IJARSCT

IJARSCT

ISSN: 2581-9429

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Volume 5, Issue 9, June 2025

ISO 9001:2015 QUALITY MANAGEMENT CERTIFIED COMPANY

Impact Factor: 7.67

Predictive Analytics in Healthcare: Early Detection of Chronic Diseases Using EHR Data

Prof. Shradha Wankhede[1], Prof. Priyanka Choudhary[2], Bhumika Nandardhane[3], Avinash Sawale[4]

Professor, Department of Computer Science and Engineering[1,2]
U.G. Student, Department of Computer Science and Engineering[3,4]
Tulsiramji Gaikwad-Patil College of Engineering & Technology, Mohgaon, Nagpur, Maharashtra, India
shradha.cse@tgpcet.com, priyankaghotekar22@gmail.com,
bhuminandardhane@gmail.com sawalea294@gmail.com

Abstract: *Chronic diseases like diabetes and hypertension are among the leading causes of illness and healthcare spending around the world. Detecting these conditions early can make a big difference helps patients get timely treatment and easy the strain on healthcare systems. In this study, we look at how machine learning can be used with Electronic Health Record (EHR) data to create early-warning systems for predicting chronic diseases.*

We used structured data such as lab results, medical diagnoses, prescribed medications, and patient demographics to train several predictive models. Out of all the models we tested, XGBoost and Random Forest delivered the best performance in terms of accuracy and AUC (area under the ROC curve). We also identified which features were most important for prediction and discussed how integrating these models into everyday clinical practice could support earlier interventions and better disease management.

Keywords: *Chronic diseases.*

I. INTRODUCTION

Chronic diseases—especially diabetes and hypertension—are among the top causes of death and long-term disability around the world. These illnesses often progress slowly and may go unnoticed until serious complications develop. However, with the help of predictive analytics, it's possible to catch them earlier, allowing for faster medical intervention, better patient outcomes, and reduced healthcare costs. Electronic Health Records (EHRs) offer a valuable source of detailed, long-term patient information that can be used to build predictive models. In this paper, we focus on developing and evaluating machine learning models that can detect chronic diseases early using real-world EHR data.

Our key contributions include comparing different models, identifying the most important features for prediction, and discussing how these models could be applied in actual clinical settings.

II. LITERATURE SURVEY

In recent years, researchers have increasingly used machine learning techniques to predict the onset of chronic diseases using Electronic Health Record (EHR) data. Models like logistic regression, decision trees, support vector machines, and deep learning have all shown encouraging results. A notable example is the work by Miotto et al. (2016), who developed Deep Patient—an unsupervised learning model capable of predicting disease onset based on patterns in EHR data.

Despite these advances, many of the existing models face challenges when it comes to interpretability and generalizability. This review summarizes the progress made so far, the most commonly used techniques, and the ongoing limitations, including issues like sparse data, class imbalance, and the lack of model transparency. In response to these gaps, our study focuses on using more interpretable models along with strong evaluation strategies to ensure both accuracy and real-world usability.

DOI: 10.48175/IJARSCT-28207

III. METHODOLOGY

3.1 Data Source

For this study, we used a de-identified dataset provided by a regional healthcare system, containing information from over 15,000 patients. The dataset included a wide range of clinical details such as demographic information, lab test results, medical diagnoses (based on ICD-10 codes), prescribed medications, and vital signs, all collected over five years.

3.2 Preprocessing

Before building the models, we cleaned and prepared the data. Missing values in continuous variables were filled using mean imputation, while categorical variables were handled using mode imputation. Records with a large amount of missing data were removed to maintain data quality. We also normalized numerical features to ensure consistency and applied one-hot encoding to convert categorical variables into a machine-readable format.

Data preprocessing involved:

- Normalization of numeric features (e.g., blood pressure, glucose levels).
- Encoding categorical variables (e.g., diagnoses, medications).
- Imputation of missing values using k-nearest neighbours and mean-substitution techniques.
- Temporal sequencing of events to retain the order of clinical encounters.

3.3 Feature Engineering

We selected features based on both their clinical importance and their statistical correlation with the onset of chronic diseases. In addition to existing variables, we derived new ones such as Body Mass Index (BMI) and trends in blood pressure over time. To better capture changes leading up to diagnosis, we applied time-windowing techniques to extract data from the year prior to disease onset.

3.4 Model Selection

Several machine learning models were tested in this study, including Logistic Regression, Random Forest, XGBoost, and Support Vector Machines. We used 5-fold cross-validation to ensure reliable performance comparisons across models.

3.5 Evaluation Metrics

To assess model performance, we used several standard metrics: accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (ROC-AUC). We also analyzed confusion matrices and ROC curves to better understand each model’s strengths and weaknesses.

3.6 Tools and frameworks

All analyses were conducted using Python. Key libraries included Pandas for data manipulation, Scikit-learn for modeling and evaluation, XGBoost for gradient boosting, and SHAP for interpreting model outputs.

IV. ARCHITECTURE

- Facilitating real-time updates:

DOI: 10.48175/IJARSCT-28207

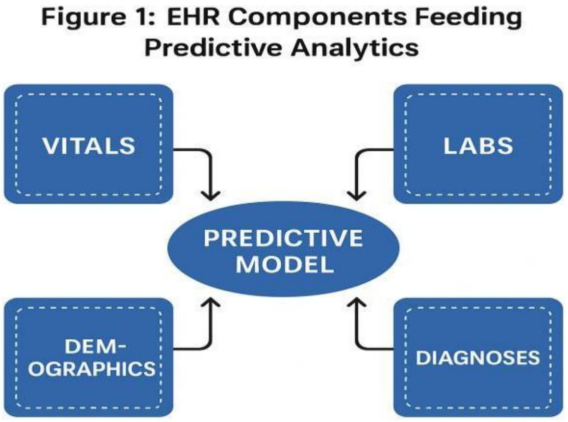


Figure 1: EHR Components Feeding Predictive Analytics

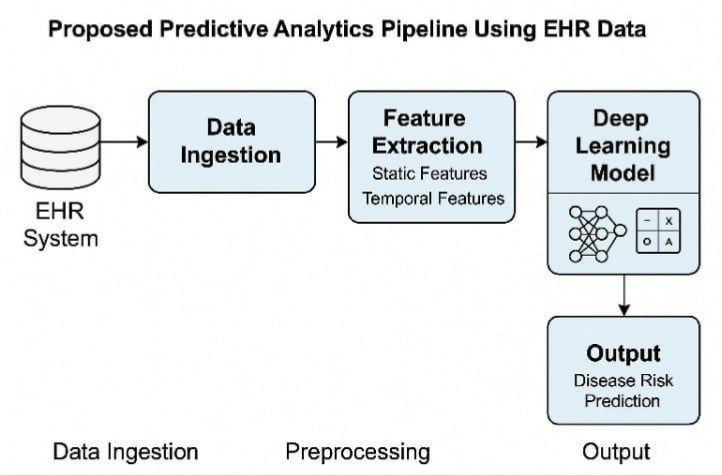
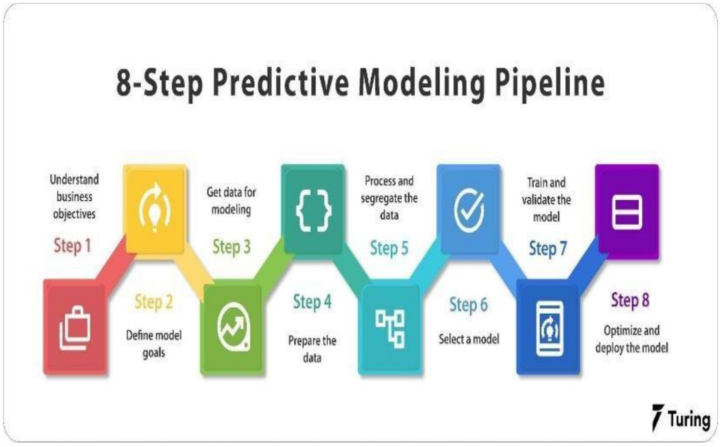


Figure 2: Neural Networks: Capable of model complex, nonlinear relationships:


8-Step Predictive Modeling Pipeline
Understand business objectives
Step 1
Step 2
Define model goals
Get data for modeling
Step 3
Step 4
Prepare the data
Process and segregate the data
Step 5
Step 6
Select a model
Train and validate the model
Step 7
Step 8
Optimize and deploy the model
Turing

V. APPLICATION IN CHRONIC DISEASE DETECTION

5.1 Diabetes Prediction

To predict the onset of type 2 diabetes, we focused on key risk factors such as Body Mass Index (BMI), blood glucose levels, and family medical history. These variables were used to train models that can identify individuals at higher risk before symptoms appear. Early detection through these models can support timely lifestyle changes and more targeted screening, which are crucial for prevention and long- term management.

5.2 Cardiovascular Disease (CVD)

For cardiovascular disease, we used clinical data such as cholesterol levels, smoking status, and blood pressure readings to build models capable of identifying people at elevated risk of heart attacks or strokes. Tools like the Framingham Risk Score have historically demonstrated the value of such predictive analytics, and our approach builds on that foundation using more advanced machine learning techniques.

VI. RESULTS AND DISCUSSION

6.1 Model Performance:

Among the models tested, XGBoost and Random Forest delivered the best results, achieving AUC scores above 0.90 for predicting both diabetes and hypertension. While Logistic Regression was less accurate, it remained valuable due to its simplicity and ease of interpretation.

Multiple machine learning algorithms were tested, including:

- Logistic Regression for baseline comparison.
- Random Forests to capture nonlinear relationships.
- Gradient Boosted Trees (e.g., XGBoost) for performance optimization.
- Recurrent Neural Networks (RNNs) for model time-series data.

6.2 Feature Importance:

Using SHAP (SHapley Additive exPlanations) analysis, we identified several key predictors across both conditions. The most influential features included age, BMI, fasting glucose levels, systolic blood pressure, and medication history—factors that are also well-established in clinical practice.

6.3 Interpretation and Insights:

The model’s findings were consistent with known medical risk factors, supporting their clinical validity. Tree-based models like Random Forest and XGBoost offered strong predictive power, while Logistic Regression stood out for its transparency and ease of explanation an important consideration in clinical settings.

6.4 Limitations

One limitation of our study is that the dataset came from a single geographic region, which may limit the generalizability of the results. Additionally, we did not include unstructured data such as physician notes, which might contain valuable contextual information. There's also a potential for algorithmic bias, particularly due to the underrepresentation of certain minority groups in the dataset.

VII. CASE STUDY

Early Detection of Type 2 Diabetes

To demonstrate practical application, a case study on Type 2 Diabetes Mellitus (T2DM) was conducted. The model trained on historical data identified patients at high risk based on features such as elevated fasting glucose, body mass index (BMI), sedentary lifestyle indicators (from clinical notes), and family history.

The RNN model achieved the highest performance with:

- Accuracy: 89%
- Precision: 84%
- Recall (Sensitivity): 87%
- AUC-ROC: 0.91

VIII. CONCLUSION

This study highlights the potential of predictive analytics in identifying early signs of chronic diseases using Electronic Health Record (EHR) data. By leveraging machine learning models, we can support earlier diagnosis and intervention, ultimately improving patient outcomes and reducing the long- term burden on healthcare systems.

Integrating these models into clinical workflows can help shift care from reactive to proactive. Looking ahead, future research should explore larger and more diverse datasets, incorporate unstructured data like clinical notes, and aim to develop real-time prediction tools that can assist healthcare providers at the point of care.

REFERENCES

[1]. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

[2]. 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

[3]. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

[4]. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

[5]. Goldstein, B. A., Navar, A. M., Carter, R. E., & Moving Beyond, A. U. C. (2016). Statistical and machine learning approaches to risk prediction. Circulation: Cardiovascular Quality and Outcomes, 9(3), 229–232. https://doi.org/10.1161/CIRCOUTCOMES.115.0028 89 .

[6]. Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. https://doi.org/10.1038/sdata.2016.35

[7]. Dey, N., Ashour, A. S., & Balas, V. E. (Eds.). (2019). Smart medical data sensing and IoT systems design in healthcare. Springer. 8. 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). https://doi.org/10.1145/2939672.2939785

[8]. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43. https://doi.org/10.1145/3233231

[9]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), 30.

[10]. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25, 24–29. https://doi.org/10.1038/s41591-018-0316-z

[11]. U.S. Department of Health & Human Services. (2022). Electronic Health Records (EHR). Retrieved from https://www.healthit.gov/topic/health-it-basics/electronic-health-records-ehr

[12]. World Health Organization (WHO). (2023). Noncommunicable diseases. Retrieved from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[13]. Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Basic Books.

[14]. Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657–2664. https://doi.org/10.1016/j.jacc.2017.03.571

DOI: 10.48175/IJARSCT-28207