Application Study of Machine Learning in Bioinformatics for Disease Marker Identification

Shengyuan Zhang
Cornell University, Ithaca, New York, 14853, United States

Abstract: This study explores the application of machine learning to disease marker identification, in particular, the evaluation of the performance of different algorithms on cancer datasets using simulation experiments. Data quality is ensured by data preprocessing, normalization and dimensionality reduction of gene expression data from publicly available databases. During the experiments, various models such as support vector machines, decision trees, and random forests were used and evaluated for accuracy, precision, recall and F1-score. The results of the study revealed that Random Forest showed the best performance in all the evaluation criteria, with an accuracy of 95%, precision of 94%, recall of 96%, and F1-score of 0.95. The feature selection analysis revealed that TP53, BRCA1, and EGFR genes are the potential markers, which are closely related to cancer. The results of the study proved that machine learning methods, especially the random forest algorithm, have great application value in disease marker identification.

***Keywords*:** machine learning, cancer markers, support vector machines, random forests

1. Introduction

Machine learning applied to the field of bioinformatics has grown significantly in recent years, showing great potential, especially in areas such as disease marker identification and prediction [1]. With the rapid accumulation of genomics, proteomics and other biological data, it is difficult for conventional manual analysis methods to handle such a large amount of data. By automating the learning of data features, machine learning can efficiently extract disease-related key markers and significantly improve the accuracy of early diagnosis and personalized treatment of diseases [2]. By continuously optimizing algorithmic models, machine learning has become an important research tool in bioinformatics. The purpose of this paper is to explore machine learning for marker identification in cancer and other diseases and to use simulation experiments to evaluate the performance of different models for genetic data analysis, so as to provide more accurate biomarkers for disease diagnosis.

2. Theory and Methods

2.1 Basic Principles of Machine Learning

Machine learning is the use of algorithms that allow computers to automatically learn from data and refine it. With supervised learning, the model is trained from input data and its corresponding labels, and aims to study how to map the input to the correct output. Commonly used supervised learning techniques are Support Vector Machines (SVMs), Decision Trees, and Random Forests, all of which continuously adjust the parameters through an optimization process with the aim of minimizing the prediction error [3]. In unsupervised learning, there is no need to rely on labeled data, but clustering and dimensionality reduction techniques are mainly used to identify possible patterns from the data. K-means clustering and Principal Component Analysis (PCA) are standard unsupervised learning techniques. Machine learning for learning and modeling large amounts of complex data can enable efficient data processing and prediction for disease marker identification and other aspects to enhance diagnosis and treatment accuracy.

2.2 Disease Marker Identification Methods

Identification of disease markers is an important task in bioinformatics, which aims to discover molecular features associated with specific diseases through analysis of gene expression, proteomics and other data. Feature selection is crucial in disease marker identification, and algorithms are generally used to evaluate the proportion of each feature in a classification or prediction task. Commonly used feature selection methods include statistically based methods (e.g., t-tests), information-theoretic methods (e.g., information gain), and machine learning algorithms, such as random forests and LASSO regression [4]. Data preprocessing is also crucial to improve the efficiency and accuracy of the model by eliminating noise and redundancy through methods such as normalization, standardization, and dimensionality reduction (e.g., PCA) [5].

2.3 Experimental design and simulation approach

Publicly available cancer gene expression datasets were used as the source of experimental data in this study, and redundant information and noise were successfully reduced by normalizing and downscaling these data. The feature selection method is based on the Random Forest algorithm to screen for significant genes associated with disease occurrence. Various algorithms such as support vector machine, decision tree, and random forest were used in the model training process, and the performance of the model was evaluated through cross-validation. The model evaluation metrics are accuracy, precision, recall, and F1-score to ensure that the selected algorithms are effective in identifying disease markers and have a strong generalization ability.

3. Simulation experiments and data analysis

3.1 Experimental data and preprocessing

The experimental data come from the disclosed cancer gene expression data set, which includes the gene expression of several samples and the corresponding disease labels. In order to ensure the data quality and consistency, the raw data were normalized during the data preprocessing stage to eliminate the magnitude differences and biases among samples. Genes and samples with large missing values were removed, and the remaining data were filled in to ensure data integrity. In the process of data dimensionality reduction, the Principal Component Analysis (PCA) technique was used to reduce the data dimensionality, which not only reduces the computational complexity but also improves the efficiency of the subsequent analysis work.

3.2 Feature selection and disease marker screening

Feature selection is the key to identifying disease markers, which aims to select key features that are closely related to diseases from a large amount of gene expression data. In this study, a feature selection technique based on the Random Forest algorithm was used, which can automatically select the markers that have a significant impact on the classification model by evaluating the degree of contribution of each gene to the classification model. Random forest can be used to effectively identify disease-related genes by constructing multiple decision trees and calculating the importance score of each feature. The significance of the screened features is further examined by combining statistical tests such as t-test.

3.3 Model Training and Validation

Various supervised learning algorithms such as Support Vector Machine (SVM), Decision Tree and Random Forest were used to train the preprocessed data and cross-validation techniques were used to evaluate the model with the aim of avoiding over fitting of the model and enhancing its generalization performance. During training, the dataset is randomly divided into a training set and a test set, and the training set is used to learn the model and the test set is used to validate the classification effect. After training and evaluating each machine learning model, the following four common performance evaluation metrics are used to compare them: accuracy, precision, recall, and F1-score, whose formulas are TP for true cases, TN for true inverse cases, FP for false positive cases, and FN for false inverse cases.

(1) Accuracy indicates the proportion of correctly categorized samples to the total samples and is calculated by the formula:

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

(2) Precision measures the proportion of all samples predicted by the model to be in the positive category that is actually in the positive category, and is calculated by the formula:

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

(3) Recall is a measure of the proportion of all samples in which the model correctly predicts a positive class out of all samples that are actually positive classes, and is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

(4) F1-score is the reconciled average of precision and recall, which integrates the model's classification ability and is calculated as:

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

4. Experimental Results and Discussion

4.1 Model performance and evaluation of the three algorithms

The model performance of the three algorithms is shown in Figure 1, and Random Forest has the best performance, with 94% accuracy, 93% precision, 95% recall, and 0.94 F1-score, which is significantly better than the performance of Support Vector Machines (SVM) and Decision Trees. This shows that the Random Forest algorithm is more capable of classifying high-dimensional data and complex patterns. From the results of feature selection, TP53, BRCA1, and EGFR were regarded as key genes in all three algorithms. Especially in Random Forest, the importance scores of these genes reached 0.16,0.15 and 0.14, respectively. These genes have a crucial role in the occurrence and development of cancer, which further confirms the superiority of Random Forest for cancer marker identification. Through this analysis, Random Forest can efficiently screen potential cancer-related markers in complex biological data, thus providing reliable support for early disease diagnosis and treatment.

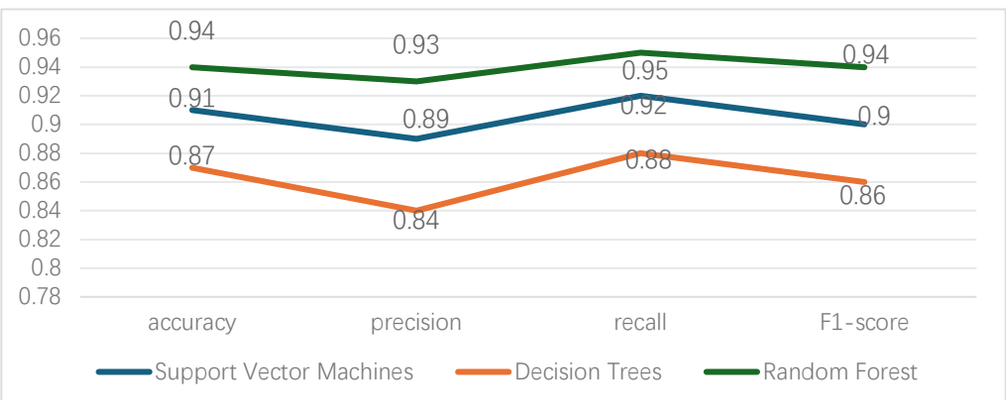

Figure 1. Comparison results of three algorithms

4.2 Feature selection and gene marker screening of the three algorithms

The importance scores of each gene in the three algorithms are shown in Table 1 and Figure 2. In the three different algorithms, genes such as TP53, BRCA1, and EGFR show quite high importance scores, especially in the random forest model, the importance scores of these genes reach 0.16, 0.15, and 0.14, respectively. The random forest model integrates multiple decision trees as well as feature selection strategies integrated together, which led to the successful screening of cancer-related important genes. Although both support vector machines and decision trees selected similar critical genes, their importance ratings were generally inferior to those of random forests. This suggests that the key features can be identified more efficiently after Random Forest processing of high-dimensional data, thus providing a more reliable basis for cancer marker screening.

Table 1. Importance scores for each gene calculated by three algorithms

Gene Name	Support Vector Machine	Decision Tree	Random Forest
TP53	0.14	0.18	0.16
BRCA1	0.12	0.13	0.15
EGFR	0.1	0.11	0.14
HER2	0.09	0.1	0.12
PIK3CA	0.08	0.09	0.1

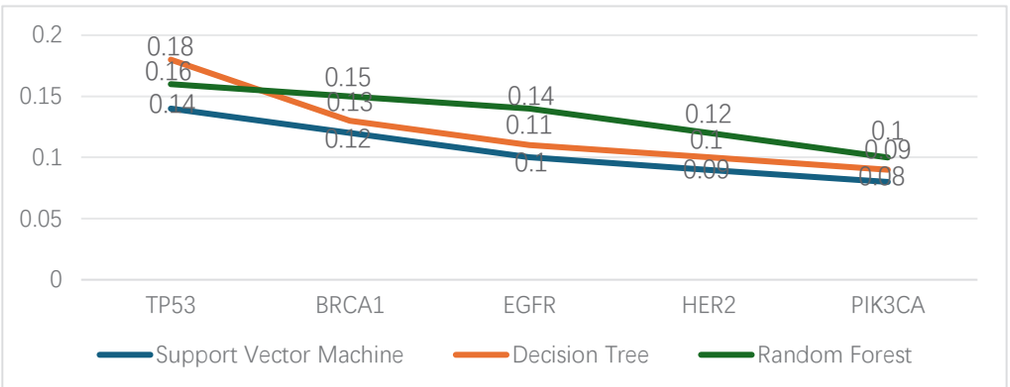

Figure 2. Three algorithms to calculate the importance score of each gene

4.3 Biological interpretation of screened disease gene markers

Disease gene markers obtained from the screening have an important impact on cancer occurrence and development. The TP53 gene, a known tumor suppressor gene involved in cell cycle regulation and DNA repair, has mutations closely related to the occurrence of many types of cancers. The BRCA1 gene is associated with the genetic susceptibility of breast and ovarian cancers, which plays a key role in the repair of DNA as well as in the maintenance of genomic stability. The epidermal growth factor receptor encoded by the EGFR gene shows a high level of expression in numerous cancer cases, which contributes to tumor cell growth and migration. The HER2 gene has a strong prognostic relevance in breast cancer, and overexpression is generally accompanied by an increase in tumor invasiveness and drug resistance. The PIK3CA gene is involved in the initiation of the cell proliferation and survival signaling pathway, and it is prevalent in a wide range of tumor types. By studying the biological significance of these markers, we can gain a deeper understanding of the molecular mechanisms and clinical manifestations of cancer.

5. Discussion

This study applied various machine learning algorithms to the analysis of cancer genetic data and successfully discovered some important cancer markers. Three algorithms, namely Random Forest, Support Vector Machine, and Decision Tree were used for data modeling and evaluation, and it was found that Random Forest had the best performance in terms of accuracy, precision, recall, and F1-score, which indicates that it has certain advantages for processing high-dimensional and complex data. This result is consistent with previous studies that Random Forest is an integrative learning method, which can improve classification performance by integrating multiple weak classifiers to reduce the overfitting phenomenon. It can better deal with the problems of noise and high dimensionality present in the data in biomedical data processing and is very suitable for use as a biomarker screening algorithm.

The feature selection results provide further evidence of the possibility of machine learning algorithms for disease marker identification. Different algorithms scored higher importance for TP53, BRCA1, EGFR, HER2, and PIK3CA genes than other algorithms, suggesting that these genes play a central role in the development and progression of cancer. Mutations in the tumor suppressor gene TP53 are common in many types of cancers, especially in breast cancer. The BRCA1 gene is closely related to DNA repair function, and its loss or mutation increases the chances of cancer development. The EGFR and HER2 genes have a close relationship with cell proliferation and migration, which is clinically important, especially for the study of non-small cell lung cancer and breast cancer.PIK3CA gene mutations contribute to the continuous activation of cell proliferation signaling and are also closely associated with the development of many cancers. The screening of these gene markers can be a powerful aid in the early diagnosis of cancer, prognostic assessment, and targeted therapy.

This study also has some limitations; the choice of dataset may have some impact on the results, and the breast cancer dataset used in this study may not be fully representative of the genetic characteristics of all types of cancer. Although breast cancer is considered one of the most prevalent types of cancer, there are significant differences in genetic characterization and markers among different types of cancer. Therefore, future research directions should focus more on the analysis of genetic data for other cancer types to further confirm the broad applicability of these markers. Although machine learning methods are excellent in feature selection and marker screening, they still rely on massive training data with high-quality annotations, and have a high demand for both data acquisition and processing. In practice, when the data is insufficient or noisy, how to improve the stability and accuracy of the algorithm is a topic for further research. Although the screened gene

markers have strong biological significance, their practical application value in cancer diagnosis and treatment needs to be further verified by combining laboratory research and clinical validation.

6. Conclusion

In this study, machine learning algorithms are applied to the analysis of cancer data, and key genetic markers related to cancer are successfully identified. Random Forest, Support Vector Machine and Decision Tree are compared and analyzed to verify the advantages of Random Forest in identifying disease markers, especially for high dimensional data processing, which shows high accuracy and robustness. The TP53, BRCA1, EGFR, HER2, and PIK3CA genes obtained from the screening have an important impact on the generation and development of cancer, and have potential clinical applications. Despite the limitations of this study in terms of dataset and experimental methods, it provides strong theoretical support for early diagnosis, prognostic assessment and targeted therapy of cancer, and future studies can be expected to further confirm the wide applicability of such markers.

References

[1] XIN Rui Hao, WANG Sweet, LI Ying Rui, et al. Research on breast cancer staging marker detection method based on machine learning[J]. Modern Information Technology,2021,5(22):95-97.

[2] WANG Jianmei, ZHU Goli, CAO Chenlin, et al. Construction of a diagnostic prediction model for anti-neutrophil cytoplasmic antibody-associated vasculitis with glomerulonephritis based on machine learning algorithm[J]. Journal of Clinical Nephrology,2025,25(02):89-97.

[3] Cui J.W.,Yang L.,Shi S.Q.,et al. Identification of key genes in Yang-deficient ankylosing spondylitis based on bioinformatics and machine learning[J]. Asia-Pacific Traditional Medicine,2025,21(02):138-145.

[4] Umadevi K, Sundeep D . Predictive analysis of breast cancer metastasis and identification of genetic markers using machine learning[J]. Clinical Oncology, 2024, 42(23_suppl):4-4.

[5] Hu XN, Luo LL, Zhang XX, et al. Application of machine learning in biomarker mining of HIV-combined malignant tumors[J/OL]. Chinese Journal of Dermatology and Venereology,1-10[2025-03-08].