiple subsets from the original data set. After that, the decision trees on each subset are trained according to specific parameters. Finally, the predictions of each tree are summarized and the results are decided by majority vote (Liu *et al.*, 2024; Zou *et al.*, 2024). Among them, the quality of the decision tree directly affects the performance of the whole random forest, especially the selection of feature segmentation rules is crucial. The model introduces the concept of out-of-bag error (OOB error); that is, all samples are not used when building a tree and a part is reserved for subsequent evaluation to ensure that about 1/3 of the data does not participate in training, so as to objectively evaluate the accuracy of the model.

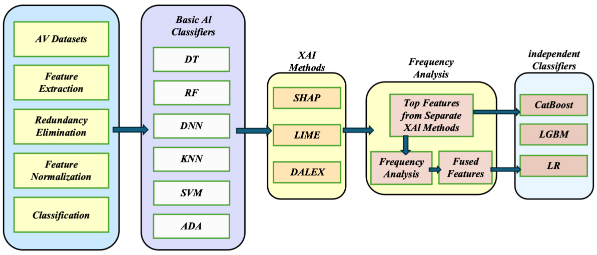

Fig. 1: Random forest processing flow

The strength of random forest lies in its extremely high randomness ability to be suitable for large-scale data and its high prediction accuracy (Sankaranarayanan and Senthilkumar, 2024). Of course, it may also be overfitted on some data sets, resulting in poor actual performance. Especially when processing feature dimensions are high, the classification ability of random forest may be limited and additional tuning is required to give full play to its effectiveness.

Genomic Data Analysis Process

Data Preparation and Preprocessing

Genomics research often faces the problem of missing data due to experimental errors, difficulty in obtaining samples, or failure of collection technology, which will seriously interfere with random forest-based genomic data analysis and disease association research, resulting in biased or erroneous results, so it is extremely important to properly handle the missing data to ensure the integrity and accuracy of the analysis. In this study, the random forest-based Multiple Imputation Method (MICE-RF) was adopted, which combined the advantages of multiple imputation and random forest and could accurately estimate the missing values. The process is as follows: First, the original genomic data is standardized and the outliers are processed with boxplots; Then, the complete data is used to construct the initial model of random forest, the number of decision trees is set to a large value and the maximum number of features considered when each tree is split is taken as the square root of the total number of variables. Then, other variables except the variables where the missing values are located are taken as inputs and the missing values are predicted by the model to complete the preliminary filling. Then, in the initial fill dataset, 5-10% of the data points for each variable were randomly selected each time and set as missing and filled again and multiple complete data sets were generated by repeating multiple times and the number of interpolations was adjusted between 5 and 20 times according to the data situation. Finally, genomic data analysis and disease association studies were carried out on multiple complete datasets, such as calculating the strength of gene and disease association, making feature selection, etc. and the results were combined with mean and median to improve reliability. The effect of this method can be judged by plotting the distribution histogram and scatter plot of the data before and after interpolation to see whether the distribution of the dataset is similar, calculating statistical indicators such as mean, variance and correlation coefficient to see the size of the difference and using the imputed data to construct the disease association prediction model and the model with the original complete data to compare the accuracy, recall, F1 value and other performance indicators.

Preparing data plays an important role in exploring Random Forest (RF)-based genomic data analysis and disease association research (Zhang *et al.*, 2024).

In terms of data collection and integration, the first task is to obtain genomic data from multiple channels, including but not limited to gene sequences, mutation information, protein structure data, etc., in public databases (such as GEO, dbSNP) and personal laboratory databases. Then, convert data from different sources in a unified way and establish a compatible and standard data framework for subsequent analysis.

In terms of data cleaning and verification, all obtained data are preliminarily screened and non-specific data, invalid samples and records with serious deviations are eliminated to ensure the overall quality of the data set. Statistical analysis is performed for missing values and deletion or imputation strategies are adopted according to the specific situation to avoid analysis bias caused by data blanks. By implementing standardization or normalization operations to deal with the huge magnitude gap among various biological indicators, so as to eliminate the interference caused by unit inconsistency. For categorical variables or sequential data, monothermal coding or ordinal coding is used to convert them into machine-readable forms to enhance the mathematical representation power of the data.

In terms of feature engineering, with the help of statistical methods such as PCA and LASSO regression, subsets of features highly related to target diseases are mined from massive genomic data, simplifying data dimensions and improving computational efficiency. Combining domain knowledge with model importance scores, the feature combination is further optimized to ensure that the model can capture complex patterns without overfitting existing data. In view of the phenomenon that the number of samples in the disease state is much smaller than that in the healthy control group, the proportions of the two groups are equalized by undersampling, oversampling or SMOTE techniques to improve the prediction accuracy.

Analysis Operation Based on Random Forest Algorithm

Stem covers a variety of different types of entities and various types of interrelationships established among these entities (Brennan *et al.*, 2024; Qiu *et al.*, 2024). In recent years, data construction and analysis techniques in heterogeneous information networks have been widely used in the research of biology (Barnett *et al.*, 2024). In order to deeply explore the interconnection between lncRNAs and diseases, a heterogeneous network of lncRNAs as well as a heterogeneous network of diseases, called MHILDA, was designed. MHILDA calculated the functional similarity of lncRNA and the semantic similarity of diseases and saved these data in matrices LL and DD. In addition, in order to enhance the connectivity between heterogeneous networks, MHILDA also integrated the LDAs network, the lncRNA-miRNA association network and the miRNA-disease associations network and stored them in the matrix LD, LM and MD (Du *et al.*, 2024; Gatasheh, 2024).

In random forest-based genomic data analysis studies, the identified genomic markers have significant value in clinical applications. It can assist in disease diagnosis, such as in complex genetic diseases, the combination of specific gene markers as biomarkers, more specific and sensitive than traditional indicators, taking breast cancer as an example, related genetic markers can accurately identify the risk of disease at an early stage with the help of genetic testing and improve the cure rate and survival rate; For example, specific gene markers in lung cancer patients can help doctors judge the possibility of postoperative recurrence, so as to facilitate the formulation of plans in advance to improve the prognosis. Due to the different genetic backgrounds of patients, the genetic markers related to drug efficacy and adverse reactions determined by random forest analysis can help doctors select the most suitable drugs and dosages for patients. At the same time, it can be used for disease risk assessment, for diseases with genetic predisposition, the detection of specific genetic markers can predict the likelihood of disease, such as people with a family history of cardiovascular disease can assess the risk through the detection of relevant genetic markers and high-risk individuals can use this to adjust their lifestyles and regular physical examinations to prevent diseases.

In this research on genome data analysis and disease association based on random forests, we have successfully identified a series of key Single Nucleotide Polymorphism (SNP) loci significantly associated with the target disease. To present the distribution characteristics of these key SNPs more intuitively, we visualized their genomic positions. Using specialized genomic visualization tools, each key SNP was accurately located and labeled on the corresponding chromosomal region with chromosomes as units, clearly showing their distribution in the genome, including chromosome numbers and specific position coordinates. In terms of biological relevance, the genomic regions where these key SNP loci are located are not randomly distributed. Some SNPs are located in the known gene coding regions, which means they may directly affect the coding sequence of proteins, thus changing the structure and function of proteins and influencing normal cell physiological activities, thereby being related to the occurrence and development of diseases. Other SNPs are located in gene regulatory regions such as promoters and enhancers, where they can regulate the activity of related genes by affecting gene expression levels during transcription or translation, indirectly participating in the pathological process of diseases. In addition, some key SNPs are in non - coding RNA regions and may regulate the function of non-coding RNAs through interactions with them, playing an important role in the pathogenesis of diseases. Through the visualization of the genomic positions of these key SNPs and in - depth biological correlation analysis, important clues and bases are provided for us to further understand the genetic mechanisms of diseases.

In order to predict potential LDAs more accurately, MHILDA adopted the already constructed disease model and lncRNA heterogeneous network to integrate the sample vector representation of LDAs and successfully fused the disease heterogeneous network with the heterogeneous network of lncRNAs.

Each sample is presented as a feature vector and a feature matrix of training samples is constructed on this basis. The integrated sample feature matrix contains a lot of correlation information and at the same time, it may produce noise and redundant data, which may also cause the predictor to encounter difficulties in capturing key information. Therefore, in view of this situation, MHILDA adopted the Lasso algorithm developed by Robert as a reference. Lasso designed a penalty function to reduce some regression coefficients to ensure that the sum of the absolute values of these coefficients does not exceed a certain fixed value so as to construct a more accurate model. In addition, on the one hand, Lasso uses regularization technology to achieve sparsity and feature selection and resets the weights of some irrelevant features to 0. On the other hand, by quantifying the correlation between each feature and the sample label, the weight coefficient of irrelevant or redundant vectors is reduced to zero and then the matrix is compressed to extract key features more effectively (de la Cruz-Ruiz *et al.*, 2024).

This project first uses the training sample feature matrix with dimensionality reduction processing to train the RF and uses the trained model to score the potential LDAs relationship. A 5-fold experiment was used afterward to evaluate the predictive power of MHILDA for unknown LDAs. In each 5-fold validation cycle, the MHILDA model randomly selects negative samples from the unknown sample of LDAs that match the positive sample size and scores other unknown relationships after these negative samples are removed. Moreover, ambiguous LDAs will not be applied to the training and testing of the MHILDA model at the same time, preventing the problem of model overfitting. The next step is to ensure that the MHILDA model can accurately predict all unknown associations, select 10 5-fold verification operations and score ambiguous associations. Finally, MHILDA selected lncRNA related to specific diseases as cases for in-depth analysis based on the average score of candidate sample associations. The MHILDA prediction model flow is shown in Figure (2).

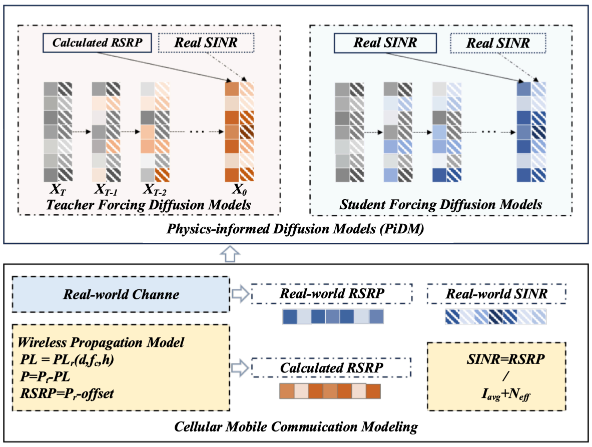

Fig. 2: Flowchart of prediction model

In the study of genomic data analysis and the use of random forests for disease association, a common problem of missing data in genomic datasets was encountered. To solve this problem, the experiment uses a multi-imputation method, in which an algorithm generates multiple reasonable values for each missing data point based on the existing data pattern. By creating these imputed datasets, we manage the uncertainty associated with missing values. Applying the random forest algorithm to each dataset not only enhances the stability of the model, but also provides more comprehensive insights into the relationships between the data. We've also enhanced our validation strategy to go beyond simple cross-validation by implementing a nested cross-validation approach. In the outer loop, the data is divided into model evaluations and in the inner loop, the hyperparameters of the random forest model are adjusted to optimize accuracy and generalization ability. Our findings are biologically important because key genomic markers associated with specific diseases were identified through random forest analysis. These markers may be involved in fundamental biological processes such as cell signaling, gene regulation and metabolism; For example, the discovery of certain disease-associated Single Nucleotide Polymorphisms (SNPS) may imply disruption of key cellular functions associated with that disease. From a clinical perspective, these findings may be relevant. Identifying these disease-relevant genomic markers can play a role in biomarkers for early disease detection, allowing clinicians to develop more targeted screening tests for earlier intervention and improved patient outcomes. In addition, understanding the affected biological pathways may lead to new treatment strategies in which drugs can be designed to target specific dysregulated pathways and correct underlying biological defects to provide more effective treatment options for these patients.

Materials and Methods

In this study, a large-scale case-control dataset was constructed, in which the case groups were screened strictly according to the International Classification of Diseases (ICD) and the control group was matched with the case groups by 1:1 frequency in terms of age, gender and geography to ensure that the samples were well comparable. The sample covered 5 ethnic groups from 3 different geographical regions, with a total of 2000 samples (1000 cases and 1000 controls) and detailed demographic and lifestyle characteristics such as age, gender, smoking history and alcohol history were recorded. If there is significant population stratification, the Principal Component Analysis (PCA) method is used to correct. In the process of phenotypic data collection, the disease phenotype is comprehensively determined through various means such as diagnosis by professional physicians and laboratory testing to ensure the accuracy and reliability of the data. Genomic data was acquired using the Illumina Infinium Global Screening Array array platform, which detected a total of 850,000 SNP loci. In the data preprocessing stage, samples with a deletion rate of more than 5% are directly eliminated; SNPs are filtered if the deletion rate is >5%, the Minimum Allele Frequency (MAF) is <0.01, or the Hardy-Weinberg Equilibrium (HWE) test p-value $<1\times10^{-6}$. The genotype was coded using an additive model (0/1/2 representing homozygous, heterozygous and variant alleles homozygous for the reference allele, respectively) and 620,345 high-quality SNP loci were ultimately retained for subsequent analysis.

In this study, random forest algorithm was used to analyze genomic data. Based on the idea of ensemble learning, random forests can reduce model variance and improve prediction accuracy by constructing multiple decision trees and summarizing prediction results. In the process of decision tree construction, the Gini impurity was used as the node splitting criterion and some samples and features were randomly selected for the growth of each tree to avoid overfitting. The key parameters are set to 1000 to ensure the stability of the model. Set the minimum number of nodes (nodesize) to 5 to ensure that the branches of the tree are sufficiently representative. The feature sampling ratio (mtry) is set to the square root of the total number of features, i.e., about 788 SNPs, balancing the complexity of the model with the generalization ability.

The 5-fold cross-validation strategy is used to divide the training set and the test set and the dataset is divided into 5 subsets each time, of which 4 subsets are used for training the model and the remaining 1 subset is used for testing and the cycle is 5 times to ensure that each sample has the opportunity to participate in training and testing and finally the average of the 5 results is taken as the model performance indicator. In the feature selection process, a preliminary screening was carried out based on MAF and HWE to remove SNPs that did not meet the conditions. The Mean Decrease Accuracy and Mean Decrease Gini index of each SNP were calculated by random forest to quantify their importance for disease prediction. In order to control the false positives caused by multiple tests, the Benjamini-Hochberg procedure was used for False Detection Rate (FDR) control and the FDR threshold was set to 0.05.

Disease association analysis consists of single SNP association analysis and multiple SNP combination analysis. The single SNP association analysis used the traditional logistic regression model, with the case-control status as the dependent variable and the genotype of a single SNP as the independent variable, while adjusting for potential confounding factors such as age and gender and calculating the Odds Ratio (OR) and 95% Confidence Interval (CI) of each SNP's association with the disease and the p-value $< 5\times10^{-8}$ was considered to reach the genome-wide significance level.

Multi-SNP combinatorial analysis uses random forest algorithm to mine SNP interactions and epistatic effects. By ranking the importance of SNP features in the random forest model, the top 100 SNP loci were screened out and the Multivariate Dimensionality Reduction (MDR) method was used to further explore the association pattern between these SNP combinations and disease risk. Combined with the analysis results of random forest and MDR, the complex association between SNPs and diseases is comprehensively analyzed and accurate disease association research and SNP marker identification are realized, which provides a theoretical basis for the exploration of disease genetic mechanism and biomarker discovery.

Results

This project provides an in-depth performance evaluation of the predictive capabilities of the MHILDA model. To verify the prediction accuracy of MHILDA and ensure the fairness of the experimental results, we divide the dataset into benchmark datasets and independent test sets. The data set studied included 2697 positive samples and 96183 unlabeled samples. Of these, 80% of the datasets were selected as benchmark datasets, while the remaining 20% were used as independent test sets. In order to achieve a fairer classification, 20% positive samples and 20% unlabeled samples are divided into two independent groups of test data. The remaining 80% of positive samples and unlabeled samples were classified as the benchmark dataset. Table (2) has showed the 5-fold validation results on benchmark dataset.

Genomic markers determined based on random forest analysis play a significant role in clinical applications and in the diagnosis of adjuvant diseases, their specific combinations are used as biomarkers in complex genetic diseases, which are more specific and sensitive than traditional indicators, such as breast cancer-related gene markers can accurately identify the risk of disease at an early stage through genetic testing and improve the cure rate and survival rate. For example, specific gene markers in lung cancer patients can help doctors judge the possibility of postoperative recurrence and formulate plans in advance to improve long-term prognosis. In guiding personalized treatment, due to the different genetic backgrounds of patients, the genetic markers related to drug efficacy and adverse reactions determined by random forest help doctors select appropriate drugs and dosages for patients and adjust the plan according to the sensitivity of genetic markers to chemotherapy drugs in cancer chemotherapy, so as to achieve precision medicine. For example, people with a family history of cardiovascular disease can assess the risk by testing relevant markers, so that high-risk individuals can adjust their lifestyle and have regular physical examinations to prevent disease. In order to deeply evaluate the predictive ability of MHILDA, a 5-fold check was performed on the standard dataset and 2158 positive samples and 76946 unlabeled samples were found in the benchmark dataset. The experimental data showed that MHILDA had an ACC performance of 89.69% on the standard dataset, Sen had a 90.46% accuracy in the classification of true positive samples and its F1 and MCC data were 89.45% and 80.23%, respectively. Some SNP loci are of great significance in random forest-based genomic data analysis and disease association studies.

Table 2: 5-fold validation results on benchmark dataset

Test Set	ACC (%)	Sen (%)	F1 (%)	MCC (%)	AUC (%)
1	93.683	95.191	93.122	84.625	92.851
2	92.695	90.917	91.884	80.142	93.361
3	91.842	94.380	92.310	80.163	94.016
4	93.496	93.610	92.914	86.018	95.098
5	94.650	96.304	94.890	86.268	94.994
Average	93.278	94.078	93.028	83.439	94.068

Table 3: Functional roles identified by SNPs

SNP	Functional role
rs6511723	Located in the APOE-C1/C4/C2 region, it is involved in the transport process of cholesterol
rs1121980	Similar to the FTO gene, the FTO gene can affect weight and obesity risk by encoding fatty acid oxidase
Top 50 High-Risk SNPs (Overall)	The main focus is on genes known to be involved in lipid metabolism, inflammatory responses and immunomodulatory pathways

As shown in Table (3), e.g. rs6511723 is located in the APOE - C1/C4/C2 region and is involved in the transport process of cholesterol; rs1121980 is similar to the FTO gene, which can affect weight and obesity risk by encoding fatty acid oxidase. In addition, the top 50 high-risk SNPs (overall) are primarily focused on genes that have been clearly involved in lipid metabolism, inflammatory responses and immunomodulatory pathways. These findings could help further explore the association between genomes and disease.

It can be observed from Figure (3) that the ROC curve of MHILDA based on 5-fold validation is close to the true prediction interval and the AUC reaches 89.28, 89.77, 90.40, 91.44 and 91.34%, respectively, with an average value of 90.45%, which indicates that MHILDA has stable performance. According to the above experimental data, the MHILDA model shows excellent performance on the reference dataset and can accurately predict potential LDAs.

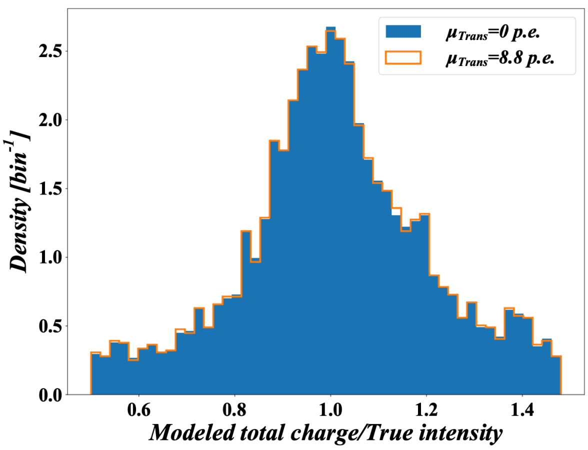

Fig. 3: ROC diagram based on 5-fold verification

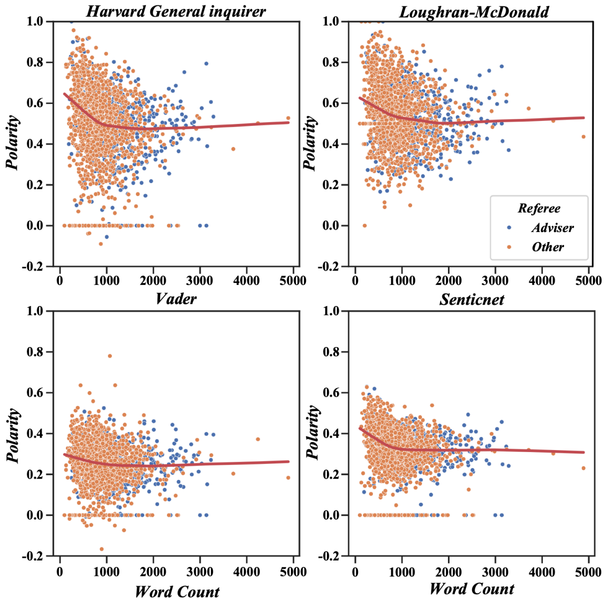

Fig. 4: Comparison with other feature extraction methods

Multi-information fusion technology can extract various kinds of information from LDAs in all directions, but it may also lead to redundancy and noise of information, which adversely affects the predictive ability of the model. In the computing model of this study, the core link is to efficiently extract key features and reasonably eliminate redundant information in the data set, deeply explore the potential value of the data and further improve the computing speed of the model. In this part, Lasso and four other similar feature extraction techniques are experimentally compared with extra-Trees (ETS), SVD, Multiple Dimensional Scaling (MDS) and Local Linear Embedding (LLE). The datasets studying the key features extracted by the Lasso, ETS, SVD, MDS and LLE methods are respectively put into the predictor, the AUC values are compared and the ROC curve shown in Figure (4) is plotted. In these methods, the AUC values of 0.9045, 0.8930, 0.9123 and 0.9215 were achieved for Lasso, ETS, SVD, MDS and LLE, respectively.

In order to achieve the best prediction effect, MHILDA uses miRNAs as the source of its biological data compared with previous studies that only focused on lncRNAs and diseases. To examine the validity of the miRNAs data source, MHILDA was trained and tested on the miRNAs-removed dataset and the miRNAs-unremoved dataset. As shown in the experimental data in Figure (5), the ROC curve area of the miRNAs-added dataset significantly exceeds the ROC curve area of the miRNAs-removed dataset. In addition, the AUC values obtained by MHILDA were 0.8613 and 0.9045 in the data sets with miRNAs removed and without miRNAs removed, respectively. The data showed that by introducing miRNAs biological data sources, the AUC value of the MHILDA model was improved by 0.0432, further confirming that the introduction of miRNAs is both appropriate and efficient.

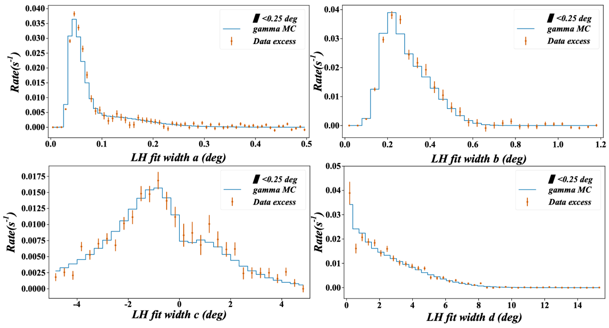

Fig. 5: ROC comparison with other predictors

To ensure the fairness of the experimental process, all models were predictively analyzed on their respective test sets and the corresponding AUC values are shown in Figure (6). On the independent test set, the AUC value of MHILDA was 0.9203, while the AUC values of the other models were 0.8579, 0.8822, 0.8116, 0.8289, 0.7313 and 0.8750, respectively, which were 0.0624, 0.0381, 0.1087, 0.0914, 0.189 and 0.0453 higher than the AUC values of the other six prediction models of the same type, respectively. Under the 5-fold condition, the comparison results of MHILDA with six other methods on independent test sets show that MHILDA shows significant superiority in predicting possible LDAs, proving that it is an effective strategy to mine potential LDAs.

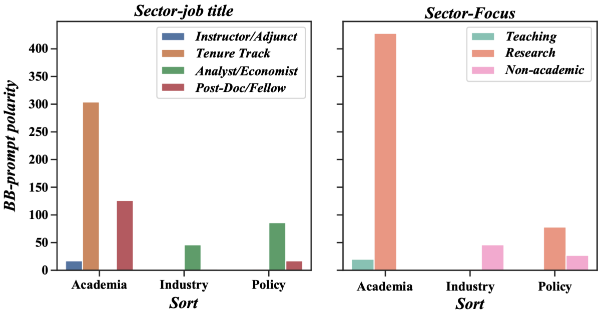

Fig. 6: AUC values with other prediction models

As shown in Figure (7), RF achieved the highest AUC value of 0.9196, while ETS, XGB and AdaBoost achieved AUC values of 0.8965, 0.9054 and 0.9012, respectively. Experimental data show that compared with other learners, RF is more suitable as a learner for the idenLD-AREL model and has advantages in prediction.

Figure (8) shows pairs of gene numbers in the five datasets before and after performing gene filtering. It can be observed from this that the number of features in the dataset decreases by about 80%-90°C after filtering the features of Filter. By filtering out a large number of noisy genes and irrelevant genes, the burden of subsequent Wrapper feature selection is successfully reduced.

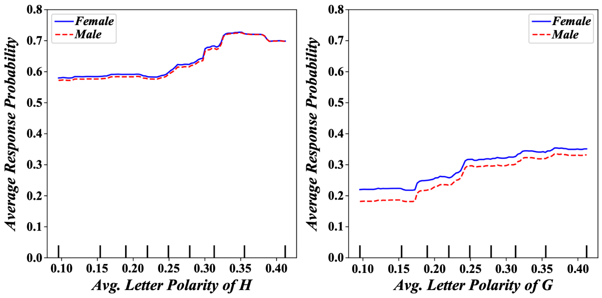

Fig. 7: Comparison of experimental results

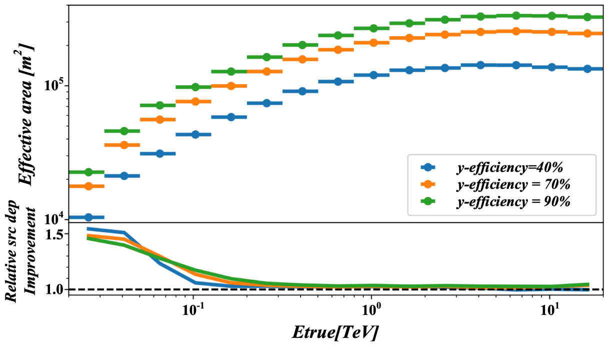

Fig. 8: Characteristics of the dataset after gene filtering

Figure (9) shows that on the Leukaemia, Prostate, Breast and DLBCL datasets, the classification accuracy of the Stratify-RF method far exceeds that of the SVM-RFE, GA-KNN, mRMR, SVM-RCE, Ensemble and Multi-stage methods; In Nervous dataset, although the classification accuracy of Stratify-RF algorithm is slightly inferior to that of Multi-stage method, it is overall superior to other technologies, which proves that the feature selection strategy in this study has an excellent performance in classification.

Figure (10) shows the classification accuracy of various feature search methods. Based on the average results of 10 experiments, the original data of each experiment is randomly divided into 80% training set and 20% test set. The data show that on the Leukaemia, Breast, Nervous and DLBCL datasets, the hierarchical feature search method combined with SVM classification in this study performs better than SBS and SFS. However, on the Prostate dataset, the classification accuracy of SVM is slightly lower than that of SFS.

Figure (11) shows that the classification performance of the Stratify-SVM algorithm on the Leukemia dataset is first rising, then falling and then rising to stability. It shows that with the removal of redundant and irrelevant genes, the performance of the classifier is enhanced, but the performance of the system decreases significantly when the number of genes is less than 50. This finding is consistent with the conclusion of calculation methods based on SFS and SBS. The algorithm for the hierarchical elimination of redundant features has higher classification accuracy than the algorithms based on SFS and SBS.

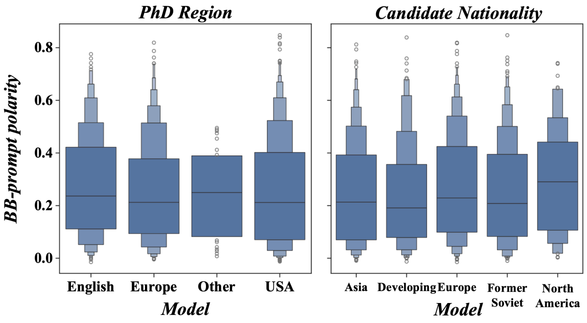

Fig. 9: Classification performance of different feature selection algorithms

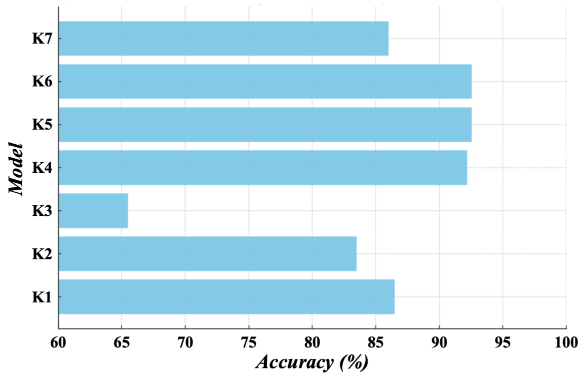

Fig. 10: Classification accuracy of different feature search strategies

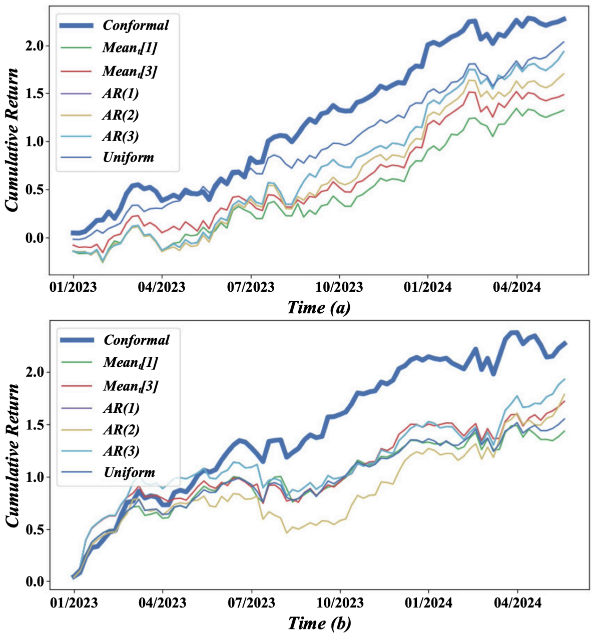

Fig. 11: Comparison of algorithm classification performance

Discussion

In this study, a large-scale case-control dataset containing 2000 samples (1000 cases and 1000 controls) from 5 ethnic groups in 3 geographical regions was constructed and the case groups were screened according to the International Classification of Diseases (ICD), the control group was matched by 1:1 frequency in terms of age, gender, geography, etc. and demographic and lifestyle characteristics were recorded, the Principal Component Analysis (PCA) was used to correct the population stratification and the phenotypic data were reliable through professional diagnosis and detection. Data from 850,000 SNPs were acquired using the Illumina Infinium Global Screening Array array platform, filtered to an additive model encoding genotype with a sample deletion rate of >5%, SNP loci deletion rate of >5%, Minimum Allele Frequency (MAF) <0.01, or Hadi-Weinberg Equilibrium (HWE) test p-values $<1\times10^{-6}$ to encode genotypes in additive models, ultimately retaining 620,345 high-quality SNPs for analysis. The random forest algorithm was used for genomic data analysis, 1000 decision trees were constructed based on ensemble learning, nodes were split by Gini impurity, the performance of the model was balanced by setting the minimum number of samples of the nodes to 5 and the feature sampling ratio of about 788 SNPs, the model was evaluated by 5-fold cross-validation, the SNP importance was quantified by the average reduction accuracy and the average reduction Gini index and the False Discovery Rate (FDR) was controlled by the Benjamini-Hochberg program. In the disease association analysis, the logistic regression model was used to adjust for confounding factors for single SNP associations and the p-value was significant $<5\times10^{-8}$. Multi-SNP combination analysis uses random forest to screen the top 100 important SNP loci, combined with Multivariate Dimensionality Reduction (MDR) to explore the association between SNP combination and disease risk, so as to achieve accurate disease association research and SNP marker identification.

Conclusion

As a powerful machine-learning model, random forest has shown excellent performance in bioinformatics, especially in the field of genomic data analysis. This study focuses on exploring its potential to identify the relationship between gene expression patterns and specific diseases, with a focus on its ability to improve prediction accuracy and handle complex biological networks.

The study used gene expression profiling data from publicly available databases, covering tens of thousands of human samples and covering a wide range of common conditions. A subset of data containing healthy controls and patient populations is carefully selected to ensure comprehensiveness and diversity in model training and testing. In the data preprocessing stage, the influence of dimensionality is eliminated through standardization, normalization and other operations and the analysis quality is improved. During the analysis, the random forest algorithm was used to select features, identify key gene markers and construct a prediction model. At the same time, the grid search strategy is used to optimize the number of trees, node splitting threshold and other hyperparameters to ensure the generalization ability and stability of the model.

Experimental results show that random forests perform well in classification tasks, with an average accuracy of 87%, far exceeding the baseline model. In cases such as lung cancer and diabetes, the sensitivity and specificity were 92% and 88%, respectively, fully confirming the potential value of this model in diagnosing complex diseases. In addition, the ranking of feature importance not only revealed several known risk genes, but also discovered some novel candidate sites, which provided clues for subsequent biological verification.

In conclusion, random forests are an effective tool for interpreting genomic data, which can accurately characterize disease-related gene expression patterns and assist clinical decision-making. In the future, we will expand the dataset and integrate multiple sets of biological information to further improve the prediction accuracy of the model and build a more comprehensive and reliable disease early warning system. At the same time, the exploration is combined with advanced algorithms such as deep learning to dig deep into the hidden connections between genes and phenotypes, opening up a new path for personalized medicine.

Acknowledgment

We the publisher for their support in the publication of this research article. We appreciate the efforts of the editorial team in reviewing and editing our work and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Barnett, E. J., Onete, D. G., Salekin, A., & Faraone, S. V. (2024). Genomic Machine Learning Meta-regression: Insights on Associations of Study Features With Reported Model Performance. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *21*(1), 169-177. https://doi.org/10.1109/tcbb.2023.3343808

Brennan, I. G., Singhal, S., & Bkhetan, Ziad Al. (2024). pipesnake: generalized software for the assembly and analysis of phylogenomic datasets from conserved genomic loci. *Bioinformatics*, *40*(5), btae195. https://doi.org/10.1093/bioinformatics/btae195

Cai, H., Liao, Y., Zhu, L., Wang, Z., & Song, J. (2024). Improving Cancer Survival Prediction via Graph Convolutional Neural Network Learning on Protein-Protein Interaction Networks. *IEEE Journal of Biomedical and Health Informatics*, *28*(2), 1134-1143. https://doi.org/10.1109/jbhi.2023.3332640

de la Cruz-Ruiz, F., Canul-Reich, J., Rivera-López, R., & de la Cruz-Hernández, E. (2024). Impact of data balancing a multiclass dataset before the creation of association rules to study bacterial vaginosis. *Intelligent Medicine*, *4*(3), 188-199. https://doi.org/10.1016/j.imed.2023.02.001

Du, L., Zhao, Y., Zhang, J., Shang, M., Zhang, J., & Han, J. (2024). Identification of Genetic Risk Factors Based on Disease Progression Derived From Longitudinal Brain Imaging Phenotypes. *IEEE Transactions on Medical Imaging*, *43*(3), 928-939. https://doi.org/10.1109/tmi.2023.3325380

Gatasheh, M. K. (2024). Identifying key genes against rutin on human colorectal cancer cells via ROS pathway by integrated bioinformatic analysis and experimental validation. *Computational Biology and Chemistry*, *112*, 108178. https://doi.org/10.1016/j.compbiolchem.2024.108178

Gómez-Méndez, I., & Joly, E. (2023). Regression with missing data, a comparison study of techniques based on random forests. *Journal of Statistical Computation and Simulation*, *93*(12), 1924-1949. https://doi.org/10.1080/00949655.2022.2163646

Gómez-Méndez, I., & Joly, E. (2024). On the consistency of a random forest algorithm in the presence of missing entries. *Journal of Nonparametric Statistics*, *36*(2), 400-434. https://doi.org/10.1080/10485252.2023.2219783

Gronowski, A., Arness, D. C., Ng, J., Qu, Z., Lau, C. W., Catchpoole, D., & Nguyen, Q. V. (2024). The impact of virtual and augmented reality on presence, user experience and performance of Information Visualisation. *Virtual Reality*, *28*(3), 133. https://doi.org/10.1007/s10055-024-01032-w

Gualdi, F., Oliva, B., & Piñero, J. (2024). Genopyc: a Python library for investigating the functional effects of genomic variants associated to complex diseases. *Bioinformatics*, *40*(6), btae379. https://doi.org/10.1093/bioinformatics/btae379

Lemane, T., Lezzoche, N., Lecubin, J., Pelletier, E., Lescot, M., Chikhi, R., & Peterlongo, P. (2024). Indexing and real-time user-friendly queries in terabyte-sized complex genomic datasets with kmindex and ORA. *Nature Computational Science*, *4*(2), 104-109. https://doi.org/10.1038/s43588-024-00596-6

Liu, F., Yang, Y., Xu, X. S., & Yuan, M. (2024). MESBC: A novel mutually exclusive spectral biclustering method for cancer subtyping. *Computational Biology and Chemistry*, *109*, 108009. https://doi.org/10.1016/j.compbiolchem.2023.108009

Luo, Y., Chen, Y., Xie, H., Zhu, W., & Zhang, G. (2024). Interpretable CRISPR/Cas9 off-target activities with mismatches and indels prediction using BERT. *Computers in Biology and Medicine*, *169*, 107932. https://doi.org/10.1016/j.compbiomed.2024.107932

Ma, C., Li, W., Ke, S., Lv, J., Zhou, T., & Zou, L. (2024). Identification of autism spectrum disorder using multiple functional connectivity-based graph convolutional network. *Medical & Biological Engineering & Computing*, *62*(7), 2133-2144. https://doi.org/10.1007/s11517-024-03060-9

Ou, H., Yao, Y., & He, Y. (2024). Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network. *Sensors*, *24*(4), 1112. https://doi.org/10.3390/s24041112

Pinheiro, D., Santander-Jimenéz, S., & Ilic, A. (2022). PhyloMissForest: a random forest framework to construct phylogenetic trees with missing data. *BMC Genomics*, *23*(1), 377. https://doi.org/10.1186/s12864-022-08540-6

Qiu, S., Sun, Y., Guo, J., Zhang, Y., & Hu, Y. (2024). Genome-wide analysis reveals extensive genetic overlap between childhood phenotypes and later-life type 2 diabetes. *Computers in Biology and Medicine*, *171*, 108065. https://doi.org/10.1016/j.compbiomed.2024.108065

R, S., & M, S. (2024). Optimized twin spatio-temporal convolutional neural network with squeezenet transfer learning model for cancer miRNA biomarker classification. *Applied Soft Computing*, *167*, 112270. https://doi.org/10.1016/j.asoc.2024.112270

Rabiei, N., Soltanian, A. R., Farhadian, M., & Bahreini, F. (2023). The Performance Evaluation of The Random Forest Algorithm for A Gene Selection in Identifying Genes Associated with Resectable Pancreatic Cancer in Microarray Dataset: A Retrospective Study. *Cell Journal*, *25*(5), 347-353. https://doi.org/10.22074/CELLJ.2023.1971852.1156

Ranjan, A., Bess, A., Alvin, C., & Mukhopadhyay, S. (2024). MDF-DTA: A Multi-Dimensional Fusion Approach for Drug-Target Binding Affinity Prediction. *Journal of Chemical Information and Modeling*, *64*(13), 4980-4990. https://doi.org/10.1021/acs.jcim.4c00310

Ren, S., Cooper, G. F., Chen, L., & Lu, X. (2024). An interpretable deep learning framework for genome-informed precision oncology. *Nature Machine Intelligence*, *6*(8), 864-875. https://doi.org/10.1038/s42256-024-00866-y

Seyedmirzaei, H., Salmannezhad, A., Ashayeri, H., Shushtari, A., Farazinia, B., Heidari, M. M., Momayezi, A., & Shaki Baher, S. (2024). Growth-Associated Protein 43 and Tensor-Based Morphometry Indices in Mild Cognitive Impairment. *Neuroinformatics*, *22*(3), 239-250. https://doi.org/10.1007/s12021-024-09663-9

Usha Ruby, A., George Chellin Chandran, J., Swasthika Jain, T., Chaithanya, B., & Patil, R. (2023). RFFE - Random Forest Fuzzy Entropy for the classification of Diabetes Mellitus. *AIMS Public Health*, *10*(2), 422-442. https://doi.org/10.3934/publichealth.2023030

Wang, L., Li, Z.-W., You, Z.-H., Huang, D.-S., & Wong, L. (2024). GSLCDA: An Unsupervised Deep Graph Structure Learning Method for Predicting CircRNA-Disease Association. *IEEE Journal of Biomedical and Health Informatics*, *28*(3), 1742-1751. https://doi.org/10.1109/jbhi.2023.3344714

Wang, Y., Song, J., Dai, Q., & Duan, X. (2024). Hierarchical Negative Sampling Based Graph Contrastive Learning Approach for Drug-Disease Association Prediction. *IEEE Journal of Biomedical and Health Informatics*, *28*(5), 3146-3157. https://doi.org/10.1109/jbhi.2024.3360437

You, X., Yang, J., Qin, C., & Liu, H. (2023). Missing data analysis in cognitive diagnostic models: Random forest threshold imputation method. *Acta Psychologica Sinica*, *55*(7), 1192-1206. https://doi.org/10.3724/sp.j.1041.2023.01192

Zhang, L., Li, A., Chen, S., Ren, W., & Choo, K.-K. R. (2024). A Secure, Flexible, and PPG-Based Biometric Scheme for Healthy IoT Using Homomorphic Random Forest. *IEEE Internet of Things Journal*, *11*(1), 612-622. https://doi.org/10.1109/jiot.2023.3285796

Zhang, S., Strayer, N., Vessels, T., Choi, K., Wang, G. W., Li, Y., Bejan, C. A., Hsi, R. S., Bick, A. G., Velez Edwards, D. R., Savona, M. R., Phillips, E. J., Pulley, J. M., Self, W. H., Hopkins, W. C., Roden, D. M., Smoller, J. W., Ruderfer, D. M., & Xu, Y. (2024). PheMIME: an interactive web app and knowledge base for phenome-wide, multi-institutional multimorbidity analysis. *Journal of the American Medical Informatics Association*, *31*(11), 2440-2446. https://doi.org/10.1093/jamia/ocae182

Zhao, C., Esposito, M., Xu, Z., & Zhou, W. (2025). HAGMN-UQ: Hyper association graph matching network with uncertainty quantification for coronary artery semantic labeling. *Medical Image Analysis*, *99*, 103374. https://doi.org/10.1016/j.media.2024.103374

Zou, B., Wang, J., Ding, Y., Zhang, Z., Huang, Y., Fang, X., Cheung, K. C., See, S., & Zhang, L. (2024). A multi-modal deep language model for contaminant removal from metagenome-assembled genomes. *Nature Machine Intelligence*, *6*(10), 1245-1255. https://doi.org/10.1038/s42256-024-00908-5