iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.84094

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Machine Learning Approaches for Early Disease Prediction Using Electronic Health Records (EHR)

Swapan Banerjee
Faculty of Management & Commerce, (Hospital and Health Management), Poornima University, Jaipur, India
**Corresponding Email: sbanerjee.researcher.21@gmail.com*
https://orcid.org/0000-0001-5781-5436

Abstract: ***Machine learning is highly versatile in health care, especially for early disease prediction, leveraging electronic health records (EHRs). EHRs provide a wealth of longitudinal clinical, laboratory, diagnosis, and medication data, vital signs, and clinical notes. Heterogeneous data sources offer opportunities to detect disease patterns before obvious clinical deterioration. In recent years, machine learning models have been increasingly used to transform raw EHR data to actionable predictions, including for diabetes, sepsis, chronic kidney disease, cardiovascular disease, and cancer. In this article, the main machine learning methods for early disease prediction from EHR data, including logistic regression, random forests, gradient boosting machines, support vector machines, recurrent neural networks, transformers, and multimodal deep learning architectures, are reviewed. The end-to-end pipeline, spanning data extraction and preprocessing, feature engineering, model training and evaluation, interpretability, and clinical deployment, is discussed. It also points out several important pitfalls, including class imbalance, temporal irregularity, label leakage, reduced generalizability across institutions, coding variability, and missing data. Representative model families, strengths, and challenges are summarized in a table. A conceptual figure of the workflow of early prediction using EHR. The article ultimately argues that, in addition to a high predictive performance, calibration, interpretability, fairness, and workflow integration are required for successful implementation. Future research should focus on prospective validation, federated learning, and clinically informed evaluation to facilitate safe and scalable translation to practice.***

Keywords: ***Electronic health records, Machine learning, early Disease prediction, Clinical informatics, Predictive modeling, Healthcare AI.***

I. INTRODUCTION

Early disease prediction aims to predict disease in patients even before they develop it or before it causes any damage. Electronic health records (EHRs) in today's healthcare systems deliver a constantly expanding collection of structured and unstructured patient data. Unlike traditional clinical studies, which capture snapshots of patient status, EHRs capture a record of patient history, diagnoses, procedures, lab values, medications, vital signs, radiology reports, clinician notes, and more. This longitudinal richness makes EHRs highly attractive for longitudinal pattern recognition and risk estimation in machine learning applications [1, 3].

There has been significant progress in applying machine learning to predictive modeling in the healthcare sector over the past decade. Despite their simplicity and comprehensibility, conventional statistical models such as logistic regression remain popular. However, they can't handle either nonlinear interactions or complex temporal dependencies. Ensemble methods, such as random forests and gradient boosting, are now common alternatives that can be highly effective with tabular EHR data [11]. With the field's development, deep learning models such as Recurrent neural networks (RNNs) and Transformers, which learn latent representations of sequential data and text, have become increasingly popular [1, 2]. Useful when revisiting the model allows for considering changes in patient state over time, and/or when there are clues to prediction that are not readily captured in coded data in clinical notes [3].

Models are performing well, but it is still difficult to apply them in the clinical scenario. Predictive systems need to work in the context of missingness, irregular sampling, coding changes, changes in case mix, and changes in care practices. A successful model in retrospective data might not perform as well in prospective data, and/or might be inadequate when applied to another hospital [8]. Many studies also focus solely on metrics of discrimination and fail to account for calibration, interpretability, fairness, and actionability [14]. These operational constraints must be taken into account when moving predictive algorithms to the clinic, as they can influence clinical decisions and affect patient safety [9, 15].

A brief overview of machine learning techniques for early disease prediction is provided, along with a discussion of clinical and technical workflows for robust modeling with EHRs. It will provide insights into data sources, feature engineering, model families, evaluation methods, and interpretability tools. Practical deployment issues are also discussed, and directions for future research to enhance translational impact are suggested.

II. METHODS

A. *Pre-processing and Feature Engineering*

Prediction studies are usually based on EHRs, starting with the cohort definition. The target population may include all adult inpatients, ED visitors, outpatients in general practice, or sub-populations of patients with a specific disease. The researchers select an index time point, a prediction horizon, and an outcome definition. These may include predicting the onset of sepsis within 6, 12, or 24 hours of observation, predicting the onset of diabetes within 1 year, or predicting readmission within 30 days [6,7]. The construction of cohorts must be carefully done to prevent selection bias and "leakage" of labels [8].

Populated data comprises demographics, diagnosis codes, procedure codes, medication orders, lab results, vital signs, hospitalization history, and unstructured clinical notes [7]. Usually, structured data is converted into machine-readable features, and text is processed via natural language processing (NLP) pipelines or new language models [3]. Even though streams of events can be modeled using timestamps, they still require harmonization because measurements can be sampled at different time points [16].

The most time-consuming step of modeling EHRs is pre-processing. Tests not requested, tests not captured, or tests lost during system integration are all reasons for missing data. There are several methods for imputing missing values, such as the mean or median approach, model-based methods, forward filling, and clinically informed rules [16]. But the missingness itself might be predictive information. For instance, low disease suspicion may lead to a lack of testing, whereas the clinician's concerns may drive frequent testing. Therefore, modern models often use indicators of missingness in addition to imputed values [16].

Several feature engineering strategies differ depending on the type of model. A typical machine learning model will require the data to be summarized over time, such as by minimum, maximum, mean, trend, or by counting or measuring the variance over a specified period. Diagnosis codes can be binary flags, counts, or embeddings [2], active prescriptions, or cumulative medication effects. Temporal models can be used to preserve the raw sequence and directly feed it to models such as recurrent neural networks or transformers [1, 12].

B. *Machine learning models*

Logistic regression is commonly used as a baseline because it is easy to interpret and has well-understood behavior. Applies when the number of predictors is moderate, and relationships are roughly linear on the log-odds scale. Random forests and gradient boosting methods (XGBoost, LightGBM) are effective for handling high-dimensional tabular data and nonlinear interactions [11]. These methods are flexible with respect to variable types and, in many cases, offer good performance out of the box [8].

In some high-dimensional environments, Support Vector Machine (SVMs) can be effective but may not scale well in very large ones. As the use of sequential, multimodal, or very large data sets has increased, deep learning-based approaches are increasingly adopted [1]. Recurrent neural networks process event sequences as a series of events, one at a time, and model temporal dependencies [16]. Long short-term memory networks (LSTMs) have been widely used in the field of clinical time-series, where they can learn and retain information over long time scales [6]. Transformer-based models have gained popularity for large-scale EHR representation learning. With the assistance of an attention mechanism, the importance of previous events can be taken into account. A multimodal or hybrid approach attempts to merge traditional data, clinical notes, and image data [4, 5]. These systems can be independent encoders for each modality, then combine the learned representations for classification. Such a strategy can improve predictive performance if the disease signatures are distributed across heterogeneous record types [3].

C. *Model Evaluation*

For early prediction evaluation, more than overall accuracy is required. The majority of disease results are rare, so the accuracy is questionable [8]. The most commonly used discrimination measures are the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). Sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and calibration (Brier score and calibration slope) provide greater insight.

Temporal validation (when possible) is better than random splits, and it is more representative of future deployment [8]. Independent validation across hospitals/health systems is even more important for assessing transportability [9]. Further, the net clinical benefit of a model can be assessed at different action thresholds using decision curve analysis.

D. Interpretability and Fairness

A key factor in the clinical adoption is the interpretability. Feature importance, Shapley Additive Explanations (SHAP) values, partial dependence plots, attention visualization, and counterfactual explanations are examples of techniques that can be used to explain to the clinical user why a model made the decision it did and assigned a certain risk score [12, 13]. Interpretability, however, should not be interpreted as correctness - explanations should be checked to make sure they are faithful to the underlying model [13]. It's also crucial that it's fair. EHR data may also indicate structural inequities in access to care, diagnosis, and treatment. There may be unknown differences between models related to race, sex, age, socioeconomic factors, language, or insurance type [14]. The behavior of subgroups, calibration, and threshold behavior should be included in a fairness-aware evaluation. After identifying differences, some mitigation options include balanced sampling, reweighting, threshold adjustment, or causal analysis of the source(s) of bias [14].

III. RESULTS

Many works reviewed in this paper have demonstrated that, in many early disease prediction tasks, machine learning models are more successful than simple rule-based models when adequate data is available. For structured EHR data, tree-based ensemble methods have often been used to model nonlinear relationships and interactions among different lab trajectories, medication history, and comorbidities, providing good discrimination [11]. Gradient boosting is effective for predicting mortality, readmission, sepsis risk, and chronic disease [6, 7] many times. Temporal dynamics or unstructured text prediction tasks have particularly benefited from deep learning techniques [1, 3]. Recurrent and attention-based architectures can model symptom trajectories, repeated measurements, and changes in symptom profiles [12,16]. Transformer models have proven especially attractive in large longitudinal datasets, as they can model long-range dependencies and are easily adapted to the powerful computational resources available today. But success in the retrospective performance does not guarantee success in clinical use. For a model to be properly validated, it needs to be evaluated on external datasets, and substantial discrimination drops in a few studies may suggest that the model is overfitting the institution's coding practices or the patient mix [8]. Calibration can be lower than discrimination, meaning the predicted probability can be less representative of the actual event rates [9, 15]. That's a problem for early disease prediction, when clinicians seek a reliable prediction to inform intervention decisions.

The following table summarizes common machine learning model families used in EHR-based early disease prediction, together with their strengths and limitations.

Table 1. Summary of common machine learning approaches for early disease prediction using EHR data.

Model family	Typical data type	Main strengths	Common limitations
Logistic regression	Structured tabular features	Highly interpretable, fast, and easy to calibrate	Limited nonlinear modeling is weaker with complex interactions
Random forest	Structured tabular features	Robust, handles mixed variables, captures nonlinearities	Less interpretable than regression, it can be computationally heavy
Gradient boosting machines	Structured tabular features	Strong predictive power, handles missingness well	Requires tuning; explanations may be less transparent
Support vector machines	High-dimensional features	Effective in some sparse settings	Scalability issues, limited probabilistic interpretability
Recurrent neural networks	Sequential EHR trajectories	Models temporal dependencies and event order	Harder to interpret, may require large datasets
Transformers	Longitudinal sequences and multimodal inputs	Captures long-range dependencies, scalable representation learning	Data-intensive, high computational cost, and explanation challenges

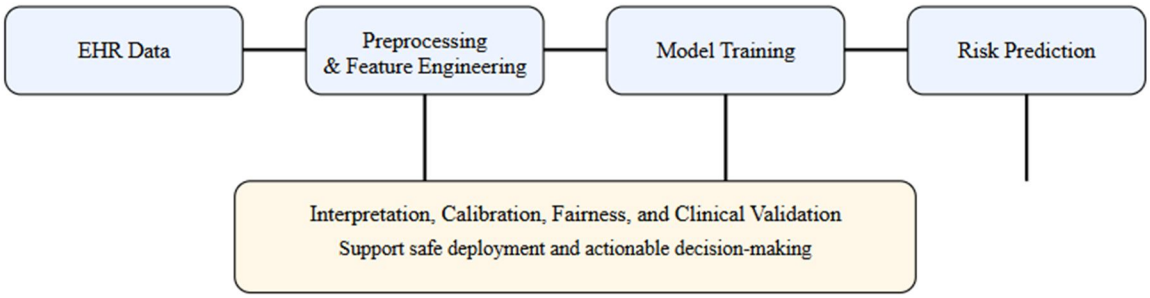


Figure 1. Conceptual workflow for early disease prediction using electronic health records.

Figure 1 shows the electronic health record data and their processing through model training and risk prediction, with interpretation of fairness and clinical validation.

IV. DISCUSSION

The results show that disease prediction can be significantly improved using machine learning. If the models are developed using clinical data [1, 8], then early disease prediction can be achieved. EHRs are not lab data; they are products of care activities, billing, and workflow problems. Irregularity and sparsity in the data are important, as is confounding or changing documentation behavior, and it's important to be able to handle these with successful models. Often, the best methods will be those with a strong statistical basis and flexible representation learning [3]. One of the main compromises mentioned in the literature is between the predictive performance and the interpretability [13]. Models that are explainable, auditable, and can be correlated with Medical reasoning [12] increase the trustworthiness. But in complex data structures, overly crude models may be lacking. A step-up approach might be best, beginning with the simplest, baseline model, and moving to more sophisticated models, and then using the simplest model that is robust and suitable for the clinical needs.

Another key concern is generalizability. The models can be trained at a single institution, where they learn patterns specific to that institution, such as local coding patterns, hospital ordering patterns, or referral pathways [8]. So, it's crucial to get outside perspectives and tests over time. In some cases, techniques such as domain adaptation, transfer learning, or federated learning can improve model performance by training without sharing patient-level information [9]. Unstructured text is also increasingly used. However, structured fields are not as rich as clinical notes in describing symptoms, clinician impressions, changes in the patient's state, and so on [3]. There are new risks, such as hallucination, prompt sensitivity, and opaque reasoning that come with this hidden signal, which could be discovered and revealed by NLP techniques and large language models. However, when using text-based components, care should be taken to benchmark and monitor them to a high level of safety in safety-critical applications [15].

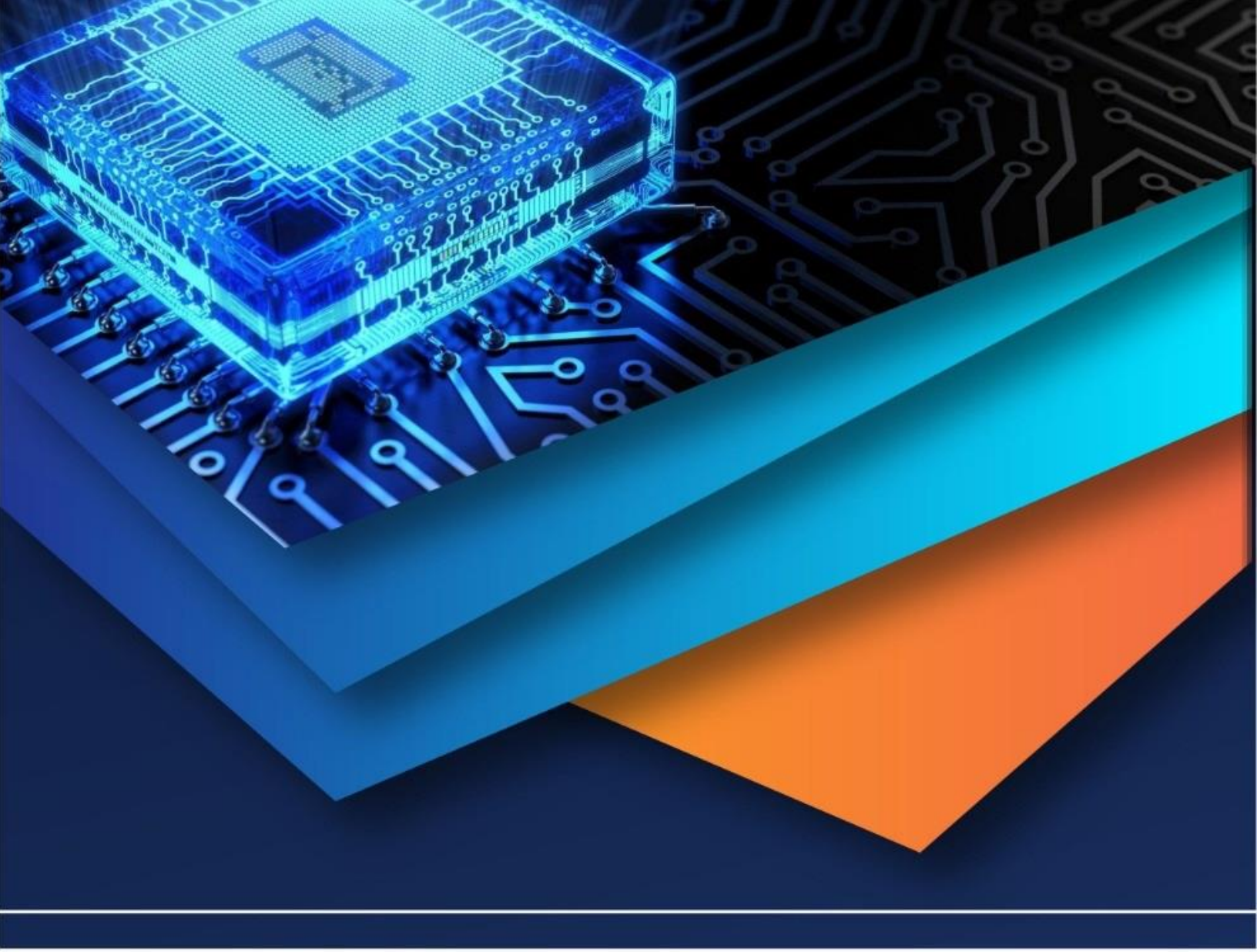
The last and most challenging stage is operational deployment. If a model can't be used in the clinical workflow when needed, in the right user interface, and in the right way to act upon it, then it's no good [9]. Even if a system is technically sound, alert fatigue, unclear ownership, and poor usability can be issues. Thus, implementation science should be considered from the outset when developing a model.

V. CONCLUSION

The potential of machine learning methods for predicting disease in its early stages from EHRs is significant [1, 8]. These will be able to predict risk earlier than any rule-based system and will help with proactive care based on longitudinal clinical information. This can only be achieved, however, if careful attention is paid to data quality [13, 14], temporal validation, interpretability, fairness, and clinical utility. The architectures can be categorized into tabular, temporal, and multimodal, which are suitable for EHR prediction problems, temporal domain problems, and multimodal problems, respectively [11,12]. Prospective studies, multicenter validation, standardized reporting, and improved connections between model outputs and actionable interventions will be crucial to future progress. The focus on achieving a maximum AUROC is insufficient; reliable, explainable, equitable, and deployable models should be pursued. These ideas could be used to leverage machine learning to deliver practical solutions for early diagnosis and treatment, ultimately improving patient outcomes [16].

REFERENCES

[1] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*. 2018;1:18.
[2] Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*. 2016;6:26094.
[3] Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances on deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics*. 2018;22(5):1589-1604.
[4] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*. 2018;19(6):1236-1246.
[5] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.
[6] Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Scientific Data*. 2019;6:96.
[7] Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*. 2016;3:160035.
[8] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*. 2017;24(1):198-208.
[9] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. 2019;25(1):44-56.
[10] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nature Medicine*. 2019;25(1):24-29.
[11] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. *KDD*. 2016:785-794.
[12] Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in Neural Information Processing Systems*. 2016;29:3504-3512.
[13] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*. 2017;30:4765-4774.
[14] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453.
[15] Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New England Journal of Medicine*. 2019;380(14):1347-1358.
[16] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*. 2018;8:6085.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY