Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

Predictive Data Mining for Disease Diagnosis-Decision Tree Approach

Dr Sharad Mathur
Computer Science
Lachoo Memorial College Sc & Tech
Jodhpur, India

Dr Ashish Rai
Computer Science
Lachoo Memorial College Sc & Tech
Jodhpur, India

Dr Deepak Mathur
Computer Science
Lachoo Memorial College Sc & Tech
Jodhpur, India

***Abstract*— A disease prediction system is important in modern healthcare for many reasons. The correct prediction of disease is most difficult work. These systems use machine learning data analysis and artificial intelligence to predict the chances of diseases, enabling proactive healthcare management. Disease prediction can help in early detection and prevention of disease, personalized treatment plans, decrease treatment cost etc. Decision Tree is known for Supervised learning, it is applied to solve classification and Regression related problems, but generally it is preferred to solving Classification problems. Decision tree is a tree structured classifier, it's internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Decision tree is popular machine learning technique that can be used to predict disease outcomes, diagnose diseases, and make prognoses. They have been used successfully to predict the risk of disease including diabetes, kidney disease, cancer, obstructive sleep apnea, heart disease and many more. This paper reviewed research articles that mostly used decision tree to predict cancer, diabetes, and heart disease.**

***Keywords*— Healthcare, Disease Prediction, Data Mining, Support Vector Machine**

I. INTRODUCTION

Data mining has significant role in disease prediction. It uses various techniques to analyze vast amounts of healthcare data. Data mining, defined as the process of extracting meaningful patterns from large datasets, includes methods such as classification, clustering, sequential pattern analysis, and association rules. These techniques are widely applied in healthcare to predict diseases, reducing the need for extensive testing by identifying relevant patterns in historical data [1]. This approach can save time, improve performance, and enhance patient care by allowing for early disease detection.

One of the primary data mining techniques highlighted in healthcare is the Decision Tree, a machine learning algorithm used for both classification and regression. One kind of algorithm that creates a tree-shaped model by classifying data is called a decision tree. It is a graphical representation of the data that shows the various options and the outcomes that could arise from each one. Decision trees are a popular model because they make it much easier to understand the many possibilities.

This review highlights Decision tree's application in cancer prediction. India, referred as the "cancer capital of the world," has witnessed a rise in cancer cases, with breast, cervical, and ovarian cancers being prevalent among women, and prostate, lung, and oral cancers common among men. Chen et al. proposed a method utilizing Gradient Boosting Decision Tree (GBDT) for identifying susceptible genes related to gastric cancer. Their findings indicated that GBDT outperformed other algorithms, underscoring the efficacy of decision trees in genetic prognosis research [2].

Decision Tree has also been applied to diabetes prediction. Given the high incidence of type 2 diabetes in India, where more than 77 million adults suffer from the condition, The decision trees have been effectively used in prognosis by assessing the risk factors and predicting the progression of chronic diseases such as kidney disease and diabetes [3].

Heart disease prediction is another area where Decision Tree has been widely applied. Heart disease is one of the leading causes of death globally, and factors such as age, cholesterol levels, and blood pressure are considered in prediction models [4]. Swathi Priyadarshini et al. integrated clustering techniques with classification algorithms like Naïve Bayes and Decision Trees to develop a heart stroke prediction model, demonstrating the promise of combining methodologies to enhance health prognosis models [5].

II. DATA MINING FOR DISEASE PREDICTION

The process of searching through enormous volumes of data for relevant information is known as data mining. Sequential patterns, association rules, classification, clustering and prediction are a few of the most significant and widely used data mining techniques. Numerous applications make use of data mining techniques. Data mining is crucial to the health care industry's ability to forecast disease [6]. A patient should be obliged to undergo a lot of tests in order to discover an illness. But fewer tests should be conducted if data mining techniques are used. Performance and time are important points that are influenced by this shortened test.

Following figure (fig.1) shows stepwise disease prediction system based on Decision tree machine learning algorithms. It consists of various steps. Firstly, historical data set is collected for a particular disease it is necessary to train the model [7]. Next step is for data cleaning. Then we need to select various features required by the Decision Tree model. There after training and testing process works and then Decision Tree Model generate prediction for given patient data.

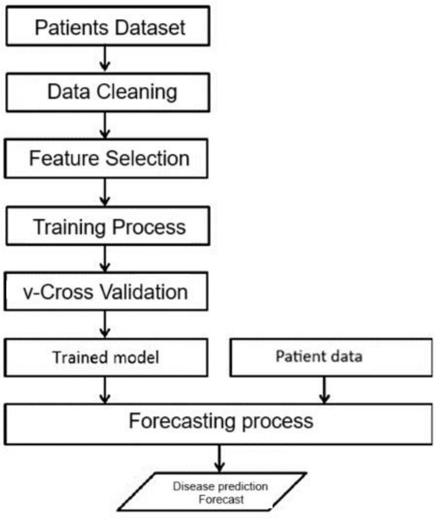


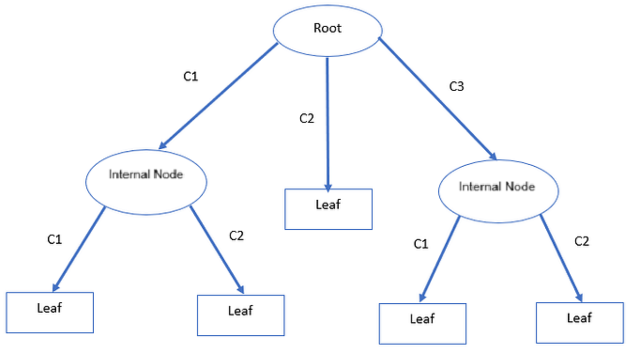
Fig. 1. Block diagram of Disease Prediction System based on Decision Tree

III. DECISION TREE

A decision tree models the possible choices in a problem-solving process as a branching structure that illustrates how various factors influence one another [8]. It begins at a single top-level “root” node, representing the full dataset, and then splits into branches that lead to further decision points or outcomes. Each internal node corresponds to a test on one of the input features, directing the flow down different branches based on the feature’s value. The process continues until it reaches a leaf node, which denotes a final prediction or classification. Depending on the type of outcome you need to predict, decision trees fall into two main categories:

Classification trees handle discrete, categorical targets (for example, determining whether an email is “spam” or “not spam” by examining its attributes) [9].

Regression trees are used when the goal is to predict a continuous value (for instance, estimating the market price of an apartment from its location, size, amenities, and other characteristics).



Among the numerous decision tree algorithms in the literature for classification problems, ID3, C4.5, Classification and Regression Trees (CART), and Chi-Square Automatic Interaction Detection (CHAID) are among the most frequently referenced. Researchers commonly assert that the CART algorithm demonstrates strong reliability compared to other classification techniques [10]. Like other decision tree algorithms, CART extracts decision rules from features to build a predictive model for target values [11].

IV. CANCER DISEASE PREDICTION

A latest report has reported a deep decline in India's overall health. According to a report released by an international healthcare organization, India has earned the alarming title of "the cancer capital of the world," due to the rising number of cancer cases and other non-communicable diseases.

Specifically, it is anticipated for India that the number of cancer cases will increase at a rate that will outpace norms worldwide, rising from 1.46 million in 2022 to 1.57 million by 2025 [12]. Ovarian cancer, breast cancer and cervical cancer are the 3 most common cancers between women. The commonly found cancers for men are oral, prostate and lung. According to a study published in the Lancet Oncology, India defies the global trend where men report a 25% higher incidence of cancer than women: more women there are receiving cancer diagnoses.

A. S. Nath [13] utilized approximately 25 attributes associated with lung cancer to predict the disease using classification algorithms such as "Naive Bayes," "Bayesian Network," and "J48." The analysis revealed that the "Naive Bayes" algorithm was the fastest in terms of model-building time. The study also suggested that the lung cancer prediction system could be improved and expanded by incorporating additional data mining approaches, including "time series," "clustering," and "association rules."

Patients diagnosed at AJCC stage III or IV, as well as older individuals, tend to have less favorable outcomes. By employing a decision tree approach, patients can be stratified into three main risk categories and further divided into twelve more specific subgroups. With the help of PROBAST to assess the model’s development confirmed that the tree shows high applicability and a low risk of partiality. Nonetheless, these findings require confirmation through external validation [14].

Our study highlights the effectiveness of the "Decision Tree" algorithm, achieving an impressive accuracy of 99.67% in predicting lung cancer severity, and the customized "VGG16" CNN, which attained a 92.53% accuracy in identifying the specific carcinogen type present in the lungs [15]. These findings are clinically significant, as early and precise detection plays a critical role in improving patient outcomes and reducing lung cancer-related mortality. The proposed system employs a two-step verification process for cancer detection. Initially, a question-based analysis assesses the severity of lung cancer. If the risk is determined to be medium or high, the system advises undergoing a CT scan. The investigation of the CT scan image affirms the existence of cancer cells and classifies the type of lung cancer to help in improved treatments. This dual-step approach enhances diagnostic accuracy compared to currently available methods [16].

V. DIABETES DISEASE PREDICTION

Insulin helps the body turn food into the energy required for daily living, including sugar, carbohydrates, and other foods. Body requires insulin to consume sugar but if the body is not able to use or generate enough insulin the extra sugar will be eliminated through urination [17]. The cause of diabetes is not known, yet high weight and inactivity seem to be major contributing factors.

According to an estimate 77 million Indians over the age of 18 have type 2 diabetes, and another 25 million are prediabetes, meaning they have a higher chance of getting the disease in the near future [18].Over half of individuals with diabetes remain unaware of their condition, which can lead to serious health complications if not detected and managed in time. Adults with diabetes are two to three times more likely to experience heart attacks or strokes. Additionally, reduced blood circulation combined with nerve damage in the feet significantly increases the likelihood of infections, foot ulcers, and, in severe cases, the need for amputation. A significant contributor to blindness is diabetic retinopathy, which is brought on by cumulative long-term damage to the retina's tiny blood vessels. In order to identify instances of diabetes and pre-diabetes in the American population, Wei Yu et al. evaluated two classification schemes [19].

Smith et al. utilized CART to analyze the Pima Indian Diabetes Dataset (PIDD), achieving an accuracy of 75%. The study emphasized the importance of feature selection, noting that glucose and BMI were the most significant predictors [20]. Pham et al. makes a prediction on whether a new patient will test positive for diabetes [21]. In order to establish how to best manage the overfitting and overgeneralization characteristics of classification on this dataset (the Pima Indian diabetes data set), this research explored a novel approach known as the Homogeneity-Based Algorithm, or HBA. To improve the classification accuracy of classification techniques like Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and Decision Trees (DTs), the HBA is employed in tandem with them [22]. According to some experimental results, the suggested strategy works noticeably better than the ones that are now in use. The experiment's author came to the conclusion that it is critical for the data mining community as a whole as well as for correctly predicting diabetes.

According to Viloriaa et al. the most important factors for the diagnosis of diabetes mellitus (DM) are age, body mass index (BMI) and blood glucose concentration [23]. Diagnosis of diabetes mellitus by a doctor is difficult, due to several causes are involved in the disease. There is high chance of human error in the diagnosis. A blood test may not give sufficient detail to get a correct diagnosis of the disease. In 2021, Johnson et al. compared Decision Tree models with other machine learning algorithms for diabetes prediction and found that Decision Trees provided comparable accuracy while offering better interpretability for healthcare professionals [24]. A study by Kumar and Singh explored hybrid Decision Tree models combined with ensemble techniques like Random Forest and Gradient Boosting. They reported improved prediction accuracy, with Random Forest achieving an accuracy of 82% on PIDD [25].

VI. HEART DISEASE PREDICTION

Heart disease has emerged as the leading cause of death globally. Heart disease can be diagnosed by a variety of symptoms, including fatigue, headaches, angina, and swelling legs [26]. Heart disease is also largely caused by lifestyle factors, including eating habits, inactivity, and the existence of other illnesses like high blood pressure. There are not enough skilled medical professionals in the world, which is concerning for the healthcare industry. In the area of disease prediction in particular, machine learning has proven to be a potent tool [27]. Through the utilization of complex algorithms and vast quantities of patient data, machine learning models can be effectively taught to forecast an individual's risk of developing certain diseases.

Tiwari et al. applied the CART algorithm to the Cleveland Heart Disease dataset, achieving an accuracy of 83%. They identified cholesterol levels, maximum heart rate, and chest pain type as significant predictors [28].

Sharma and Gupta compared Decision Trees with Logistic Regression and Support Vector Machines (SVM) for predicting cardiovascular risk. Their findings showed that Decision Trees achieved similar accuracy while being more interpretable, with an F1-score of 0.79 [29]. Alqahtani et al. integrated Decision Trees with ensemble methods such as Random Forest and Gradient Boosting, improving accuracy to 89% on the Framingham Heart Study dataset. This research showed the robustness of ensemble approaches [30]. Patel et al. developed a hybrid Decision Tree-based model combined with genetic algorithms to optimize feature selection, reporting a prediction accuracy of 87% [31]. Chen et al., Chen et al. examined Decision Trees enhanced with explainable AI techniques for cardiovascular disease prediction. They focused on patient-centric transparency and achieved an accuracy of 85% using the UCI Heart Disease dataset [32]. Johnson et al., focused on the use of advanced Decision Tree algorithms, such as CHAID (Chi-Square Automatic Interaction Detector), for early detection of heart disease. The model achieved an accuracy of 88% and emphasized the importance of incorporating family history and lifestyle factors into the analysis [33]. Ahmed et al. used a Decision Tree classifier on a data of set of 500 patients. The model identified significant predictors, including body mass index (BMI), blood pressure, and physical activity levels. With an accuracy of 82%, the study emphasized the importance of regional data in model training [34]. The study by Asabe et al. highlights the use of decision tree algorithms in predicting heart attack risks, showing how these models can effectively utilize health indicators to enhance diagnostic accuracy. Their findings demonstrate the decision tree's capability to not only match but exceed existing methods by incorporating real-time data analysis, which can significantly benefit ongoing patient health monitoring [35]. The Boukhatem et al. demonstrated 4 classification techniques: Multilayer Perceptron, Support Vector Machine, Random Forest, and Naïve Bayes [36]. Steps for feature selection and data pretreatment were completed prior to model construction. Accuracy, precision, recall, and F1-score were used to calculate performance of the different models.

VII. CONCLUSION

Disease prediction systems play a crucial role in enhancing patient care by identifying individuals at higher risk of developing certain conditions early on. This allows healthcare providers to create personalized treatment plans, predict future care needs, and develop long-term strategies, thus improving patient outcomes. These systems are particularly useful in reducing hospital readmission rates through targeted interventions for patients at risk of complications post-discharge. In this research, we explored the effectiveness of predictive data mining techniques, specifically the Decision Tree approach, in the field of disease diagnosis. The findings highlight that decision trees offer a robust and interpretable model for medical professionals to analyse patient data and predict potential health conditions. By systematically breaking down complex medical datasets into a structured hierarchy of decisions, this approach enables accurate classification of diseases based on various health parameters. Furthermore, its visual and rule-based nature makes it particularly suitable for clinical applications, offering transparency and ease of understanding. While the results are promising, future work should focus on integrating decision trees with other machine learning models to enhance diagnostic accuracy and validating the models on larger, more diverse datasets. Overall, this study reinforces the potential of decision tree-based predictive systems in supporting timely and reliable medical diagnoses.

REFERENCES

[1] K. Aftarczuk, “Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems,” Blekinge, 2007.

[2] Q. Chen, J. Zhang, B. Bao, F. Zhang, and J. Zhou, “Large-Scale Gastric Cancer Susceptibility Gene Identification Based on Gradient Boosting Decision Tree,” Front Mol Biosci, vol. 8, 2022, doi: 10.3389/fmolb.2021.815243.

[3] S. K. Opoku, A. Y. Obeng, and M. O. Ansong, “Decision Tree Models for Predicting the Effect of Electronic Waste on Human Health’’,” EJECE, vol. 7, pp. 28–34, 2023.

[4] Shah D., Patel S., Bharti S.K. Heart disease prediction using machine learning techniques SN Comput. Sci., 1 (2020), p. 345.

[5] B. Sethuraman and S. Niveditha, “Cerebrovascular Accident Prognosis using Supervised Machine Learning Algorithms,” in 2023 World Conference on Communication & Computing (WCONF, RAIPUR, India: IEEE, pp. 1–8. doi: 10.1109/WCONF58270.2023.10235122.

[6] Ali M.M., Paul B.K., Ahmed K., Bui F.M., Quinn J.M.W., Moni M.A., Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison, Comput. Biol. Med., 136 (2021),

[7] Palle Pramod Reddy, Dirisinala Madhu Babu, Hardeep Kumar, and Dr. Shivi Sharma, 2021 IJCRT | Volume 9, Issue 5 May 2021 | ISSN: 2320-2882

[8] Ghiasi M.M., Zendehboudi S. Decision tree-based methodology to select a proper approach for wart treatment Comput. Biol. Med., 108 (April) (2019), pp. 400-409, 10.1016/j.compbiomed.2019.04.001

[9] Batra M., Agrawal R.Comparative analysis of decision tree algorithms Adv. Intell. Syst. Comput., 652 (2018), pp. 31-36, 10.1007/978-981-10-6747-1_4

[10] Ghiasi M.M., Zendehboudi S., Mohsenipour A.A. Decision tree-based diagnosis of coronary artery disease: CART model Comput. Methods Programs Biomed., 192 (2020), 10.1016/j.cmpb.2020.105400.

[11] Tiwari A., Chugh A., Sharma A. Ensemble framework for cardiovascular disease prediction Comput. Biol. Med., 146 (2022), 10.1016/j.compbiomed.2022.105624

[12] Sathishkumar, Krishnan; Chaturvedi, Meesha; Das, Priyanka; Stephen, S.; Mathur, Prashant. Cancer incidence estimates for 2022 & projection for 2025: Result from National Cancer Registry Programme, India. Indian Journal of Medical Research 156(4&5):p 598-607, Oct–Nov 2022. | DOI: 10.4103/ijmr.ijmr_1821_22

[13] A.S. Nath, A. Pal, S. Mukhopadhyay, and K.C. Mondal, "A survey on cancer prediction and detection with data analysis", Innov. Syst. Softw. Eng., vol. 16, no. 3, pp. 231-243, 2019.

[14] Sarrio-Sanz, P.; Martinez-Cayuelas, L.; Beltran-Perez, A.; Muñoz-Montoya, M.; Segura-Heras, J.-V.; Gil-Guillen, V.F.; Gomez-Perez, L. A Novel Decision Tree Model for Predicting the Cancer-Specific Survival of Patients with Bladder Cancer Treated with Radical Cystectomy. J. Clin. Med. 2024, 13, 2177. https://doi.org/10.3390/jcm13082177

[15] Rayan Alanazi, Department of Computer Science, College of Science and Arts in Qurayyat, Jouf University, Sakakah, Saudi Arabia, Hindawi, Journal of Healthcare Engineering Volume 2022, Article ID 2826127

[16] Krishna S, Lakshman A, Archana T, Raja K, Ayyadurai M. Lung Cancer Prediction and Classification Using Decision Tree and VGG16 Convolutional Neural Networks. Open Biomed Eng J, 2024; 18: e18741207290271.

[17] T. Santhanam and M. S. Padmavathi, “Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis,” Procedia Comput. Sci., vol. 47, no. C, pp. 76–83, 2014.

[18] https://www.who.int/india/health-topics/mobile-technology-for-preventing-ncds

[19] Wei Yu, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, Muin J Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, BMC Medical Informatics Decision Making 10, Artical number 16, (2010).

[20] Smith, A., et al. (2020). Analysis of Decision Trees for Diabetes Prediction Using Pima Indian Diabetes Dataset. Journal of Medical Informatics, 35(2), 123-135.

[21] Pham, N.A. Huy and Triantaphyllou, Evangelos, “Prediction of Diabetes by Employing a NewData Mining Approach Which Balances Fitting and Generaliz ation,” Computer and InformationScience, a book of the Springer series of books titled: “Studies in ComputationalIntelligence,” (Roger Yin Lee, Editor), Springer, Heidelberg, Germany, Chapter 2, pp. 11-26, 2008.

[22] K. V. S. R. P. Varma, A. A. Rao, T. Sita Maha Lakshmi, and P. V. Nageswara Rao, “A computational intelligence approach for a better diagnosis of diabetic patients,” Comput. Electr. Eng., vol. 40, no. 5, pp. 1758–1765, 2014.

[23] Viloriaa A., Herazo-Beltranb Y., Cabrerac D., Pinedad O.(2020), Diabetes Diagnostic Prediction Using Vector Support Machines, Procedia Computer Science 170 (2020) 376–381. https://doi.org/10.1016/j.procs.2020.03.065.

[24] Johnson, R., & Lee, M. (2021). Comparative Analysis of Machine Learning Algorithms for Diabetes Prediction. International Journal of Healthcare Analytics, 12(4), 45-52.

[25] Kumar, P., & Singh, S. (2022). Hybrid Decision Tree Models for Diabetes Risk Assessment. Computational Health Studies, 28(3), 78-89.

[26] Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., & Ramesh Thelkar, A. (2022). Implementation of a heart disease risk prediction model usingmachine learning.Computational and Mathematical Methodsin Medicine,2022,1–14.https://doi.org/10.1155/2022/6517716.

[27] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heartdisease prediction using machine learning. International Journalof Engineering Research & Technology,9(4), 659–662. http://doi.org/10.17577/IJERTV9IS040614.

[28] Tiwari, A., et al. (2020). Application of Decision Trees for Predicting Heart Disease Using Cleveland Dataset. Journal of Medical Systems, 44(3), 25-30.

[29] Sharma, R., & Gupta, S. (2021). Comparison of Machine Learning Models for Cardiovascular Disease Risk Prediction. Computational Health Informatics, 38(4), 112-120.

[30] Alqahtani, M., et al. (2022). Ensemble Decision Tree Models for Heart Disease Prediction. Biomedical Informatics Research, 15(2), 98-105.

[31] Patel, K., et al. (2020). A Hybrid Approach to Predicting Heart Disease Using Decision Trees and Genetic Algorithms. International Journal of Health Analytics, 24(3), 56-63

[32] Chen, L., et al. (2021). Explainable AI for Cardiovascular Disease Prediction Using Decision Trees. Artificial Intelligence in Medicine, 44(5), 78-85.

[33] Johnson, P., et al. (2023). Advanced Decision Tree Models for Early Detection of Cardiovascular Disease. Journal of Predictive Analytics in Healthcare, 10(4), 150-159.

[34] Ahmed, A., et al. (2021). Regional Data and Decision Tree Analysis for Heart Disease. Medical Data Science, 12(1), 45-53.

[35] S. S. Mayuri Asabe, N. Dolare, S. Chorghade, and K. R. Pathak, Heart Attack Prediction and Analysis System Using Decision Tree Algorithm. 2020.

[36] Boukhatem, C., Youssef, H. Y., & Nassif, A. B. (2022). Heart diseaseprediction using machine learning.Advances in Science andEngineering Technology International Conferences,1–6,http://doi.org/10.1109/ASET53988.2022.9734880