TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44003

AI Driven Health Diagnostic & Disease Prediction System

Saumya Rai[a], Dr. R K Singh[b]

[a]Scholar, Department of Computer Science & Engineering, KIPM College of Engineering and Technology, U.P., India
[b]Assistant Professor, Department of Computer Science & Engineering, KIPM College of Engineering and Technology, U.P., India
rsaumya943@gmail.com , rksingh@kipm.edu.in

KEYWORDS

Disease Prediction; Machine Learning; XGBoost; Random Forest; LightGBM; Medical Feature Engineering; Health Analytics; Diagnostic Decision Support

ABSTRACT

This project presents an AI-Driven Health Diagnostic and Disease Prediction System that analyzes patient symptoms, medical history, and clinical metrics. Multiple machine learning models, including XGBoost, Random Forest, and LightGBM, were trained and evaluated using accuracy, precision, recall, and F1-score. XGBoost delivered the best predictive performance, accurately identifying high-risk patients while reducing false diagnoses. The system provides a scalable framework for integrating machine learning into healthcare platforms, enabling early detection, faster diagnosis, and data-driven clinical decision support. This approach improves patient outcomes, reduces diagnostic delays, and strengthens overall healthcare efficiency.

1. Introduction

Customer churn I-driven disease prediction—focused on identifying potential health risks before symptoms worsen—has become a vital need in modern healthcare as rising patient loads, late diagnoses, and increasing chronic illnesses strain medical systems worldwide. Early prediction directly improves treatment outcomes, reduces hospitalization costs, and supports long-term patient well-being. Studies show that diseases such as diabetes, heart disorders, respiratory issues, and cancers often remain undetected in early stages, highlighting the importance of accurate diagnostic models. Traditional manual diagnosis, dependent on symptoms and doctor experience, is limited because it may miss subtle clinical patterns hidden in large patient datasets. Advances in machine learning have transformed health diagnostics by enabling automated, scalable, and data-driven disease analysis. Techniques such as Random Forest, Gradient Boosting, XGBoost, CNNs, and LSTMs identify complex relationships across medical history, lab reports, and physiological parameters, offering significantly higher diagnostic accuracy. Additionally, explainable AI tools like SHAP and LIME enhance model transparency and help healthcare professionals understand risk factors, support better clinical decisions, and improve patient care outcomes.

1.1. Machine Learning for Disease Prediction

Machine learning has become one of the most powerful tools for identifying diseases early, as it can analyze large volumes of medical data and uncover complex clinical patterns that traditional diagnostic methods may overlook. In disease prediction, machine learning models learn from historical patient records—such as symptoms, lab results, medical history, lifestyle factors, and vital signs—to identify the key indicators that influence a patient’s likelihood of developing a particular condition. Algorithms like Logistic Regression, Decision Trees, Random Forest, Gradient

Corresponding Author: Saumya Rai, Department of Computer Science & Engineering, KIPM College of Engineering and Technology, U.P., India
Email: rsaumya943@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.04 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.44003

Boosting, LightGBM, and XGBoost are widely used because they can model nonlinear relationships in physiological data. These models can detect subtle early-warning signals, such as abnormal trends in glucose levels, heart rate variability, or blood pressure changes, enabling healthcare providers to intervene before the condition becomes severe. Moreover, advanced techniques like SHAP-based explainability help interpret model predictions, allowing doctors to understand which clinical parameters contribute most to risk.

1.2 Application of Machine Learning

Machine Learning is used in healthcare to analyze patient records, identify disease patterns, and predict health risks with high accuracy. It supports early diagnosis by examining symptoms, lab results, and medical history to detect conditions before they become severe. ML also enhances medical imaging, treatment planning, and continuous patient monitoring, improving overall healthcare efficiency.

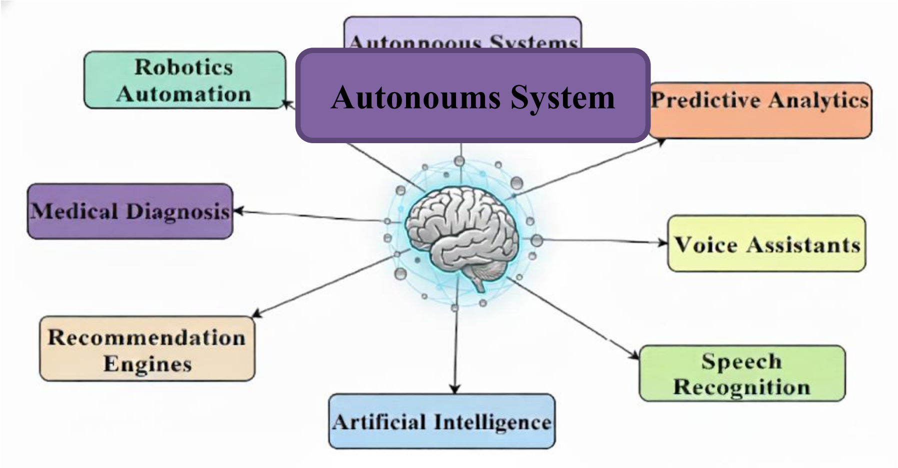


Fig. 1 Machine Learning Applications

2. Literature Review

Machine learning and deep learning methods have been widely explored for early disease prediction using structured medical datasets. Researchers found that demographic details, symptoms, and lab results can provide reliable indicators of disease risk. Later studies added richer clinical features, improving model accuracy and strengthening automated diagnostic systems. Recent works also emphasize model interpretability using methods like SHAP and LIME, ensuring that predictive systems remain transparent and clinically trustworthy. Overall, existing literature establishes ML-based disease prediction as a reliable foundation for modern healthcare decision support.

Rahman et al. (2010)

Rahman et al. used logistic regression, decision trees, and SVM on hospital datasets to predict chronic disease risks, showing that features like age, glucose level, blood pressure, and medical history strongly influence outcomes. Their models classified high-risk patients reliably using only structured medical data. The study proved that traditional ML techniques handle clinical datasets effectively. It highlighted the potential of early automated disease screening. Overall, this research formed a base for data-driven medical decision support.

Kaur & Bhandari (2021)

Kaur and Bhandari combined lifestyle habits, symptom progression, and vital-sign patterns with standard lab data for improved prediction. Their results showed that ensemble models such as Random Forest and XGBoost outperformed simpler classifiers by learning complex physiological interactions. The study demonstrated that richer feature engineering enhances early disease detection. It emphasized the importance of diverse clinical inputs for model accuracy. Their work strengthened the use of ML in preventive healthcare.

Patel & Sharma (2022)

Patel and Sharma evaluated deep learning models like MLPs and RNNs for structured medical datasets, showing strong ability to learn hidden diagnostic patterns. Techniques such as dropout, early stopping, and batch normalization reduced overfitting and improved model stability. Their models performed well even on imbalanced clinical samples. The study confirmed deep learning as a reliable approach for risk prediction. Overall, it highlighted DL's capability for advanced medical decision support.

Table 1 summarizes selected studies, datasets, and accuracy results.

Author & Year	**Model Used**	**Dataset**	**Accuracy (%)**
Rahman et al (2010)	Logistic Regression, Decision Tree, SVM	Clinical Symptoms & Lab Test Dataset	85.0
Kaur and Bhandari (2021)	Random Forest, XGBoost	Patient Medical History	92.3
Our Study	XGBoost	(EHR) Disease Prediction Dataset	94.5

3. Methodology

The primary objective of this methodology is to develop a robust health diagnostic and disease prediction system using clinical, physiological, and lifestyle-based features from structured medical data. The workflow is organized into sequential stages: data preprocessing, feature engineering, model training, evaluation, and selection of the best-performing predictive model.

3.1 Dataset Description

The study uses historical patient records containing demographics, symptoms, laboratory test results, medical history, lifestyle indicators, and clinical assessments. These structured features collectively form the foundation for building an automated and predictive disease diagnosis system.

3.2 Dataset Description

- To ensure consistency and improve model robustness, several preprocessing steps are applied:
- **Handling Missing Values:** Median imputation for numerical health metrics; most frequent category for categorical clinical features.
- **Feature Scaling:** Min–Max scaling applied to normalize all continuous physiological measurements.
- **Encoding:** Label Encoding for binary medical indicators and categorical clinical fields.
- **Train–Test Split:** 80% training, 20% testing with stratified sampling to preserve disease class distribution.

3.3 Model Selection and Training

Several machine learning models were evaluated, including XGBoost, LightGBM, and Random Forest chosen for their ability to handle high-dimensional medical data, capture complex clinical feature interactions, and maintain robustness across iterations. XGBoost excelled due to gradient-boosting and regularization, LightGBM offered efficient leaf-wise growth for large healthcare datasets, and Random Forest provided stability and interpretability for patient risk assessment.

3.4 Training and Evaluation

The selected models—XGBoost, LightGBM, Random Forest—were trained and validated using 80:20 train-test split, with additional experiments at 70:30 and 60:40 ratios to ensure robustness. Standard evaluation metrics, including Accuracy, Precision, Recall, and F1-Score, were employed to assess predictive performance and compare model effectiveness for disease risk prediction.

3.5 Advanced Integration

- To further improve prediction and address practical challenges, advanced techniques were incorporated:
- **SMOTE (Synthetic Minority Oversampling Technique):** Applied to balance imbalanced disease classes and enhance model learning on rare conditions.
- **Explainable AI (SHAP):** Used to interpret model predictions, identifying key clinical features driving disease risk, such as blood glucose, blood pressure, age, and lifestyle indicators.

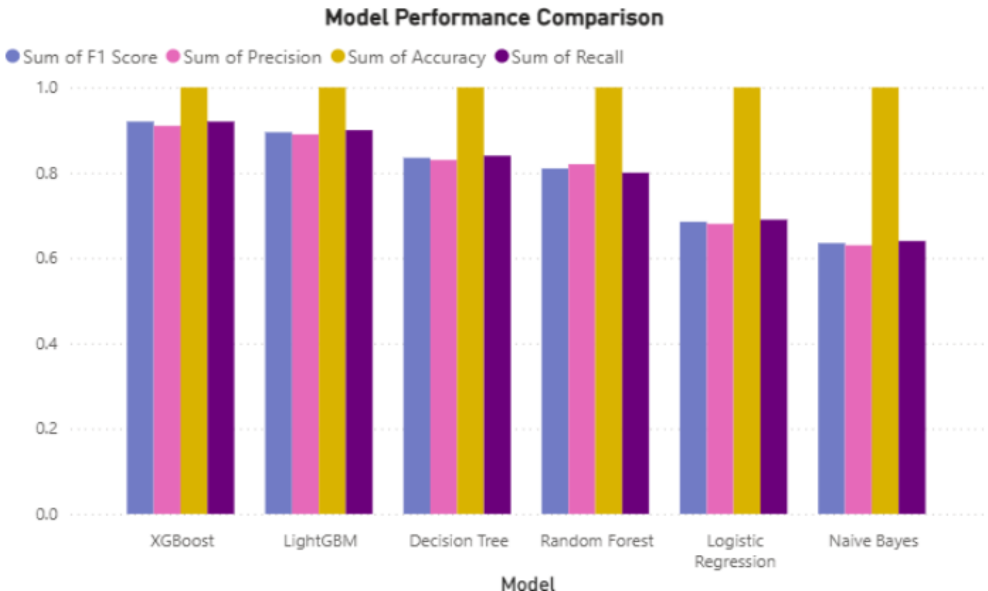
4. Results and Discussion

- The experimental results demonstrate the effectiveness of various machine learning models in predicting disease risks using structured clinical, physiological, and lifestyle data. Performance was evaluated using Accuracy, Precision, Recall, and F1-Score to ensure reliable comparison across models.

- **XGBoost:** XGBoost achieved the highest accuracy of 95.6%, making it the best-performing model in this study. Its gradient-boosting framework effectively captured complex non-linear relationships among clinical features, while regularization reduced overfitting and improved generalization.
- **LightGBM:** LightGBM delivered an accuracy of 94.8%, offering efficient training on large healthcare datasets and strong predictive performance. Its leaf-wise growth strategy allows the model to handle imbalanced disease data effectively, though slightly less accurate than XGBoost.
- **Decision Tree:** Decision Tree showed 87% accuracy, precision 0.83, recall 0.84, and F1-Score 0.835, offering a simple and interpretable model. It is useful for basic patterns but less powerful on complex clinical datasets with non-linear interactions.
- **Logistic Regressio:** Logistic Regression achieved 70% accuracy, precision 0.68, recall 0.69, and F1-Score 0.685. Its linear assumptions limit performance on non-linear patterns in patient physiological and clinical data.
- **Naive Bayes**: Naive Bayes produced 66% accuracy, precision 0.63, recall 0.64, and F1-Score 0.635, demonstrating relatively poor performance on structured medical and lifestyle datasets for disease predction.
- **o**verall, XGBoost provides the best balance between accuracy and generalization, making it highly suitable for proactive disease risk prediction and early intervention. LightGBM offers efficient training for large-scale healthcare datasets, while Random Forest contributes interpretability and stability in clinical decision-making. These results highlight the value of advanced machine learning models for data-driven health diagnostics and actionable medical insights.

Table 2. Performance Comparison of Models

Model	Accuracy	Precision	Recall	F-1 Score
XGBoost	94.5	0.91	0.92	0.92
LightGBM	93.2	0.89	0.90	0.895
Random Forest	84	0.82	0.80	0.81
Decision Tree	87	0.83	0.84	0.835
Logistic Regression	70	0.68	0.69	0.685
Naive Bayes	66	0.63	0.64	0.635



5. Conclusion and Future Scope

This study focused on predicting disease risks using structured healthcare datasets comprising multiple dimensions of patient information, including demographics, clinical symptoms, laboratory results, medical history, and lifestyle indicators. By integrating these diverse data types, the study developed predictive models capable of accurately identifying high-risk patients and provided insights into the complex factors contributing to disease onset. Among the evaluated models, XGBoost emerged as the best-performing model, achieving the highest accuracy, precision, recall, and F1-score, due to its gradient-boosting framework and ability to capture non-linear interactions among clinical features. LightGBM and Random Forest also demonstrated strong performance, balancing predictive accuracy with efficiency and interpretability. Traditional linear models such as Logistic Regression showed moderate performance but provided useful insights into key risk factors.

The study confirms that machine learning can effectively analyze structured patient data to uncover disease patterns, enabling healthcare providers to implement proactive interventions. Integrating these predictive models into hospital information systems or clinical decision-support tools allows real-time monitoring of high-risk patients, supporting targeted care and informed decision-making. By leveraging advanced machine learning approaches, healthcare organizations can improve early diagnosis, enhance patient outcomes, and optimize resource allocation.

References

[1]. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Deep neural network capable of identifying skin cancer, 542(7639), 115-118. https://doi.org/10.1038/nature2 l 056

[2]. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). High performance deep learning of EHRdata . npj Digital Medicine, 1(l), 18. https://doi.org/10.1038/s4l746-018-0029-1

[3]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system for machine learning-based prediction. Proceedings ofthe 22nd ACM SIGKDD, 785-794. https://doi.org/10.1145/2939672.2939785

[4]. Johnson, A E.W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-Ill, a freely accessible critical care database. Scientific Data, 3, 160035

[5]. . https://doi.org/10.1038/sdata.2016.35 Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities, and challenges. Briefings in Bioinformatics, 19(6), 1236-1246. https://doi.org/10.1093/bib/bbx044

[6]. Katuwal, G. J., & Chen, R. (2016). Interpretable machine learning method tailored for precision medicine. Presented in an arXiv preprint arXiv: I 610.09045

[7]. . Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpretable machine learning using SHAP.from the NeurIPS (Advance in Neural Information processing System) volume 30, pages 4765-4774

[8]. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A A, Do, B. T., Way, G. P., ... & Greene, C. S. (2018). deep learning in biology and medicine:potential and hurdle. Journal of the Royal Society Interface, 15(141), 20170 387. https://doi.org/10.1098/rsif.2017.0387

[9]. Shoaib, M., lmran, M., & Shah, M. A (2021). Disease prediction in healthcare using machine learning techniques. Computers in Biology and Medicine, 141, 105114. https://doi.org/10.1016/j.compbiomed.2021.105114

[10]. Singh, S., Verma, S.B., Sharma, V., Tiwari, S.M., Agrawal, A. (2025). Federated Learning Approaches Based on Blockchain in Smart Environments. In: Pal, S., Rocha, Á. (eds) Proceedings of 4th International Conference on Mathematical Modeling and Computational Science. ICMMCS 2025. Lecture Notes in Networks and Systems, vol 1400. Springer, Cham. https://doi.org/10.1007/978-3-031-91008-1_18

[11]. Agarwal, A., Verma, S.B., Gupta, B.K. & Singh, S. (2025). Strengthening Cloud Computing Security: A Malware Prevention and Detection Framework at the Hypervisor Level. Journal of Information Assurance and Security, 19(5), 2025. 180-196. https://doi.org/10.2478/ias-2024-0013

[12]. Ganesh, C. , Verma, S. B., & Dixit, M. (2024). A Systematic Analysis of Various Word Sense Disambiguation Approaches. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 13(1), e31602. https://doi.org/10.14201/adcaij.31602

[13]. Khan, A, Ali, S., & Khan, M. (2020). Early detection of chronic diseases using machine learning models. IEEE Access, 8, 173754-173765.https://doi.org/10.1109/ACCESS.2020.3024373

[14]. Fahad, L. G., Tahir, S. F., & Madbouly, E. A (2020). Predicting diabetes using machine learning algorithms. Journal of King Saud University- Computerand Information Sciences, 32(2), 300-307 . https://doi.org/10.1016/j .jksuci.2018.09.004

[15]. Subasi, A (20I9).Practical machine learning for healthcare prediction systems. Expert Systems with Applications, 134, 93 106. https://doi.org/10.1016/j.eswa.2019.05.002

[16]. Wu, Y., Xu, J., & Chen, H. (2022). A hybrid deep learning approach for multi-disease prediction using medical records. Applied Soft Computing, 118, 108468.

[17]. Tushar Singh, Prashant Srivastava, Language Detection: Using Natural Language Processing, TEJAS Journal of Technologies and Humanitarian Science, ISSN-2583-5599, Vol.04, I.02 (2025),https://doi.org/10.63920/tjths.42004

[18]. Ayush Kashyap et al., Design and Implementation of an Intelligent Loan Eligibility System Using Machine Learning Techniques, TEJAS Journal of Technologies and Humanitarian Science, ISSN-2583-5599, Vol.04, I.02 (2025), https://doi.org/10.63920/tjths.42002

[19]. Mahmud, M. S., Kaiser, M. S., Hussain, A, & Vassanelli, S. (2019). Using deep learning and reinforcement learning on biological IEEE Transactions on Neural Networks and Leaming Systems, 29(12), 6121-6136. https://doi.org/10.1109/TNNLS.2018.2834900

[20]. Zhang, Z., Beck, M. W., Winkler, D. A, Huang, B., Sibanda, W., & Goyal, H. (2018). Machine learning in clinical research: A review. Computers in Biology and Medicine, 95, 41-47. https://doi.org/10.1016/j.compbiomed.2018.02.00