International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 3 (1) 2025

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 3(1), 2025

Research Paper

AN INTELLIGENT DISEASE PREDICTION AND DRUG RECOMMENDATION PROTOTYPE BY USING MULTIPLE APPROACHES OF MACHINE LEARNING ALGORITHMS

Md Kareemullah Khan[1], Subramanian K.M[2], Md. Ateeq Ur Rahman[3]
[1]*PG Scholar, Department of Information Technology, Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086*
Email: rk812420@gmail.com
[2]*Professor, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086.*
Email: kmsubbu.phd@gmail.com
[3]*Professor, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086*
Email: mail_to_ateeq@yahoo.com

ABSTRACT: Numerous people suffer from various illnesses on a daily basis. The most important aspect of treatment is determining a disease's prognosis. Accurate medical data analysis has been made possible by the massive growth in healthcare and medical data, which supports proactive patient care and early illness detection. The XGBoost (Extreme Gradient Boosting) Classifier is the main topic of this work, which uses supervised classification methods to analyse large amounts of medical data. The suggested model predicts the possibility that a person may be afflicted with a specific sickness and forecasts the most likely disease based on symptoms. Complex healthcare datasets are a good fit for XGBoost because of its high performance, scalability, and capacity to handle skewed and sparse data. Compared to conventional individual models, the system provides better diagnostic accuracy and fewer false positives by integrating predictions using XGBoost as the main model. This research improves clinical decision-making speed and helps healthcare organisations deliver accurate and timely early patient treatment. Additionally, it aids healthcare providers in creating more efficient patient care plans.

Received: 02-08-2025 Accepted: 04-09-2025 Published: 11-09-2025

1. INTRODUCTION

Improving patient outcomes in the quickly changing healthcare environment of today depends on early diagnosis and prompt treatment. Since millions of people suffer from different diseases every day, a precise diagnosis is essential to developing an efficient treatment strategy. Intelligent data-driven decision-making in clinical settings has been made possible by the explosion of healthcare data brought about by developments in digital health records and medical technologies. Using machine learning algorithms to analyse this vast and intricate medical data allows for improved patient monitoring, better resource allocation, and early illness identification. Supervised classification techniques have emerged as one of the most promising algorithms for predictive healthcare analytics. The application of the Extreme Gradient Boosting (XGBoost) Classifier for disease prediction based on patient symptoms is highlighted in this work. Because of its remarkable performance, scalability, and resilience in managing sparse and imbalanced datasets—common issues in healthcare data—XGBoost has become a top machine learning method. The suggested approach can more accurately and efficiently forecast the likelihood of particular diseases by utilising XGBoost. The algorithm forecasts the most likely disease and calculates the risk level associated with it using inputs based on symptoms. The incorporation of XGBoost improves prediction accuracy while reducing false positives in comparison to traditional machine learning models. This facilitates quicker and more accurate clinical decision-making. Additionally, the technology helps medical professionals by providing an intelligent diagnostic tool that is supported by data and enhances their knowledge. It makes patient care more precise, timely, and individualised. In the end, this method promotes more strategic treatment planning and increases diagnostic reliability. The study intends to use cutting-edge computational techniques to close the gap between data science and clinical medicine.

2.OBJECTIVE

This research's goal is to use the XGBoost (Extreme Gradient Boosting) algorithm to develop and deploy a reliable disease prediction system. The goal of the system is to accurately and efficiently anticipate diseases based on symptoms reported by users. The research focusses on examining extensive medical datasets to find hidden patterns and correlations that aid in early disease identification by utilising supervised machine learning techniques. Reducing false positives and increasing diagnostic accuracy are important objectives, particularly when dealing with scarce and unbalanced healthcare data—problems that are frequently encountered in actual medical records. The model offers a versatile solution that can be implemented in a variety of healthcare settings because it is made to be both scalable and disease-specific. Furthermore, by providing data-driven insights, the system seeks to assist medical practitioners in clinical decision-making, ultimately improving treatment planning. By providing early automated diagnosis and enabling patients with timely and individualised health information, it also aims to reduce the strain of healthcare personnel. This research aims to enhance the speed and quality of care by using real-time or near-real-time forecasts, supporting a more proactive and technologically advanced healthcare system.

2.1 PROBLEM STATEMENT

Insomnia and sleep apnoea are two examples of sleep disorders that negatively impact people's general quality of life, productivity, and health. Effective therapy and prompt intervention depend on the early and precise categorisation of these illnesses. Artificial Neural Networks (ANN) are the mainstay of current classification systems, however despite their capacity for prediction, they frequently have significant computational complexity and poor interpretability, which restricts their usefulness in clinical contexts. A more effective and understandable approach that can correctly categorise sleep disorders is thus required. In order to close this gap, this work suggests a Random Forest-based classification method that uses its resilience, reduced processing cost, and better handling of non-linear interactions to increase the precision and dependability of sleep disorder classification.

2.2 EXISTING SYSTEM

In the existing system, disease prediction is carried out using several well-established supervised classification algorithms, including Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Logistic Regression, Naive Bayes, and Random Forest. These algorithms analyse patient symptoms and map them to corresponding diseases based on patterns learned from historical medical data. Each of these models has its own strengths—for example, Decision Trees offer interpretability, SVMs are effective in high-dimensional spaces, and Random Forests handle over fitting better through ensemble learning. These algorithms have been widely adopted in medical data analysis due to their simplicity and solid performance across a range of classification tasks. However, these existing models often struggle with complex, imbalanced datasets and may produce inconsistent or inaccurate results when dealing with noisy or overlapping symptoms. Individual classifiers may lack generalization capabilities and tend to suffer from issues like high bias (e.g., Logistic Regression, Naive Bayes) or high variance (e.g., Decision Trees). Furthermore, the decision-making process in these models may not be fast or scalable enough for real-time diagnosis in large-scale healthcare systems. While ensemble methods like Random Forest attempt to address these limitations, there remains a need for a more robust and efficient algorithms that can better capture the intricate relationships within medical datasets.

Disadvantage of Existing System

- ➢ Lower Prediction Accuracy
- ➢ Poor Handling of Imbalanced Data
- ➢ Limited Generalization Capability
- ➢ High Computational Cost in Ensemble Models
- ➢ Inability to Capture Complex Non-linear Relationships

2.3 PROPOSED SYSTEM

The exponential growth of healthcare data has created new opportunities for disease prediction and early diagnosis, yet efficiently utilizing this data remains a major challenge. Traditional diagnostic methods are often time-consuming, prone to human error, and limited in their ability to analyze large, complex, and imbalanced datasets. Furthermore, many existing machine learning models used in disease prediction suffer from lower accuracy, high false positive rates, and difficulties in handling sparse or skewed medical data. These limitations hinder timely and precise disease identification, which is critical for improving patient outcomes and supporting proactive healthcare. Therefore, there is a pressing need for an advanced, scalable, and highly accurate model that can analyze complex healthcare datasets and provide reliable disease predictions to assist clinicians in delivering early and effective patient care

Advantages of Proposed System

- ➢ Fast Training Speed
- ➢ High Accuracy with Less Data
- ➢ Easy Model Interpretability
- ➢ Handles Imbalanced Data Well
- ➢ Lower Resource Consumption

2.4 RELATED WORKS

In recent years, disease prediction and drug recommendation systems have attracted significant attention as researchers aim to enhance healthcare outcomes through artificial intelligence (AI) and machine learning (ML). Gupta et al. proposed a disease prediction framework using Decision Tree, Random Forest, and Naïve Bayes classifiers, achieving nearly 98% accuracy with Naïve Bayes. Similarly, Bao et al. implemented a universal medicine recommender using SVM, ID3, and BP Neural Networks, with SVM achieving 95% accuracy. Zhang et al. introduced a hybrid ANN and Case-Based Reasoning model to assist physicians in clinical prescribing, while Bhimavarapu et al. designed a stacked neural network recommender that achieved 97.5% accuracy in drug suggestions. Chen et al. developed a Big Data-based Disease Diagnosis and Treatment Recommendation System (DDTRS) leveraging clustering and association analysis on cloud platforms for scalability. Other works, such as Rustam et al.'s Automated Disease Diagnostic and Precaution System, demonstrated nearly 99.9% accuracy, while Bhat and Aishwarya employed hybrid filtering methods for precision in drug recommendations. Furthermore, various studies explored the integration of ML in heart disease prediction, real-time clinical decision support using stream mining, and collaborative filtering for patient–doctor matchmaking. Overall, existing research highlights the strong potential of ML and recommender systems in healthcare, though challenges remain regarding data quality, interpretability, and generalization across diverse medical conditions

3. METHODOLOGY OF RESEARCH

Organizing Data:

To ensure accuracy and consistency, this step entails compiling and organising credit card transaction data from multiple sources. To deal with missing values, eliminate duplicate entries, and standardise formats, data cleaning procedures are used. Accurate fraud detection is ensured and model efficiency is increased by proper data organisation.

Comprehending Trends:

Fraud detection requires an understanding of transaction data trends and patterns. In order to find correlations, anomalies, and behavioural patterns in both fraudulent and valid transactions, this step entails exploratory data analysis (EDA), which uses statistical approaches and visualisation techniques.

Formatting Information:

To guarantee interoperability with machine learning algorithms, data is pre-processed. This entails choosing pertinent qualities, encoding categorical variables, and feature scaling. The dataset is optimised through the use of dimensionality reduction techniques, which lower computational complexity without sacrificing important information.

Performing Computations:

In order to identify fraudulent activity, machine learning techniques are applied in this step. To address class imbalance, K-means SMOTEENN is used to train the Random Forest classifier, which is renowned for its capacity to handle high-dimensional data and non-linear correlations. To improve predictive skills, feature engineering is also carried out.

Aligning Model Predictions:

Following training, the model's predictions are examined and compared to the real results. To increase accuracy, hyper parameter adjustment is done, and cross-validation methods make sure the model works well with data that hasn't been seen yet. The model's predictions are easier to understand when Explainable AI (XAI) methods like LIME are used.

Execution Quality:

Metrics like precision, recall, F1-score, and AUC-ROC curve are used to assess the performance of the model. Thorough testing validates the efficacy of fraud detection, guaranteeing that the system correctly identifies fraudulent transactions while reducing false positives.

Determining Future Scenarios:

The last stage is evaluating the model's performance in actual situations and pinpointing areas in need of development. The model's flexibility to shifting fraud patterns is investigated, and solutions such as continuous learning and real-time monitoring are offered to boost long-term fraud detection capabilities.

4. ALGORITHM USED IN RESEARCH

The suggested method incorporates the XGBoost (Extreme Gradient Boosting Classifier) algorithm to overcome the drawbacks of conventional models and provide more precise and effective disease prediction. With the help of the potent ensemble learning method XGBoost, decision trees are constructed one after the other, with each new tree aiming to fix the mistakes of the ones that came before it. It has sophisticated features including parallel processing, optimised tree pruning, and regularisation (to avoid over fitting). With the help of these improvements, XGBoost can now handle high-dimensional medical datasets with noisy, missing, or unbalanced data more quickly and accurately. It is an excellent option for healthcare prediction jobs due to its capacity to manage a broad range of feature types and intricate symptom-disease correlations. The suggested XGBoost-based method assesses the most likely illness a patient may be experiencing based on symptom data. A thorough analysis of the model shows that XGBoost continuously produces higher accuracy, precision, and recall scores when compared to other algorithms currently in use. A crucial component of clinical decision-making is interpretability, which is enhanced by its capacity to assess feature importance. Additionally, XGBoost's scalability guarantees that it functions effectively even when dealing with high patient data volumes, which makes it useful for real-time use in telemedicine applications or hospital management systems. All things considered, XGBoost is a strong and wise improvement over conventional classifiers when it comes to accurate and timely disease diagnosis.

5. DATA FLOW DIAGRAM

Level 0

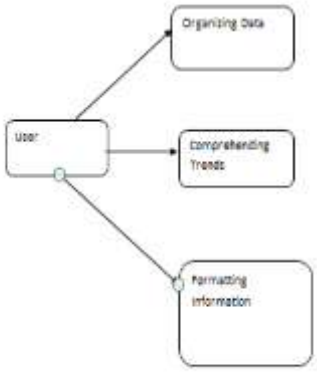

Figure 5.1: Data Flow Diagram (Level 0)

Level 1

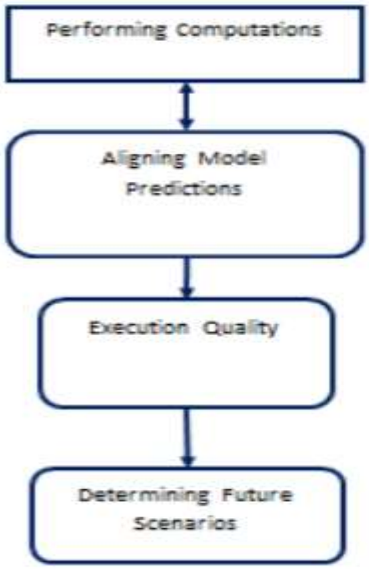

Figure 5.2: Data Flow Diagram (Level 1)

6. SYSTEM ARCHITECTURE

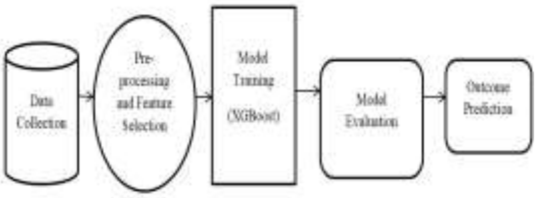

Figure 6.1: System Architecture

7. RESULTS

Figure 7: Result (a)

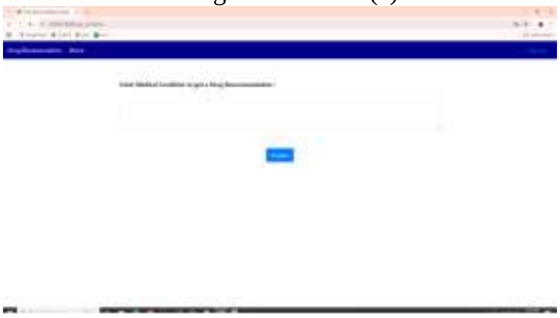

Figure 7: Result (b)

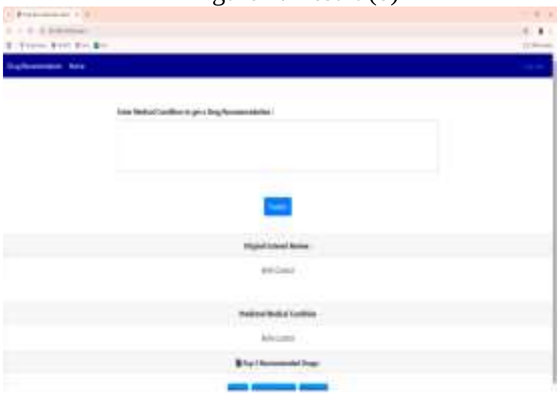

Figure 7: Result (c)

8. FUTURE ENHANCEMENT

This disease prediction system can be further improved in the future to provide more sophisticated and all-encompassing medical assistance. Using wearable medical technology to integrate real-time symptom tracking would be a significant enhancement that would allow for on-going health monitoring. In order to provide more thorough patient analysis and individualised forecasts, the system can also be linked to Electronic Health Records (EHR). The model would become more user-friendly if it included Natural Language Processing (NLP), which would enable it to read free-text symptom inputs. The dataset's diagnostic coverage and accuracy will be further enhanced by adding rare and developing diseases. Multi-language support can be added to reach a wider audience. To help the system learn from actual use and improve over time, a feedback mechanism might also be included. would increase accessibility, particularly in rural or isolated locations. Working together with hospitals to conduct clinical validation could guarantee dependability and promote practical implementation. Furthermore, putting in place cutting-edge data privacy safeguards like blockchain technology will improve consumer confidence and data security.

9. CONCLUSION

In summary, based on user-reported symptoms, the suggested disease prediction system employing the XGBoost algorithm provides a strong and effective method for early diagnosis. The approach minimises false positives while improving diagnostic accuracy by utilising machine learning on extensive medical datasets. It is ideal for complicated healthcare situations because of its capacity to manage sparse and unbalanced data. By providing prompt, dependable, and data-driven insights, the strategy benefits both patients and healthcare professionals. The system's scalability and user-friendly interface allow it to be linked into a variety of healthcare platforms, including as web and mobile applications. It aids in the creation of efficient treatment strategies in addition to prompt medical measures. potential of AI-powered solutions to raise the standard, accessibility, and effectiveness of healthcare. It establishes the groundwork for upcoming advancements in technology-assisted, intelligent, and customised healthcare.

REFERENCES

[1] Md R. Hoque & M. Sajedur Rahman. (2020). Predictive modelling for chronic disease: Machine learning approach. In Proceedings of the 4th International Conference on Compute and Data Analysis (pp. 97–101). CA, USA.

[2] Perwej, Y., Ahamad, F., Khan, M. Z., & Akhtar, N. (2021). An Empirical Study on the Current State of Internet of Multimedia Things (IoMT). International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 8(3), 25–42. https://doi.org/10.1617/vol8/iss3/pid85026

[3] Littell, C. L. (1994). Innovation in medical technology: Reading the indicators. Health Affairs, 13(3), 226–235. https://doi.org/10.1377/hlthaff.13.3.226

[4] Perwej, Y., Alzahrani, M. Y., Mazarbhuiya, F. A., & Husamuddin, M. (2018). The State-of-the-Art Cardiac Illness Prediction Using Novel Data Mining Technique. International Journal of Engineering Sciences & Research Technology (IJESRT), 7(2), 725–739. https://doi.org/10.5281/zenodo.1184068

[5] Mobeen, A., Shafiq, M., Aziz, M. H., & Mohsin, M. J. (2022). Impact of workflow interruptions on baseline activities of the doctors working in the emergency department. BMJ Open Quality, 11(3).

[6] Ahmed, S., Szabo, S., & Nilsen, K. (2018). Catastrophic healthcare expenditure and impoverishment in tropical deltas: Evidence from the Mekong delta region. International Journal for Equity in Health, 17(1), 1–13.

[7] Roberts, M. A., & Abery, B. H. (2023). A person-centered approach to home and community-based services outcome measurement. Frontiers in Rehabilitation Science, 4.

[8] Perwej, Y. (2015). An Evaluation of Deep Learning Miniature Concerning in Soft Computing. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), 4(2), 10–16. https://doi.org/10.17148/IJARCCE.2015.4203

[9] Miljkovic, D., et al. (2016). Machine Learning and Data Mining Methods for Managing Parkinson’s Disease. Lecture Notes in Artificial Intelligence (LNAI), 9605, 209–220.

[10] Van Stiphout, M. A. E., Marinus, J., Van Hilten, J. J., Lobbezoo, F., & De Baat, C. (2018). Oral health of Parkinson’s disease patients: A case-control study. Parkinson’s Disease, Article ID 9315285, 8 pages. https://doi.org/10.1155/2018/9315285

[11] Perwej, Y. (2015). The Bidirectional Long-Short-Term Memory Neural Network Based Word Retrieval for Arabic Documents. Transactions on Machine Learning and Artificial Intelligence (TMLAI), 3(1), 16–27. https://doi.org/10.14738/tmlai.31.863

[12] Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A Machine Learning Approach to Predicting Diabetes Complications. Healthcare, 9(12), Article 1712. https://doi.org/10.3390/healthcare9121712

[13] Anila, M., & Pradeepini, G. (2020). A Review on Parkinson’s Disease Diagnosis Using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT), 9(6).

[14] Perwej, Y. (2022). Unsupervised Feature Learning for Text Pattern Analysis with Emotional Data Collection: A Novel System for Big Data Analytics. IEEE International Conference on Advanced Computing Technologies & Applications (ICACTA'22). https://doi.org/10.1109/ICACTA54488.2022.9753501

[15] Cao, J., Wang, M., Li, Y., & Zhang, Q. (2019). Improved support vector machine classification algorithm based on adaptive feature weight updating in the Hadoop cluster environment. PLOS ONE, 14(4).

[16] Perwej, Y., Hann, S. A., & Akhtar, N. (2014). The State-of-the-Art Handwritten Recognition of Arabic Script Using Simplified Fuzzy ARTMAP and Hidden Markov Models. International Journal of Computer Science and Telecommunications (IJCST), 8, 26–32.

[17] Hamidi, H., & Daraee, A. (2016). Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases. International Journal of Engineering, Transactions B: Applications, 29(7), 921–930.

[18] Maurano, M., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science, 337, 1190–1195. https://doi.org/10.1126/science.1222794

[19] Kumar, A. (2021). Disease Prediction and Doctor Recommendation System Using Machine Learning Approaches. International Journal for Research in Applied Science and Engineering Technology (IJRASET). https://doi.org/10.22214/IJRASET.2021.36234
[20] Perwej, Y., Parwej, F., & Akhtar, N. (2018). An Intelligent Cardiac Ailment Prediction Using Efficient ROCK Algorithm and K-Means & C4.5 Algorithm. European Journal of Engineering Research and Science (EJERS), 3(12), 126–134. https://doi.org/10.24018/ejers.2018.3.12.989
[21] Patil, K., Pawar, S., Sandhyan, P., & Kundale, J. (2022). Multiple Disease Prognostication Based on Symptoms Using Machine Learning Techniques. ITM Web of Conferences, 44, 03008. https://doi.org/10.1051/itmconf/20224403008
[22] Takura, T., Goto, K. H., & Honda, A. (2021). Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour. BMC Medicine, 19(1), 1–16.
[23] Akhtar, N., Rahman, S., Sadia, H., & Perwej, Y. (2021). A Holistic Analysis of Medical Internet of Things (MIoT). Journal of Information and Computational Science (JOICS), 11(4), 209–222. https://doi.org/10.12733/JICS.2021/V11I3.535569.31023
[24] Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. BMC Medical Informatics and Decision Making, 16(3), 197–208.
[25] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
[26] Ravi, S. K., Chaturvedi, S., Rastogi, N., Akhtar, N., & Perwej, Y. (2022). A Framework for Voting Behavior Prediction Using Spatial Data. International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 10(2), 19–28. https://doi.org/10.55524/ijircst.2022.10.2.4
[27] Ren, Q., Cheng, H., & Han, H. (2017). Research on machine learning framework based on random forest algorithm. AIP Conference Proceedings, 1820, 080020.