Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28795

A Systematic Review of Decision Trees Application in the Medical Field

Jiaming Xing

Beijing Etown Academy, Beijing, China
xjmBJEAINTER@outlook.com

Abstract. In the contemporary field of diagnostic and clinical medical practice, the reliance on big data is steadily deepening. However, it is confronted with challenges such as the exponential growth of data volume and the diversification of data types, which impose new requirements for the efficiency of data processing. Against this backdrop, decision trees play an indispensable and pivotal role in precision medicine and disease diagnosis. Therefore, this study aims to systematically sort out the core application directions and underlying principles of decision trees in the medical field through literature review and case analysis methods, while identifying their existing limitations and future development trends. The research findings indicate that decision trees have been widely applied in fields such as stomatology, cardiovascular diseases, and chronic diseases, achieving remarkably positive outcomes. Their value extends beyond disease diagnosis; they also exert a beneficial and positive impact on optimizing treatment regimens and rationalizing the structure of medical treatment costs. The study further highlights the current problems and future development directions.

Keywords: Decision Tree, Disease diagnosis, ID3 Algorithm

1. Introduction

In the era of the vigorous development of big data, the fields of diagnosis and healthcare are also confronted with an unprecedented volume and diversity of data, creating an urgent demand for the efficiency and operability of data processing. Decision trees play a crucial role in disease diagnosis and precision medicine [1]. They can accommodate large datasets for training, enabling rapid judgment based on patients' conditions and the delivery of more accurate results compared with previous methods, while being easy to understand [2]. Prior to this, as the data science discipline had not yet matured, there were relatively few papers that comprehensively summarized the role of decision trees in disease diagnosis and treatment. Most existing literature focused on specific fields, for instance, explanations of decision tree models specifically constructed for diagnosing heart disease.

Therefore, this study employs a case study and literature review based on representative disease cases to outline the role of decision trees in diagnosis and treatment, illustrating their function in the broader healthcare field. Meanwhile, it aims to identify the existing shortcomings of decision trees in healthcare, thereby facilitating better improvement directions for decision tree models in the

DOI: 10.54254/2755-2721/2025.LD28795

medical field. The paper provides primarily theoretical guidance for the diagnosis and treatment of other disease cases that have not yet integrated decision tree model.

2. Related work

2.1. The basic principles of decision trees

In essence, decision tree is a supervised learning model that achieves prediction through feature selection. Its core process is as follows: the Root Node represents the complete dataset. The Internal Nodes serve as decision points for evaluating features. Subsequently, samples are classified into child nodes via Branches based on binary outcomes (i.e., “yes” or “no”). This cycle repeats until a final result is derived [3]. Owing to its straightforward workflows, even interpretable using “if...then...” logical statements, the decision tree has become a widely applied method for diagnosis and treatment in the medical field. Meanwhile, supported by big data and advanced hardware in the current era, decision trees can also learn and generate more branches for feature evaluation, thereby efficiently yielding more accurate results.

This paper will elaborate on the construction process of decision trees, taking the classic ID3 algorithm as an example.

First, for datasets containing discrete data (hereafter, each variable is referred to as a feature variable), the first step is to calculate the entropy to quantify the degree of data disorder (denoted as D). The specific formula is as follows:

$$H(D) = -\sum_{k=1}^{K} p_k log_2(p_k) \tag{1}$$

Here, K denotes the number of class labels, and P represents the proportion of each category among the final class labels. A higher entropy value indicates greater data disorder (i.e., a more uniform data distribution).

Second, select a specific feature variable A and calculate its Conditional Entropy. The formula for conditional entropy is as follow

$$H(D|A) = \sum_{v=1}^{V} \frac{|D_v|}{|D|} \cdot H(D_v) \tag{2}$$

Here, $\mathrm{D_v}$ represents the subset of samples corresponding to the $\mathrm{v-th}$ label value of the feature variable A. Among the notations, $\mathrm{H(D)}$ denotes the entropy of the root node calculated in the first step, and $\mathrm{H(D|A)}$ stands for the conditional entropy of the specific feature variable A. On this basis, the Information Gain is derived.

$$Gain(D, A) = H(D) - H(D|A) \tag{3}$$

The greater the Information Gain, the more significant the role of the feature variable A in reducing the uncertainty of the dataset. Subsequently, the above steps are repeated for all feature variables: the feature with the maximum Information Gain is selected as the basis for splitting the current node, and then the aforementioned process is recursively executed on each branch subset until there are no usable features left or the preset conditions are met. It can thus be concluded that the larger the dataset size, the more accurate the decision tree constructed by the ID3 algorithm will be. This fully demonstrates that decision trees play a crucial role in ensuring the accuracy of disease diagnosis in the era of information explosion [3].

DOI: 10.54254/2755-2721/2025.LD28795

However, the ID3 algorithm still has limitations. First, it is prone to overfitting when the number of data samples is small. Second, ID3 does not support the direct processing of continuous random variables; such variables need to be discretized beforehand. Additionally, ID3 tends to select features with more categories (multi-label features), which may introduce bias into the model, and it is unable to handle missing values in the dataset [4].

2.2. Literature review on decision trees utilization in the medical field

Existing studies have fully demonstrated the effectiveness of decision trees in medical applications. For instance, during the 2005 dengue fever outbreak in Singapore, a decision tree model based on the C4.5 algorithm achieved a negative predictive accuracy of 94.4% in high-incidence scenarios. In low-incidence scenarios (where the disease prevalence was below 10%), its positive predictive accuracy reached 51.1%, and it outperfomed large-scale negative screening in “medical resource conservation” due to its targeted identification capability.

Additionally, research findings based on an experiment using 600 patient records show that the ID3 algorithm achieved an overall predictive accuracy of 74% for common diseases, with the accuracy for “disease-positive” predictions as high as 94%. This performance was significantly superior to that of neural networks (which only achieved 46%). Moreover, due to the model’s strong interpretability, it is more conducive to clinical promotion [5].

3. Typical application scenarios of decision trees in the medical field

3.1. Decision trees in the physical diseases diagnosis

3.1.1. Decision trees in oral disease diagnosis

According to the 2003 report released by the World Health Organization (WHO), oral and dental diseases rank as the fourth most prevalent diseases in industrialized countries. Characterized by strong latency, these diseases often lead to the missed optimal treatment window due to misdiagnosis. Conventional statistical methods are insufficient in efficiency against the backdrop of the explosive growth of medical data. In contrast, decision tree algorithms, featuring rapid response for disease classification and improved diagnostic accuracy, have emerged as a crucial tool for the auxiliary diagnosis of oral diseases.

A research team from the Department of Information Technology (IT) and Computer Science (CS) at IBB University constructed an oral disease diagnostic decision tree model based on the ID3 and J4.8 algorithms. In this model, patients’ basic information—including gender, age, and physical condition—was used as input features, and information gain was employed to screen core diagnostic indicators. Diagnostic rules (e.g., “The presence of a black layer on the teeth may indicate dental calcification") were formulated as the criteria for node selection in the decision tree and the basis for disease diagnosis.

Experimental results showed that among over 200 test samples, the ID3 algorithm achieved an accuracy rate of 79%, while the J4.8 algorithm reached 86%. Notably, the model maintained high efficiency and accuracy even when processing relatively large volumes of data, fully demonstrating the advantages of decision trees in disease diagnosis—convenience, rapidity, and precision. This research not only provides valid evidence for computer-aided diagnosis (CAD) but also offers a valuable reference for the development of diagnostic models for other types of diseases.

DOI: 10.54254/2755-2721/2025.LD28795

3.1.2. Decision trees in cardiovascular disease risk prediction

According to statistical data from the European Public Health Alliance, deaths globally attributed to circulatory system diseases such as heart disease and stroke account for a staggering 41% of the total. Notably, even against the backdrop of continuous advancements in medical technology, the mortality rate of such diseases remains persistently high and even shows an upward trend. The core reason lies not in the scarcity of treatment methods, but in the lack of effective means for early and accurate diagnosis. Meanwhile, data mining in healthcare—defined as the application of algorithms like decision trees for precise diagnosis and medical practice—is a relatively emerging field. Decision tree models not only assist in disease diagnosis but also enable the timely correction of medical errors. Therefore, amid the global high incidence of cardiovascular diseases, decision trees can effectively facilitate early diagnosis and intervention, while simultaneously possessing the advantages of maintaining relatively high accuracy and being easy to operate.

For instance, the School of Engineering and Information Technology at the University of New South Wales developed a decision tree model for cardiovascular disease diagnosis. This model was constructed based on the information gain method, combined with four discretization techniques: Equal Width, Equal Frequency, ChiMerge, and Entropy. A pruning strategy was employed to remove inefficient nodes, thereby generating a concise, efficient, and accurate decision tree. Test results indicated that the initial accuracy rates of the model corresponding to the four discretization methods were 79.1%, 76.3%, 77.1%, and 78.1% respectively, demonstrating high accuracy. After optimization using the Voting ensemble method, the accuracy rates of the decision tree model increased by 3.5%, 5.7%, 1.2%, and 2.8% respectively for each method. Although this model adopts a multi-method integrated architecture, the decision tree consistently plays a core supporting role. This research outcome throughly verifies the effectiveness of decision trees in the early diagnosis and intervention of cardiovascular diseases, providing an important technical reference for global cardiovascular disease prediction [7].

3.1.3. Decision trees in early screening for Chronic Kidney Disease (CKD)

Chronic Kidney Disease (CKD) refers to a condition characterized by the gradual decline of renal function over months to years, with no apparent symptoms. For patients diagnosed only in the advanced stage of CKD, life can only be sustained through dialysis or kidney transplantation. However, if the disease is diagnosed in the early stage and treated promptly, the risk of progressing to the advanced stage can be significantly reduced. Given CKD's high latency, harmfulness, and value of early intervention, there is an urgent need to establish efficient early diagnostic methods. Currently, a large number of decision tree models have been fully applied in the diagnosis of other diseases.

For example, Hamida Ilyas and Sajid Ali developed decision tree models based on J4.8 and Random Forest [8]. The models adopted a dataset of 400 samples, covering 25 feature variables. Among them, the J4.8 algorithm follows the same construction approach as the ID3 algorithm: both calculate the Information Gain of each feature variable respectively, select the one with the maximum Information Gain as a Node, and then repeat this process until a specific condition is met or all feature variables have been selected as Nodes [8].

In contrast, Random Forest generates multiple sample subsets through Bootstrap sampling, constructs an independent decision tree for each subset, and finally outputs diagnostic results through integrated classification or regression tasks. Experimental results showed that the diagnostic

DOI: 10.54254/2755-2721/2025.LD28795

accuracy of the decision tree model based on J4.8 was higher than that of the decision tree model based on Random Forest [8].

This study fully demonstrates the advantages of decision tree models centered on ID3 and its improved algorithms (such as J4.8) in disease diagnosis. For CKD diagnosis, these models not only have practically applicable high accuracy but also possess the characteristics of convenient operation and efficient computation. They provide strong methodological support for the diagnosis of other diseases that have not yet adopted such models

3.2. Decision trees in auxiliary diagnosis and typing of mental disorders

Globally, taking depression—a prevalent mental health disorder—as an example, over 264 million people worldwide are affected by it. Without effective intervention, such a condition can develop into major depressive disorder. Studies have shown that among individuals who have previously suffered from major depressive disorder, there is still a staggering 50% chance of relapse. When a patient experiences two depressive episodes, the risk of subsequent relapse rises to 70%; if the number of episodes reaches three or more, this probability surges to nearly 90%. In existing research, auxiliary diagnostic models centered on decision tree ensemble algorithms (specifically, Random Forest) have been increasingly developed.

For instance, Sandip Roy and other scholars designed a diagnostic model that adopted six core feature variables, including therapeutic drugs, age, gender, and follow-up duration [9]. The results indicated that the accuracy rate for predicting relapse reached 78%. In multiple groups of tests, the accuracy rate ranged from 90% to 93%, with a maximum of 95%. Additionally, the researchers constructed two other models based on the K-Nearest Neighbors (KNN) algorithm and the Judgment Analysis Algorithm, respectively. After optimization, the accuracy rate of the former model ranged from 95% to 97%, while that of the latter was between 97% and 99%—both higher than the accuracy of the Random Forest model [9].

Although in the diagnosis of diseases with strong subjective characteristics such as depressive relapse, the accuracy of decision tree models is relatively lower, they still maintain the advantages of efficient computation, convenient operation, and accuracy reaching a practical level.

3.3. Decision tree applications in other healthcare scenarios

3.3.1. Decision trees in first-aid process optimization

Beyond disease diagnosis and prediction, decision trees also play a pivotal supporting role in other healthcare scenarios, such as healthcare process optimization and medical cost control.

For instance, based on historical patient data (e.g., blood pressure, routine blood test results), decision trees can be constructed to evaluate patients' physiological indicators (eliminate irrelevant or weakly correlated indicators) to optimize healthcare processes. A case in point is a research team from the Navy Medical University (Second Military Medical University), which developed a C4.5-based decision tree model to optimize the treatment process for patients with burns of different severities, using 10 pathological attributes (e.g., burn severity, blood biochemistry indicators, blood pressure) as key features [10]. Hierarchical treatment rules were fumulated accordingly:

-For patients with mild burns, treatment plans are prioritized based on blood biochemistry indicators.

-For patients with moderate burns, comprehensive judgments are required by combining blood biochemistry indicators with blood pressure and pulse.

DOI: 10.54254/2755-2721/2025.LD28795

-For patients with severe burns, the emergency treatment process is activated directly.

Due to the small training samples size of this decision tree, certain biases may exist. Nevertheless, decision trees are effective in optimizing treatment processes. Such applications are not limited to the treatment of burn patients; decision tree models can also be applied to optimize treatment plans for other diseases, ultimately achieving the goal of improving treatment outcomes.

3.3.2. Decision tree-based optimization of medical expense structure

Decision trees can integrate multi-dimensional information, such as the effectiveness of treatment items for historical patients, cost data, and the medical service capacity of hospitals, and mine the core factors affecting medical costs, thereby optimizing the cost structure.

For example, the Department of Medical Administration of Beijing Hospital studied inpatient medical expenditure composition and influencing factors for colorectal cancer (CRC) patients based on disease Diagnosis-Related Groups (DRGs) [11]. This study used 1026 CRC impatients as samples, constructing a decision tree model after data cleaning [11]. The results showed surgical type, treatment modality, and age were the key factors affecting inpatient costs, with gender having no significant impact [11]. Based on these findings, cost optimization strategies were formulated:

-For non-surgical patients, basic treatment plans should be formulated prioritizing maintenance therapy, chemotherapy, radiotherapy, and other modalities.

-For surgical patients, the age of 67 should be used as a threshold to weigh the effectiveness of different treatment modalities, and some treatment procedures should be simplified to reduce treatment costs.

Although the sample size was relatively large, with all data from Beijing hospitals, geographical sample limitations may hinder generalization to the conclusion to other regions. In addition, although the cost reduction after optimization did not meet expectations, it still verified the effectiveness of decision trees in cost control. This method can also be applied to scenarios involving the optimization of medical cost structures for other diseases.

The aforementioned experiments indicate that, relying on its characteristics of high interpretability and strong adaptability, decision trees can provide adequate technical support in multiple scenarios within the medical field, highlighting their irreplaceable application value.

4. Challenges for decision trees

4.1. Data quality challenges

Nowadays, although the volume of data is increasing rapidly— which enables decision trees to enhance the accuracy of predictions—inevitable data quality issues also emerge accordingly.

First, the relationships between data are significantly overlooked. When constructing decision tree models currently, data are often treated as “meaningless” numbers; that is, calculations are performed on each feature variable individually while the relationships between data points are ignored. For instance, when determining whether a person suffers from lung disease, the correlation between smoking status and living environment is neglected [12].

Second, data dimensionality and inter-data relationships affect classification performance. Datasets often contain various attributes that are irrelevant to the research objective, which not only increases unnecessary computations and reduces classification efficiency but also ultimately lowers the final prediction accuracy.

DOI: 10.54254/2755-2721/2025.LD28795

Third, the handling of missing data values poses a challenge. Since current decision tree algorithms are unable to directly process missing data, the absence of data can significantly compromise the accuracy of decision trees. If a large amount of data is missing, excessive sample deletion will lead to a decline in model accuracy. Even if only a small amount of data is missing—though its impact may be negligible when the total sample size is extremely large—it still requires identifying similar samples one by one to impute the missing values.

4.2. Algorithmic and technical challenges

In addition to data quality issues, decision trees also face challenges in algorithmic technology. First is overfitting. During decision tree construction, over "detailed",sample learning captures noise and irrelevant feature variables, resulting in an overly complex tree structure. Excessive depth and nodes cause biases when diagnosing new patients' information due to overfitting to the data reduces the model's generalization ability and leads to erroneous prediction results. For example, when predicting disease types, with sample sizes and no predefined rules, the decision tree may overfit to the specific sample group, deviating from general population charateristics [13].

Second is the interpretability contradiction. To improve performance, decision trees are often combined with ensemble learning (e.g., Random Forest), which increases model complexity and obscures the interpretation of specific features. This contradicts with the medical field's demand for transparent diagnostic processes, hindering their use in medical scenarios reguiring high interpretability.

Third is class imbalance. Due to the scarcity of rare disease cases, decision trees may over-trained common diseases (given their larger sample sizes). This neglects rare disease-specific features (or misclassifies them as "outliers") and potentially leading to the possibility of misdiagnosis for some individuals.

Fourth, since decision trees cannot directly process continuous random variables, these variables must be converted into discrete random variables for further analysis. This may have a certain impact on future predictions that require higher accuracy.

5. Future directions

Decision trees can mitigate the biases caused by the aforementioned issues through algorithm optimization.

5.1. Reduce model complexity

Limit decision tree complexity by regulating its construction: setting constraints such as maximum tree depth and minimum Information Gain for feature selection. This not only prevents overfitting but also significantly enhances the model's generalization ability, enabling it to adapt to more scenarios.

Optimize features: Using feature selection techniques (e.g., the filter method) to eliminate irrelevant and redundant features (e.g., excluding the "patient height" feature) when determining heart disease risk) highlighting clinically meaningful variables. Meanwhile, integrate medical domain knowledge to select representative indicators from original features. This reduces feature count, shortens tree depth, and improves both efficiency and generalization [14].

DOI: 10.54254/2755-2721/2025.LD28795

5.2. Improve interpretability of ensemble models

Given that ensemble learning improves decision tree performance but sacrifices partial interpretability, a key optimization direction is enhancing the interpretability of ensemble learning. For complex decision tree models such as Random Forest, explore interpretable methods—such as SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations)—to decompose the decision logic of complex models.

5.3. Adapt to medical data characteristics

For imbalanced data: Use oversampling minority classes (e.g., increase the number of rare disease cases). This enables the decision tree to pay more attention to minority classes during construction, improving the diagnostic accuracy for rare diseases and reducing the probability of misdiagnosis. However, such methods may also lead to reduced generalization. Therefore, special attention must be paid to the sampling ratio and techniques for rare and common diseases.

5.4. Optimize continuous variable processing

For continuous variables: To avoid information loss caused by traditional discretization, try combining decision trees with other algorithms (e.g., linear regression) to enable them to process continuous random variables, making prediction results more personalized for different individuals and significantly improving the applicability of the model in scenarios such as chronic disease screening [14].

6. Conclusion

This study reveals that decision trees, as a highly valuable machine learning algorithm, exhibit diverse and significant applications in the medical field, such as in the diagnosis and management of oral diseases, cardiovascular diseases, and other conditions. Meanwhile, their advantage—increasing interpretability and accuracy with growing data volume—aligns well with the medical field's demands for high process interpretability and diagnostic precision.

However, this study further illustrates that decision trees also face numerous challenges in medical applications. At the data level, various issues inherent in medical data (e.g., overlooked inter-data relationships) impair the efficiency and accuracy of decision tree models. At the algorithm level, decision trees are prone to overfitting and cannot directly process continuous random variables. To address these challenges, in terms of algorithm optimization, human intervention can be introduced into decision tree construction to prevent overfitting. Additionally, efforts can be made to integrate medical domain knowledge to merge feature variables, thereby improving model efficiency and other measures.

This study still has certain limitations. First, the selection of disease diagnosis cases is relatively limited, and it fails to elaborate on scenarios where decision tree models may be inapplicable for diagnosing specific diseases. Second, the lack of detailed algorithm logic demonstrations leads to overly generalized descriptions of models in some cases. In future research, this study aims to deepen the content by integrating in-depth integration of decision tree models with medical expertise, in an in-depth manner, and to provide more persuasive and applicable case studies.

DOI: 10.54254/2755-2721/2025.LD28795

References

[1] Mallappallil, M., Sabu, J., Gruessner, A., & Salifu, M. (2020). A review of big data and medical research. SAGE open medicine, 8, 2050312120934839.
[2] Abdulqader, H. A., & Abdulazeez, A. M. (2024). A review on decision tree algorithm in healthcare applications. The Indonesian Journal of Computer Science, 13(3).
[3] Slocum, M. (2012). Decision making using id3 algorithm. Insight: River Academic J, 8(2), 1-12.
[4] Di, J., & Xu, Y. (2019). Decision tree improvement algorithm and its application. International Core Journal of Engineering, 5(9), 151-158.
[5] Tanner, L., Schreiber, M., Low, J. G., Ong, A., Tolfvenstam, T., Lai, Y. L., ... & Ooi, E. E. (2008). Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PLoS neglected tropical diseases, 2(3), e196.
[6] abdulaziz Mohsen, A., Alsurori, M., Aldobai, B., & Mohsen, G. A. (2019). New approach to medical diagnosis using artificial neural network and decision tree algorithm: application to dental diseases. International Journal of Information Engineering and Electronic Business, 11(4), 52-60.
[7] Shouman, M. (2011). Using decision tree for diagnosing heart disease patients.
[8] Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M. T., Iftikhar, M., & Malik, M. H. (2021). Chronic kidney disease diagnosis using decision tree algorithms. BMC nephrology, 22(1), 273.
[9] Roy, S., Aithal, P. S., & Bose, D. R. (2021). Judging Mental Health Disorders Using Decision Tree Models. International Journal of Health Sciences and Pharmacy (IJHSP), 5(1), 11-22.
[10] Liu, W. B., Ren, D. Y., Tao, F., & Chen, G. L. (2018). Optimization of medical treatment process and rule mining for hospitalized burn patients based on decision tree. Academic Journal of Second Military Medical University, 39(12), 5-33
[11] Wu, S. W., Pan, Q., & Chen, T. (2020). Research on diagnosis-related group grouping of inpatient medical expenditure in colorectal cancer patients based on a decision tree model. World Journal of Clinical Cases, 8(12), 10.
[12] Chanmee, S., & Kesorn, K. (2020). Data quality enhancement for decision tree algorithm using knowledge-based model. Current Applied Science and Technology, 259-277.
[13] Gulati, P., Sharma, A., & Gupta, M. (2016). Theoretical study of decision tree algorithms to identify pivotal factors for performance improvement: A review. Int. J. Comput. Appl, 141(14), 19-25.
[14] Abdulqader, H. A., & Abdulazeez, A. M. (2024). A review on decision tree algorithm in healthcare applications. The Indonesian Journal of Computer Science, 13(3).