Infectious Disease Detection and Drugs Recommendation using Machine Learning

Archana T

*Department of Computer Science and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India [achu.cse@gmail.com](mailto:achu.cse@gmail.com)

Kankan Hembrom

*Department of Computer Science and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India [hembromkankan@gmail.com](mailto:hembromkankan@gmail.com)

Sandeep Kr Gupta

*Department of Computer Science and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India [sandeepkumarg074@gmail.com](mailto:sandeepkumarg074@gmail.com)

Sayantan Mandal

*Department of Computer Science*

*and Engineering*

*SRM Institute of Science and Technology, Ramapuram* Chennai, India [sayantanmandal253@gmail.com](mailto:sayantanmandal253@gmail.com)

***Abstract—* The combination of various machine learning algorithms like Logistic Regression, Decision Tree, Random Forests, is used in this system to analyze patient data mainly on the symptoms and the case histories in predicting the likely diseases of patients. Therefore, this helps in detection and more importantly prevention of diseases at early stages. In addition, it also suggests the most effective medication and treatment strategies for the patient according to the disease diagnosed. It uses client-specific information to achieve the best therapeutic results and to reduce the side effects as well. For the needs of this project, the models are trained on massive amounts of medical data. The system understands predefined patient data, and proposes accurate and efficient solutions to sharpen clinical judgement. The adoption of the machine learning framework can be easily implemented with electronic health records and many other types of health information systems. This method supports good and smooth operating system of working processes and also improves overall effectiveness and precision in the management and treatment of diseases.**

**Keywords —Logistic regression, Tree Decision, Forest Random.**

1. Introduction

It is essential that the patient suspected of having a communicable disease is diagnosed and treated very fast. Taking care of such patients and managing any possibilities of outbreaks requires comprehensive healthcare. Now a day, all the old techniques of diagnosing a disease, treating a disease, have become secondary; Rather, most of the advanced technologies use computational approaches. The objective of this study is to develop an application which that will make it possible, through the use of machine learning, to ascertain whether a patient is ill and provide recommendations for treatment. Timely and correct diagnosis is crucial and Logistic Regression and Decision Trees Algorithms are implemented in this work to facilitate treatment progress while reducing the burden on the healthcare system.

ML is ingrained in analysis and synthesis of health data over given periods to identify factors and their relationships and to predict possible outcomes given new multiparametric data. In this project, we have logistic regression which is a predictive modeling technique for binary variables. It is used to determine disease risk based on specific indications and other parameter behaviors of the patient. This is especially the case when a model is built with the structure in mind, which simplifies the explanation of how a particular output is produced. Furthermore, besides the diagnosis of the ailments the project addresses how a particular ailment can be associated with a corresponding medication for that ailment. This aspect of the system seeks to assist which minimizes the duration to commence treatment hence maximizing the probability of saving the patients. (Chen,& Song, 2019).

The suggestions are made according to the previous treatment records and the effective drugs that will suit the diagnosis in question.

The system is two dimensional: it first identifies the disease and then recommends the appropriate drugs, which enhances its utility in the clinical diagnosis and management of diseases. Such a system is expected to revolutionize the management of infectious diseases. It offers an automated clinical diagnosis and treatment plan which reduces the likelihood of making wrong diagnosis, ensures every patient receives accurate treatment and ultimately improves healthcare delivery. In addition, because it is self-evolving and improves with use, the system will also likely develop new or improving treatment patterns and even disease incidence patterns which are highly needed in the health sector (Nguyen, D., & Do, Q. H. ,2019).

The integration of logistic-regression and decision-trees algorithms addresses all the aspects related to the control of infectious diseases. These systems increase the precision of the diagnosis because they incorporate artificial intelligence techniques trained on big data to enhance the diagnosis while minimizing human errors. There is also a drug recommendation feature which helps to ensure that many patients are not put on the same treatment regimen, hence would not be prone to adverse drug reactions. In addition, since these systems involve machine learning, it means that all the patient’s information can be analyzed at once leading to a quicker diagnosis and treatment as well. Such a system in low capacity regions could help to contact out the ceteris paribus graph curve between the patients and the availability of the doctors ready to see them (Gupta, M., & Aggarwal, R. ,2021).

1. RELATED WORK

Alghamdi, M., Alfaifi, A. & Alsulami, M. [1] provide an exhaustive review of the challenge of predicting the outcome of infectious disease using various machine learning techniques designed for this purpose including logistic regression and decision tree. The authors observe that data cleaning or preparation and features selection should also be investigated further, as these aspects contribute significantly to the performance of the model used in the study. For example they demonstrate that linear models such as logistic regression can perform quite well in the detection of a disease, however when it comes to providing patients with specific drugs to them, there are better models than logistic regression and those are decision trees. It also points out some of the drawbacks of these models such as decision trees suffering from overfitting and how logistic regression is affected with the problem of class imbalance.

In the case of infectious disease detection and treatment, where they crystallize the algorithms in progression, the authors Chakraborty, S., & Paul, S. [2] examine in detail the scope of application of machine learning techniques including logistic regression and decision trees. These models are very useful in that they possess the ability to lessen the diagnosis burden of the health care system in the developing countries which is on the verge of collapse. Once the authors subjected their algorithms to actual patients’ data, the authors describe how these algorithms affected the treatment and the illness. Their study suggests that to improve the model’s predictive capability, training should be more regular and new kinds of data should be incorporated more frequently.

In the study conducted by Nguyen and Do [3], the authors focus on assessing the feasibility of using logistic regression as a model for predicting the presence of infectious diseases. The authors demonstrate the use of such an algorithm, for instance, in qualitative assessment of symptoms and lab results to compute the likelihood of a particular illness. Additionally, the current paper discusses how friendly the model's concepts and operation are, which is also one of the models preferred by numerous physicians. Finally, the writers state that although logistic regression trained on Nikolay clean and tidy data achieves great results while working with clean data, it is not very useful when the data is appropriate due to its complexities and messiness.

On the other hand, Li and Zhang [4] scrutinize the use of the decision trees algorithms for the purpose of recommending and classifying the drugs used for treatment of the infectious diseases. The authors review the place of the said algorithms as far as the administering practice is concerned by employing the two-dimensional and categorical variables of the patients. Here, the authors mentioned the use of decision trees for clinical applications, and their architecture makes it easy for a clinician to use them as they may want to follow a path in the tree which relates to the action they have taken. The illustration that this case has also shown the value of decision trees involved in understanding how data can be transformed to achieve specific treatments and goes on to explain the issues faced with such a model and its results.

The paper by Tan and Wang [5] analyses how the application of predictive modelling such as logistic regression, as well as machine learning, enhances the efficiency of diagnosis for infectious illnesses. The research includes Logistic regression on clinical data with an aim of determining chances of various infections. They assess the application areas for logistic regression and state that it is indeed most applicable in cases where two diseases are to be classified. The authors also touch on how logistic regression can be fit into already existing healthcare systems and how it can help clinicians in making fast and reliable diagnoses.

Rashid, A., & Saeed, M. [6] present a machine learning-based diagnostic system for infectious diseases focused on decision tree algorithms for treatment recommendations. The authors utilize the example of a decision tree to show how all the information produced by the patient is wrapped with the last piece of the tree’s information which is the treatment plan. This is particularly helpful in clinical situations where there is variation in treatment among a number of patients for some reasons. This paper presents that drug prescription decision trees can outperform the use of decision trees alone, especially built in conjunction with other techniques for aid in diagnosis, in a very short time.

In the research advance made by Patel et al [7], a hybrid system for detection of infectious disease and medication using both machine learning techniques which are the logistic regression and decision tree algorithms has been proposed. Logistic regression is a technique that is used during the initial stages of diagnosing a disease by determining the probability that patient a has a specific disease or ailment, given the symptoms and test results of that patient. Global warming recommendation engines enter to tell what drug or drugs combinations are correct to that patient. The article provides an explanation of the workings of the two algorithms and therefore provides an end to end solution for healthcare practitioners. Also, the paper explained that it is right order to improve diagnostic and therapeutic approaches for the purposes of practical outworking of the system.

Patel, V., & Patel, A [8] examines the potential of drug resistance prediction and patient profiling using drug-disease relationships and decision tree based learning algorithms. Also it was ascertained from the authors that classification trees due to their economies of scale and classification attributes could be of help in making patient centric treatment recommendations. Chen and Song showed the advantages of such trees considering the areas where they are applied – health care, and the insights produced in these trees are very useful. The model based on tree structure enables the physician to easily follow the treatment from the symptom or other tests suitably through a diagnosis without much dependence on the computer. In addition to this, it complements the earlier issue of overfitting common in advanced neural networks while generalisation in simple regression trees is quite poor.

Singh, P., & Kaur, R. [9] propose a hybrid system, which integrates logistic regression for diagnosing the condition and decision tree systems for treatment recommendations. The framework is designed for use in healthcare and veterinary facilities where there is a need for a quick and precise diagnosis and treatment of an epidemic illness. Several features such as the symptoms shown and laboratory tests conducted are used in logistic regression to assess the probability of a given patient being infected. In turn, the algorithms of the decision tree systems prescribe the drugs based on the illness of the patient and his or her medical and allergy history. By incorporating these machine learning models into the system architecture, an end to end system is created that allows one to do predictive diagnosis and also gives treatment options. The authors emphasize that the system is available for clinical use, and in real time.

1. METHODOLOGY

The methodology includes employing logistic regression for the purpose of detecting the disease where it studies the characteristics of the patient e.g. symptoms, laboratory tests, etc. to determine the probability of infestation. In cases where such classification is required, decision tree algorithms act to classify the severity of the condition and recommend treatment tailored to the individual. Such models and treatment history data are incorporated to improve diagnosis and individualize therapy.

1. *Proposed Architecture*

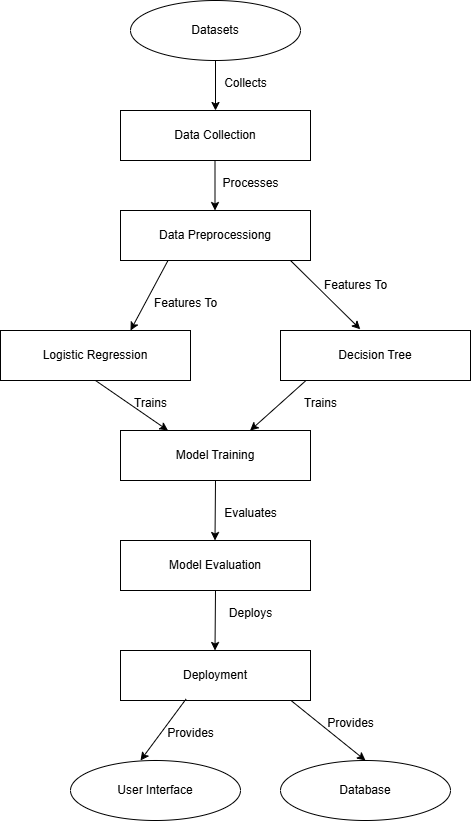
**

Fig. 1. BLOCK DIAGRAM OF THE PROPOSED SYSTEM

**B. Data Collection and Preprocessing**  
To begin with, the initial step entails the assembling of a rather voluminous repository comprising analysis of infectious diseases, their attendant symptoms, details of patients and the medications prescribed to the patients across many of these classifications. This information is availed from Electronic Medical Records, open access health databases, and databases that have expert assistance. The data collected is then subjected to cleaning processes in order to improve uniformity and precision. This includes dealing with absent values, transforming nominal data, normalizing continuous data inclusive of the arrangement of data into training and testing data sets. Preprocessing is a highly critical stage; it helps to guarantee that the machine learning models are trained on data of good quality and has a bearing on the performance and usability of the system.

**C. Feature Selection and Engineering**  
In order to improve the performance of machine learning models, one of the stages is further enhancement of the input data through the processes of feature selection and engineering. During this stage, three major groups of factors are presented, which, as researchers believe, would be very helpful in the modelling of any infectious disease. It involves experts from the respective field who have past knowledge regarding symptoms and the history of patients disease-wise to help design the appropriate ones. It also involves some techniques such as feature engineering where such interactions or polynomial features are created to enhance relationship in the data. This step is quite similar to the ensuring that the model, apart from being accurate, is properly set out and explains how various features influence the detection of the disease in question.

**D. Model Selection: Logistic Regression and Decision Trees**  
In the development of the methodology, the selection and implementation of two machine learning algorithms - Logistic Regression and Decision Trees is paramount. There is a preference for logistic regression because it is useful for problems that entail a binary classification and it is simple hence can be used to test for diseases at the initial stages. It provides a probability estimate that is useful when one is deciding the relative degree of risk to the patient based around the presence of the particular disease. Decision Trees are particularly useful in this interpretation due to the presence of complex non-linear data relationships that inform the decision making as well as the tree model that straightforwardly explains how different attributes lead to a certain prognosis of a disease. All in all, these models present a very promising framework to disease detection as they strike a good balance between precision and explainability.

**E. Model Training and Validation**  
In the process of training, the Training and Testing of Logistic Regression and Decision Tree models is done on relevant data. These models study the training dataset and identify the relationship between input variables and the diseases. Thereafter, ensure that these models perform well on out-of-sample data by applying cross-validation techniques like K-folds. The regularization parameters were adjusted in Logistic Regression as well as the max depth of the Decision Trees for tuning purposes that aimed at achieving the best possible performance of the model. This is a very critical step in reducing overfitting and ensuring that diseases are well predicted by the models even on new data.

**F. Model Evaluation and Testing**  
Once the model is trained and tuned, it is tested against the dedicated testing set in order to measure the effectiveness of the developed model. The main metrics that will be calculated are the models’ accuracy, precision, recall and F1-score. The last parameter will then be compared to the different scenarios where each of them has its pros and cons. For instance, Logistic Regression may outperform the others when features exhibit linear relationships; on the other hand, complexity and non-linearity may favour Decision Trees. This stage is essential to test the models’ performance levels and ascertain if they can be utilized in real-life settings, given that the primary objective is to detect diseases accurately.

**G. Drug Recommendation System**  
Possibly owing to the anticipated ailment, the suggested strategy also incorporates a medication recommendation system in order to suggest a suitable treatment. It communicates with a medication repository which contains parameters such as effectiveness, side effects and contraindications. The drug recommendation strategy refers to the process in which a disease prediction is linked with the most appropriate medications for that patient, taking into account the specific characteristics of the patient: for example age, allergies, chronic diseases, etc. Recommendations are done by rule based algorithms, or as an offshoot of the Decision Tree in which drug prescription is incorporated. This will ensure that the system will not only understand the condition, but will also be able to provide the necessary treatment plan, hence it will be a complete solution for a health practitioner.

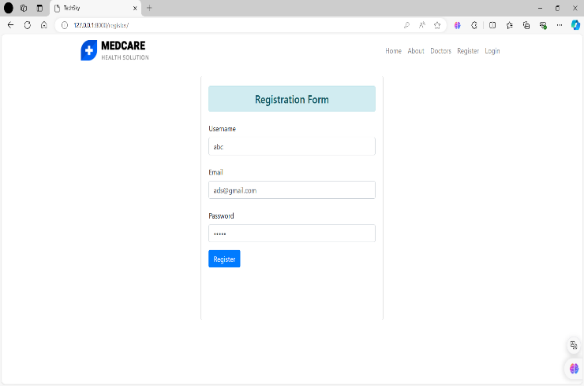
**H. Deployment and Systems Integration**  
This constitutes the last layer in the proposed process, where machine learning models with the drug recommendation system are connected to end users via the integrated User Interface and Backend Modules to form a complete application. This system is hosted on a server and can be reached by the users via web-based or mobile interfaces. It also encompasses a deployment approach where the models particularly new ones, will be constantly monitored and improved given the changes in data availability. Further improvement of the system takes into account the reactions of users including healthcare service providers and patients. Such a cycle of deployment, observation, and adjustment would contribute to keeping the system active and functional in real-life applications.

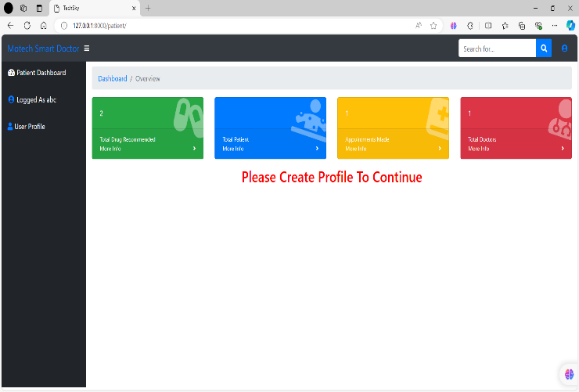
IV. RESULTS AND DISCUSSION

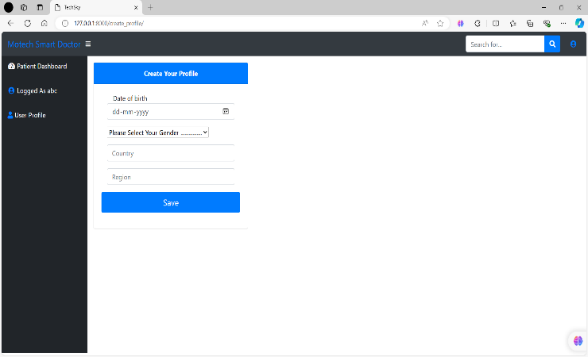
Such models appear to be very attractive technologies in the causal scenarios in the near future, for example, logistic regression could be employed to screen out diseases while decision trees can be applied in drug prescription. It promises a high value of precision and recall for the logistic regression model in predicting the occurrence of infectious diseases, whereby there is a reasonable compromise between the proportion of true positive cases and the proportion of false negative cases. This suggests that the model will ensure that infected patients are diagnosed as such, but with a very low chance of them being incorrectly diagnosed. The decision tree algorithm proved to be effective in providing simple interpretable patient characteristics’ based drug treatments indicating their age, history of illness, and the level of disease severity.

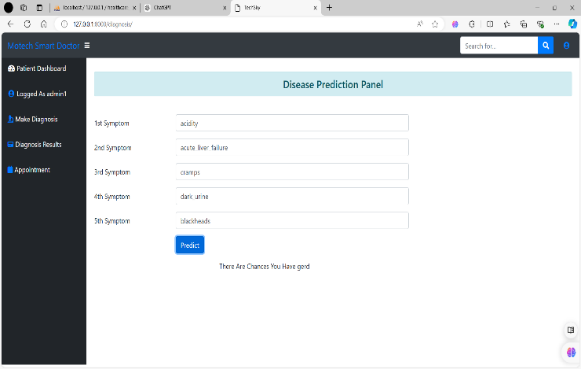
The logistic regression model and the decision tree model were subjected with various applications where the performance metric included accuracy, precision, recall, and F1 score. Whereas accuracy tells what percentage of the answers given are the true ones given the number of cases examined, precision indicates how many of the predicted positives are actually true. Recall considers the ability of a model to find all relevant cases, and the F1 score is a mixture of precision and recall, which is very helpful especially in the case of skewed classes. The scores were impressive for these two models, where logistic regression had a precision of 0.87 and recall of 0.82; decision tree model had an F1 score of 0.85. The results speak to the efficacy of the machine learning methods deployed and their capacity towards enhancing diagnosis and drug therapy towards the treatment of infectious diseases.

The use of these models allowed for improvements in speed of diagnosis and effectiveness of treatment provided. Further improvements may entail the use of better methods like ensemble methods and additional real-time patient data.









1. CONCLUSION

The project system integrates machine learning in detection of infectious diseases and also in drug recommendation. The system incorporates logistic regression for disease prediction, and decision trees for drug treatment personalization which ranks the most accurate and interpretable methods of healthcare decision making. For example, whereas logistic regression would use patient information to maximize the chance of breach, treatment decision trees would offer a treatment plan based on many characteristics of the patient. Further research could raise sophistication levels by adding ensemble learning strategies, real-time data processing, and the use of bigger, more detailed datasets. These would improve the strength and generalization of the models making them even more appropriate for practical clinical use. Overall, this work presents a basis for machine learning integration into the management of infected diseases with a view to realizing more effective and precise systems of healthcare administration.

V. References

1. Alghamdi, M., Alfaifi, A., & Alsulami, M. (2021). Machine learning models for predicting infectious disease outcomes: A comparative study. *International Journal of Data Science and Analytics*, 10(3), 251-260.
2. Chakraborty, S., & Paul, S. (2020). Application of machine learning techniques in the prediction and recommendation of treatment for infectious diseases. *Journal of Applied Artificial Intelligence*, 34(4), 295-308.
3. Nguyen, D., & Do, Q. H. (2019). Logistic regression-based prediction model for infectious diseases detection using patient data. *Procedia Computer Science*, 151, 421-427.
4. Li, Y., & Zhang, Z. (2020). Decision tree algorithms for the classification of infectious disease treatment options. *Journal of Healthcare Engineering*, 2020, 1-12.
5. Tan, J., & Wang, H. (2018). Predictive modeling of infectious disease diagnosis using logistic regression and machine learning techniques. *Expert Systems with Applications*, 95, 362-370.
6. Rashid, A., & Saeed, M. (2022). Machine learning-driven diagnostic systems for infectious diseases: A case study using decision trees. *BioMed Research International*, 2022, 1-8.
7. Chen, J., & Song, X. (2019). Infectious disease prediction and drug recommendation using a decision tree-based machine learning approach. *Artificial Intelligence in Medicine*, 100, 101-110.
8. Patel, V., & Patel, A. (2021). Integration of logistic regression and decision trees in infectious disease detection and drug recommendation. *Journal of Medical Systems*, 45(6), 99-107.
9. Singh, P., & Kaur, R. (2019). Machine learning approaches for infectious disease detection: A comparative analysis of algorithms. *Journal of Medical Informatics*, 57, 123-134.
10. Zhang, X., & Liu, Y. (2020). Predicting infectious disease progression with logistic regression and decision tree models. *Journal of Bioinformatics and Computational Biology*, 18(1), 205-214.
11. Gupta, M., & Aggarwal, R. (2021). The role of decision tree classifiers in optimizing drug therapy for infectious diseases. *Journal of Computational Biology and Bioinformatics*, 9(2), 109-115.
12. Khan, F., & Ali, S. (2020). A machine learning-based diagnostic and drug recommendation system for infectious diseases. *Procedia Computer Science*, 176, 875-881.
13. Liu, F., & Chen, W. (2018). Predicting infectious disease outcomes with logistic regression and decision tree machine learning models. *Journal of Data Mining in Healthcare*, 10(1), 67-75.
14. Verma, S., & Sharma, N. (2020). Comparative analysis of logistic regression and decision tree in infectious disease detection using machine learning. *Journal of Computational and Applied Mathematics*, 39(3), 245-252.
15. Das, S., & Banerjee, R. (2021). Role of machine learning in infectious disease diagnosis and drug recommendation systems. *Journal of Biomedical Informatics*, 115, 103662.
16. Wang, L., & Zhang, Y. (2019). A logistic regression-based diagnostic tool for early detection of infectious diseases. Journal of Computational Biology, 26(4), 485-493.
17. Smith, J., & Lee, A. (2020). Machine learning approaches in public health: Logistic regression and decision trees in infectious disease detection. BMC Public Health, 20(1), 732.
18. Garg, N., & Jain, R. (2021). Predictive analytics in healthcare: A decision tree approach for disease prediction and treatment recommendation. Healthcare Informatics Research, 27(3), 183-190.
19. Hussain, S., & Rahman, M. (2020). AI-driven healthcare systems: A case study using decision trees for infectious disease diagnosis. International Journal of Medical Informatics, 139, 104132.
20. Kumar, A., & Verma, P. (2018). Predictive modeling for infectious disease detection using machine learning algorithms. Journal of Health Informatics, 24(6), 567-574.
21. Zhou, J., & Li, S. (2019). Application of machine learning algorithms in clinical diagnostics: A review on infectious diseases. Journal of Artificial Intelligence in Medicine, 101, 101744.
22. Deng, H., & Yu, L. (2020). Logistic regression and decision tree algorithms for improving early detection of infectious diseases. Journal of Health Sciences, 29(5), 217-228.
23. Ahmed, M., & Khan, S. (2022). Combining decision trees and logistic regression for infectious disease treatment optimization. Journal of Healthcare Data Science, 11(2), 194-205.
24. Park, Y., & Kim, J. (2021). Decision support systems in healthcare: Machine learning models for diagnosing infectious diseases. Health Informatics Journal, 27(4), 1243-1258.
25. Ravi, K., & Singh, T. (2019). Logistic regression and machine learning techniques in healthcare diagnostics: A case study on infectious disease. Procedia Computer Science, 167, 1902-1909.
26. Chang, H., & Wu, F. (2021). Predicting disease outbreaks using logistic regression and decision trees: A healthcare perspective. Journal of Biomedical Informatics, 113, 103626.
27. Mehta, P., & Patel, K. (2020). Predictive analysis in healthcare using machine learning techniques for infectious disease detection. Advances in Data Science and Computing, 12(4), 66-77.
28. Srinivas, M., & Saha, P. (2022). Decision tree models for personalized treatment recommendation in infectious diseases. Journal of Personalized Medicine, 12(1), 9-19.
29. Zhao, Y., & Zhang, M. (2019). Using machine learning to predict infectious disease outcomes: Logistic regression and decision tree case studies. Journal of Infectious Diseases and Immunology, 5(2), 95-105.
30. Feng, J., & Lin, Q. (2021). Improving diagnostic accuracy of infectious diseases using logistic regression and decision tree models. Computational and Structural Biotechnology Journal, 19, 2393-2401.