National Conference on Information Technology & Digital Application (NCITSA-2025)
International Journal of Advanced Engineering, Management and Science (IJAEMS) Special Issue
https://dx.doi.org/10.22161/ijaems.ncitsa.2025.49 ISSN: 2454-1311

Healthcare predictive analytics using machine learning and deep learning techniques: a survey

Ms. Nimisha Swami[1] , Mr. Deepak Shankhala[2] , Vaishali Singh[3] , Sujata Kumari[4]

[1]Assistant Professor Department of Information Technology,Jaipur Engineering College & Research Centre, Jaipur
[1]Assistant Professor Department of ECE ,Jaipur Engineering College & Research Centre, Jaipur
[2]Student Department of Information Technology, Jaipur Engineering College and Research Centre, Jaipur
Email:vaishalisingh.it25@jecrc.ac.in
[3]Student Department of Information Technology, Jaipur Engineering College and Research Centre, Jaipur
Email : sujatakumari.it25@jecrc.ac.in

Abstract—The healthcare industry is rapidly evolving with the integration of advanced data-driven technologies. Predictive analytics, driven by Machine Learning (ML) and Deep Learning (DL) techniques, has emerged as a powerful tool to forecast health trends, diagnose diseases early, and optimize treatment plans. This survey aims to explore the landscape of predictive analytics in healthcare, highlighting various ML and DL models employed in real-world scenarios. We analyze their effectiveness in predicting diseases like diabetes, cardiovascular issues, and cancer. The study also discusses the strengths, limitations, and future scope of using such techniques in clinical decision-making and healthcare management.

Keywords—:Predictive Analytics,Healthcare Informatics, Machine Learning, Deep Learning, IoT in Healthcare, Clinical Data Mining.

I. INTRODUCTION

Healthcare systems worldwide are generating vast amounts of data daily—from electronic health records (EHRs) to imaging and genomic data. With the advancement in artificial intelligence, especially ML and DL, this data can be utilized to predict health outcomes, reduce hospital readmissions, and improve patient care quality. Predictive analytics refers to using historical and real-time data to make future predictions. In this paper, we survey various machine learning and deep learning approaches that have significantly contributed to predictive analytics in healthcare. We also investigate the challenges and ethical considerations of deploying such models in real clinical environments.

to the internet and the expansion of digital services,
cybercriminals have found new ways to exploit vulnerabilities. In India, the number of cyberattacks has escalated in recent years, with the country ranking as the third-largest source of malicious internet activity globally. Cybercrimes include a varietyofactivitiessuchasphishing,viruses, ransomware, data breaches, and denial-of-serviceattacks, all of which can cause significant harm to individuals, businesses, and even governmentinstitutions.

II. LITERATUREREVIEW

Several studies have demonstrated the potential of ML and DL in predicting health conditions. For instance, decision trees, support vector machines (SVM), and logistic regression have been widely used for structured data analysis. Deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have shown remarkable accuracy in image-based diagnostics and sequential health data. A 2020 study by Miotto et al. highlighted the use of DL in EHR analysis for predicting patient deterioration. Another prominent example is Google's DeepMind, which uses DL to detect over 50 eye diseases from retinal scans. However, challenges like data quality, interpretability, and model bias still persist.

https://dx.doi.org/10.22161/ijaems.ncitsa.2025.49 ISSN: 2454-1311

III. METHODOLOGY

3.1 Dataset

This study draws upon a variety of publicly available healthcare datasets commonly used in predictive analytics research. These include:

- MIMIC-III (Medical Information Mart for Intensive Care): A large, freely accessible critical care database comprising de-identified health data from ICU patients, including demographics, vital signs, lab tests, and diagnoses.
- UCI Heart Disease Dataset: A well-structured dataset used to predict heart disease presence based on clinical parameters.
- Diabetes 130-US hospitals dataset: Contains patient-level data of over 100,000 hospital admissions related to diabetes, including lab results, medications, and outcomes.

These datasets were selected for their diversity in data types (numerical, categorical, temporal) and relevance to real-world healthcare scenarios. They provide ample scope for comparing traditional and deep learning models for disease risk prediction, early diagnosis, and patient readmission forecasting.

3.2 Preprocessing

High-quality preprocessing is critical in the healthcare domain due to the heterogeneity and sensitivity of medical data. The following preprocessing steps were applied across datasets:

- Missing Value Handling: Imputation using median/mode values or model-based techniques such as KNN imputation.
- Normalization and Scaling: StandardScaler and MinMaxScaler were applied to normalize numerical features, ensuring uniformity across models.
- Encoding Categorical Data: One-hot encoding and label encoding were used for variables such as gender, diagnosis codes, and treatment type.
- Data Cleaning: Outliers were identified andaddressed, irrelevant fields were dropped, and timestamped data was aligned for temporalmodeling (especially for ICU datasets).
- Balancing the Dataset: In datasets with class imbalance (e.g., rare disease occurrence), techniques such as SMOTE (Synthetic Minority Oversampling Technique) were applied to avoid model bias.

3.4 Models Used

To evaluate the breadth and performance of AI approaches in healthcare analytics, both classical machine learning and state-of-the-art deep learning models were explored:

Machine Learning Models:

- Logistic Regression: Baseline binary classifier used for interpretability and benchmarking.
- Random Forest: Robust to overfitting and interpretable through feature importance, widely used in healthcare risk prediction.
- XGBoost: Gradient boosting decision trees optimized for performance on tabular clinical datasets.
- Support Vector Machines (SVM): Effective for small and medium-sized clinical datasets with high-dimensional feature spaces.

Deep Learning Models:

- LSTM (Long Short-Term Memory): Applied on time-series EHR data to capture temporal dependencies in ICU or diabetes monitoring scenarios.
- CNNs (Convolutional Neural Networks): Used for imaging datasets (e.g., X-rays, MRIs) and pattern recognition in sequential data.
- Autoencoders: Applied for unsupervised feature extraction and anomaly detection in healthcare logs.
- BERT and ClinicalBERT: Employed for clinical note classification, diagnosis prediction from discharge summaries, and medical named entity recognition.

Each model was evaluated across multiple datasets to assess its adaptability and accuracy in diverse healthcare applications. The choice of models ensures a comprehensive comparison between interpretable traditional approaches and high-performing deep neural networks.

IV. IMPLEMENTATION

The implementation phase of this study was conducted using Python due to its rich ecosystem of machine learning and deep learning libraries. The entire pipeline—ranging from data ingestion and preprocessing to model training and evaluation—was developed and tested using platforms such as **Jupyter Notebook**, **Google Colab**, and **VS Code**, with support for **GPU acceleration** where required.Traditional machine learning models were implemented using **Scikit-learn**, which provides efficient tools for classification, regression, and data preprocessing. These models included **Logistic Regression**, **Naive Bayes**, **Random Forest**, and **Support Vector Machines (SVM)**. For each, the dataset was split into **80% training** and **20% testing** sets, and **stratified sampling** was used where applicable to preserve class distributions.

https://dx.doi.org/10.22161/ijaems.ncitsa.2025.49 ISSN: 2454-1311

III. RESULTSANDDISCUSSION

The performance of the implemented models was evaluated across multiple healthcare prediction tasks, including disease classification, patient readmission prediction, and analysis of clinical notes. Evaluation metrics included **accuracy**, **precision**, **recall**, **F1-score**, and **ROC-AUC**, ensuring a comprehensive assessment of both model effectiveness and reliability.

Among the traditional machine learning models, **Random Forest** and **XGBoost** consistently delivered strong results, particularly in datasets with structured clinical features like lab results and demographic data. Random Forest achieved an average accuracy of **84–87%**, with high recall, making it suitable for identifying at-risk patients in imbalanced datasets. **Logistic Regression**, while slightly less accurate (~78–80%), offered better interpretability, which is often critical in medical decision-making contexts.

In summary, the results highlight that **no single model is universally best**, and the choice of algorithm should be tailored to the nature of the data and the specific clinical application. A hybrid or ensemble approach may provide the most practical balance between accuracy, explainability, and resource efficiency in real-world healthcare systems.

IV. CONCLUSION

This survey emphasizes the growing importance of predictive analytics in healthcare. ML and DL techniques have demonstrated great potential in improving diagnostic accuracy, treatment plans, and resource allocation. Future research should focus on improving model interpretability, ensuring ethical AI use, and creating standardized healthcare datasets. Collaboration between data scientists and healthcare professionals is vital for building trust and utility in real-world applications.

The implications of this study are significant for IoT infrastructure providers, system administrators, and AI researchers. Integrating these models into real-time monitoring systems could enhance the detection of anomalies, faults, or potential cyber threats across IoT networks. However, the trade-off between model complexity and performance must be carefully considered, especially in resource-constrained environments like edge devices.

Futureresearchcanexploremulti-modaldetectionmethods that combine text, images, and metadata. Real-time analysis, multilingual datasets, and ethical considerations such as model transparency and fairness should also be prioritized.

ACKNOWLEDGEMENTS

We would like to thank our research supervisor for their continuous guidance and valuable feedback throughout this project. We also extend our gratitude to the developers of publicly available datasets and open-source ML/DL tools that made the implementation part of this study possible.

REFERENCES

1. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236–1246.

2. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.

3. Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.

4. Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

5. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317–1318.

https://dx.doi.org/10.22161/ijaems.ncitsa.2025.49 ISSN: 2454-1311