International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 15 (2023) pp. 595-604
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Received: 03 March 2023; Accepted: 02 August, 2023; Published: 02 September, 2023

# Machine Learning Techniques for Electronic Health Records: Review of a Decade of Research

**Vibhuti Sharma[1], Anu Bajaj[1] and Ajith Abraham[2]**

[1] Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology, Patiala, India
*er.anubajaj@email.com*

[2] School of Computer Science Engineering and Technology,
Bennett University, Greater Noida, India

***Abstract*: Advancement in Machine Learning (ML) has opened new gateways for transforming the healthcare sector. This paper explores the integration of ML techniques within the healthcare domain. It also explains tools and techniques used for preprocessing and feature extraction. The review highlights various datasets, evaluation metrics, and ML techniques used by the researchers in these three domains - ICD (International Classification of Diseases) -9/10 coding, Mortality prediction, and disease prediction. Further, for disease prediction, we have reviewed four major areas: sclerosis prediction, cardiovascular disease prediction, cancer prediction, and kidney diseases. In alignment with our study, it is evident that the prevailing trend in the healthcare sector is shifting towards the adoption and integration of advanced deep learning methodologies.**

***Keywords*:** Electronic Health Records, Mortality Prediction, Deep learning, Disease Prediction, ICD-9/ICD-10 coding

## I. Introduction

The main aim of implementing Electronic Health Records (EHRs) is to enhance the working of the healthcare sector by properly managing the patient's information [1]. Researchers are working to extract meaningful information from EHR. The volume of digital data has increased rapidly during the previous 10 years [2]. Digitization of hospital records has enabled Electronic health records (EHRs) to be accessible to researchers for research purposes [3]. Electronic health records (EHRs) are a considerably essential data source for biomedical research. In recent years, EHR has supported disease genomics discovery, enabled rapid and more inclusive clinical trial recruitment, and facilitated epidemiological studies of understudied and emerging diseases [4]. EHR includes patient medical history, diagnosis, lab test results, and clinical notes that can improve healthcare outcomes, medical research and enhance the decision-making process [5].

However, EHRs can pose significant challenges like 1) equivalent length of visits, 2) equal number of observations per patient [6], 3) vast volume and complexity, and 4) Security and Privacy of patient information [7] which makes it challenging to manage and extract helpful insights. Researchers have been working on Machine learning (ML) models to overcome these issues. ML models have shown significant potential in healthcare applications such as early disease detection, healthcare automation, predictive modeling, etc.

This review aims to provide a comprehensive analysis of various ML techniques, such as support vector machine (SVM), Convolutional Neural Networks (CNN), transfer learning models, Hierarchical Multi-Label Classification with Partial Label Attention (HMC-PLA) that have been applied to EHR data in various application domains like ICD (International Classification of Diseases)-9 code assignment, disease prediction, and mortality prediction.

The contribution of this study is to get insights about the underlying trends of ML approaches in the healthcare field. By delving into these trends, the study aims to shed light on how ML approaches can be harnessed to improve results in the healthcare industry and get meaningful insights from EHRs. This study also tries to understand evaluation metrics used to validate the results of these approaches. The following research questions are framed to answer these objectives for ML approaches in EHRs.

- **RQ1**: What machine learning methods are employed for Electronic Health Records (EHR) analysis?
- **RQ2:** How EHR data is preprocessed for ML models?
- **RQ3**: Which datasets are utilized for validation?
- **RQ4**: Which evaluation metrics are employed to quantify the outcomes?
- **RQ5**: What are the potential future directions in applying ML approaches for EHR?

The paper's organization is as follows: Section 2 discusses the methodology of paper selection and Section 3 briefs about the machine learning algorithms. The ML architecture for EHR is presented in Section 4. Section 5 describes the existing literature and Section 6 gives the findings and future directions with the conclusion in Section 7.

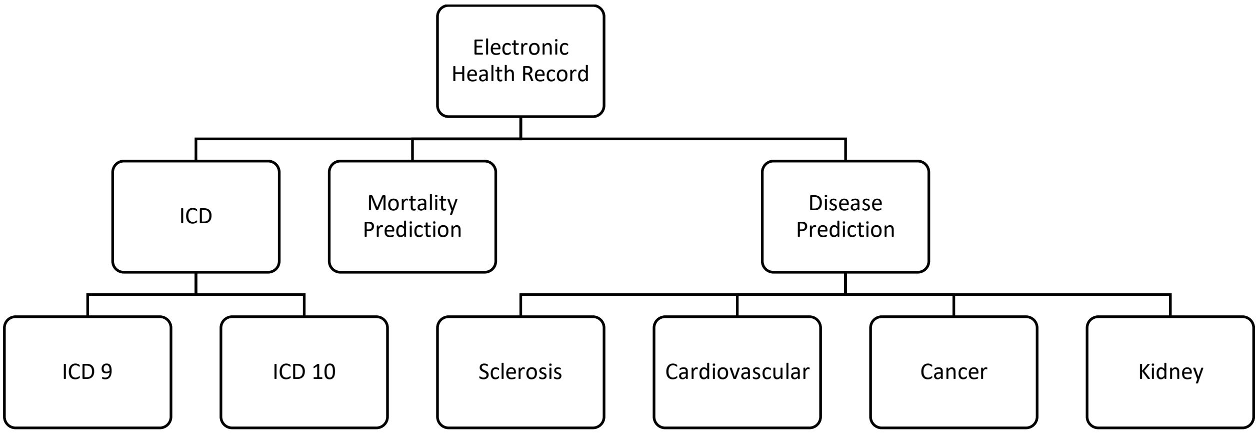

**Figure 1**. Classification of EHR data

***Table 1.*** Machine Learning Algorithms

| Models | Methods | Description |
|---|---|---|
| **Baseline Models** | Linear Regression(LR) | A simple linear model that fits a linear relationship between features and targets. |
| | Support Vector Machine (SVM) | ML algorithm used for linear or nonlinear regression classification and outlier detection tasks. It works by finding an optimal hyperplane that distinguishes various classes or makes predictions for continuous outputs. |
| | Random Forest(RF) | Ensemble learning algorithm used for both classification and regression tasks. |
| | Decision Tree (DT) | Non-linear model that makes decisions based on feature thresholds. It is a tree-like structure where internal nodes represent decisions based upon a specific feature and leaf nodes represent predicted outcomes. |
| | Naive Bayes | probabilistic classifier based on the Bayes' theorem that assumes features are conditionally independent given the class. |
| | k-Nearest Neighbors (k-NN) | It finds ‘k’ closest data point (neighbours) to a given input and make predictions based on the majority class. |
| **Deep Learning** | Convolutional Neural Networks (CNNs): | It consists of several<br>convolutional layers and subsampling layers that automatically learns hierarchical features from images, making them well-suited for tasks like image classification, object detection, and image segmentation. |
| | Recurrent Neural Networks (RNNs)) | It is designed for sequence data, such as time series, text, and speech. |

## II. Methodology and Paper Selection

This review focuses on the application of ML in EHR. We have identified three research domains in which ML techniques have been applied, such as ICD-9 code assignment, disease diagnosis, and mortality prediction as shown in Figure 1. In these three domains, we identified which dataset is used, data preprocessing techniques, feature extraction methods, algorithms used, and evaluation measures taken to evaluate the model. The literature was found by a systematic search in ScienceDirect, IEEE Xplore, Springer, MDPI, ACM, and google scholar using keywords: "EHR," "Automatic ICD-9 coding ", "Deep learning," "Transfer learning," "MIMIC-III", "Cardiovascular", "Mortality Prediction", and "sclerosis". Furthermore, the work is categorized into three categories:1) Automatic ICD-9 coding, 2) mortality prediction, and 3) Disease Prediction.

## III. Machine Learning (ML) Algorithms

Within the field of ML, there are two fundamental paradigms unsupervised learning and supervised learning. Supervised learning is training ML model on a labeled dataset where each data point is associated with the target and maps input features with its target output and generalizes this mapping to make accurate predictions on unseen data. Unsupervised learning, on the other hand, uses unlabeled dataset for training purposes. The objective is to find patterns within the data without explicit guidance. Our survey mostly focuses on supervised learning algorithms. These are further classified into two models: Baseline models and deep learning models as shown in Table 1. Baseline models are naive models that serve as reference points to determine whether complex models are actually improving performance or not. Deep learning models are specifically designed to handle more complex patterns. Various studies have highlighted that deep learning models are showing promising results for automation and prediction modelling.

## IV. ML Architecture for EHR

Our methodology involves following steps:

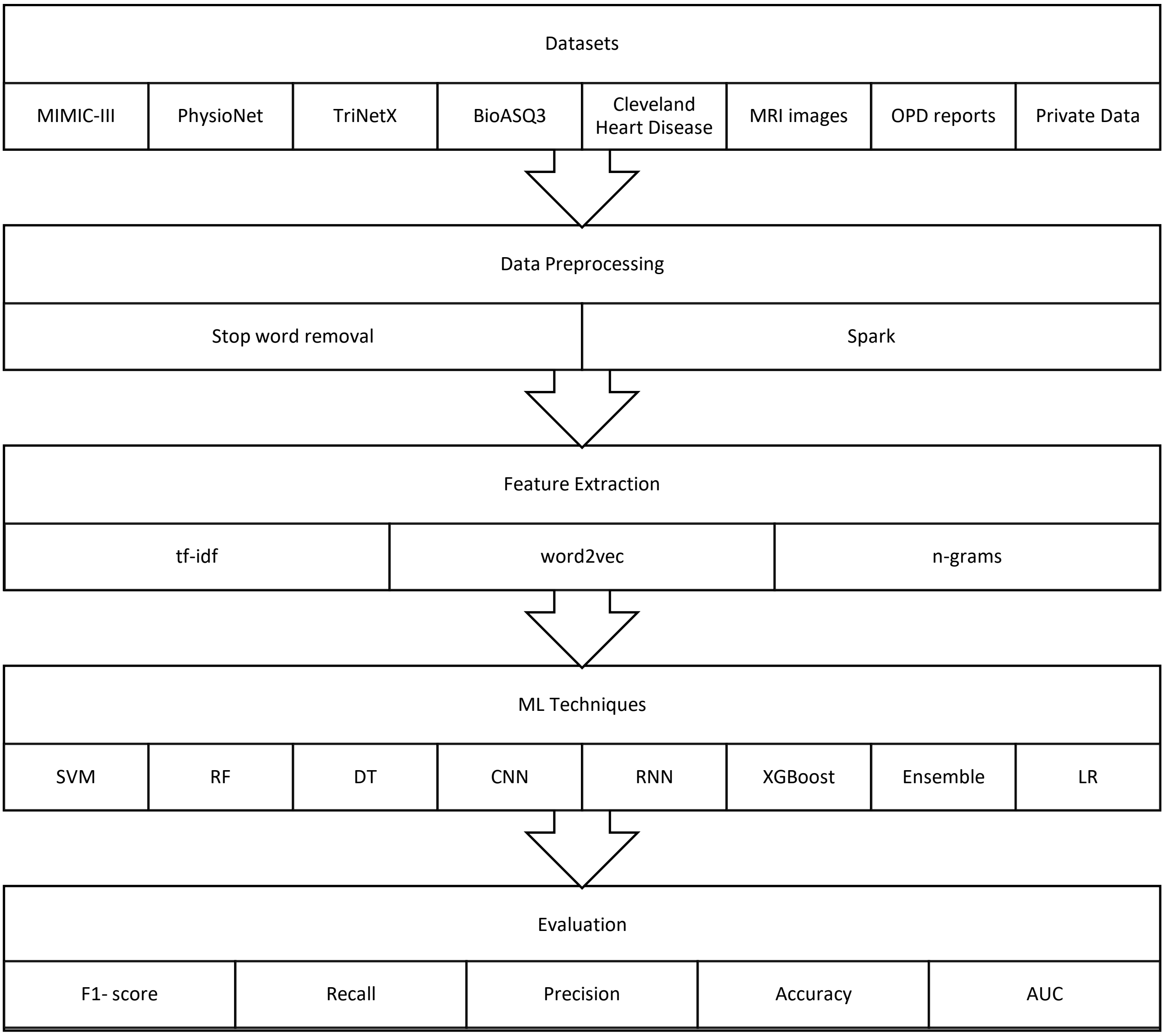

**Figure 2.** ML architecture for EHR

### A. *Dataset Selection*

Dataset selection is the most foundational step in the realm of data analysis. Due to the complex nature of EHR it is a very crucial step to find a suitable dataset for our research that should align with our research problem. Datasets mentioned in these studies are textual data and image data like MIMIC-3 is textual data and for disease prediction researchers used image data such as MRI reports.

### B. *Data Preprocessing*

It involves a series of steps that is applied to the raw data before it is used for model training. It impacts the accuracy, robustness, and effectiveness of data analysis and to get useful insights from data. Some studies highlighted that preprocessing is done by removing stop words, punctuation, and converting all words to lowercase. Stop words like “and”, “in”, “is”, “the” etc. are removed because they usually don't carry a significant meaning. By removing these words, the model can focus more upon meaningful words. Punctuation marks like comma, semicolon, etc. serve for grammatical purposes only by removing them it simplifies the text. MIMIC-III data was preprocessed using Spark [9] which provides a rich set of libraries and tools for preprocessing tasks. For image data there are various filters that can be used for preprocessing images that can help to remove noise from the data. For feature extraction image segmentation is used.

### C. *Feature Selection*

It is a strategic process that identifies relevant features from the dataset. Some approaches mentioned in the study are Term frequency - Inverse Document Frequency (tfidf) that aims to evaluate the importance of words to a given corpus. Word2vec takes input as tokenized text corpus and outputs word vectors. N-grams capture relationships between words and sequence

### D. *Model Training*

Model training is a pivotal step in ML. For that we apply various algorithms and techniques to find patterns within the dataset that gives the desired output. Various Ml approaches were also suggested for training purposes that we have discussed in the subsequent sections.

### E. *Dataset Selection*

Model evaluation provides quantitative measures to know how well our proposed model performs. Some widely used datasets, preprocessing and feature extraction techniques, model training algorithms and evaluation metrics are mentioned in Figure 2.

## V. ML Approaches for EHR

Our methodology involves following steps:
Various researchers have used ML approaches to solve various issues and challenges in EHR. A summary of their works is presented below:

### *A. Automatic ICD-9 /ICD-10 coding*

The Ninth Revision of International Classification of Diseases (ICD-9) codes assigns unique code to diseases which are used in EHR for patient disease diagnosis, mortality rate statistical analysis, and medical reimbursement for billing mechanisms [8]. Usually ICD-9\ICD-10 coding is done manually, which is a labor-intensive process. Researchers explored ML to automate mapping of ICD-9\ICD-10 codes with clinical notes. Various ML methods were applied such as k-nearest neighbour (kNN) and Naive Bayes neighbors to assign ICD-9 codes automatically [5]. To improve accuracy and efficiency of these algorithms, researchers started exploring various ML tools and techniques in this field.

Ferrão et al. [10] introduced an adaptive data processing approach using structured EHR. The process involves transforming unprocessed clinical data into feature sets that SVM classifiers use as input. SVM classifiers are trained to obtain predictions for assigning codes to each episode. Chen et al. [11] proposed an improved LCS algorithm that took word sequence and semantic similarity into account and derived a new formula of similarity measure by enhancing the weight of LCS.

The past decade has witnessed a significant impact of Deep learning algorithms on EHR [12]. Numerous research has employed CNN algorithm to tackle various research challenges. Zeng et al. [8] introduced a deep transfer learning framework for automatic ICD-9 coding that utilizes CNN with multi-scale convolutional layers for feature extraction and a shared layer of transfer learning. Huang et al. [5] compared deep learning approaches like CNNs, RNNs, and Hierarchical Attention Networks (HANs) with baseline models LR, RF, and Feed-forward Neural Networks (FNN). The study found that RNNs, specifically Gated Recurrent Units (GRUs), best predicted the top 10 ICD-9 codes.

Moons et al. [13] proposed a novel model called Hierarchical Multi-Label Classification with Partial Label Attention (HMC-PLA) to handle the imbalance and sparsity of the ICD code distribution. It also compared the proposed algorithm CNNs, RNNs, and HANs. HMC-PLA model achieves F1-score of 0.471 on the Medical Information Mart for Intensive Care III (MIMIC III) dataset and an F1-score of 0.547 on the CodiEsp corpus, which are higher than the F1-scores conducted by the other models. Li et al. [14] presented a framework called DeepLabeler that combined CNN with the 'Document to Vector' method for extracting and encoding local and global features. It compared results of the proposed model with SVM and flat SVM and concluded that DeepLabeler outperforms SVM by 14%.

Most of these works are focused on supervised framework which uses apriori labeled clinical notes, but in many instances, there is not enough labeled text which necessitates unsupervised code assignment techniques [15]. This approach focuses on finding similarity between clinical data like a diagnosis with ICD codes list: Word Mover's Similarity (WMS). Some crucial papers in this research domain are tabulated in Table 2. Observations from the table suggest that most of the work is done on MIMIC-II/III dataset and the SVM and CNN are the popular candidates of ML techniques used for ICD-9/10 coding. Mostly the performance is evaluated by calculating the F1 score.

### *B. Mortality Prediction*

Over the last decade researchers have been working on prediction modeling by utilizing the scope of ML algorithms. In the Intensive Care Unit (ICU), accurate estimation of mortality risk and patient deterioration is critical for early medical intervention and patient care. Also, precise risk assessment assists in allocating limited ICU resources [17]. Ye et al [18] showed that ML models are promising for healthcare workers for predicting the mortality risk of critically ill patients.

***Table 2.*** ICD-9/10 code assignment using ML Techniques

| Author | Dataset | Proposed technique | Algorithms compared | Performance Metric |
|---|---|---|---|---|
| Ferrão et al. [10] | Private | SVM | - | F1 Measure, Recall, Precision |
| Zeng et al. [8] | BioAS3 MIMIC-III | Deep transfer learning | flat-SVM, hierarchy-based SVM | Micro-average F-measure, Micro- -average- Precision, Micro-average Recall |
| Huang et al. [5] | MIMIC-III | RNN-GRU | CNN RNN-GRU, LSTM LR, RF, FNN | Precision, Recall, F1-score, Accuracy |
| Moons et al. [13] | MIMIC-III, CodiEsp corpus | HMC-PLA | CNN, GRU, DR-CAML, MVC-LDA, MVC-RLDA | Micro F1-score, Micro AUC |
| Singaravelan et al. [9] | Private | CNN | SVM, RF, CNN | Precision, Recall, F1-score, Accuracy |
| Che et al. [11] | gold standard | LCS algorithm | - | F-score |
| Liu et al. [16] | MIMIC-II, MIMIC-III | Deep-Labeler CNN | SVM | F-measure |

Hoogendoorn et al. [19] compared two approaches one is extracting high-level features from EHR and utilizing them from predictive modeling, and the other is a patient similarity metric. It uses K-Nearest Neighbor (KNN) for Patient Similarity and logistic regression approach for predictive modeling. Study shows that results obtained from predictive modeling are more accurate than similarity metrics with AUC of 0.84. The study also highlighted that mortality can be predicted within a median of 72 hours. Abedi et al. [20] used EHR to predict short and long-term mortality following an acute ischemic stroke. They determined that the extreme gradient boosting (XGB) model had the best Area Under the ROC curve (AUROC) of 0.82 for the 1-month prediction window.

Researchers in recent years started exploring neural networks and deep learning techniques. Nascimento et al. [21] used LSTM (Long-Short Term Memory) model. LSTMs are ideal to predict temporal sequences that can be useful for predicting mortality. Kumar et al. [22] proposed a self-explaining NN for ICU mortality prediction. Lei et al. [23] used a RNN-based denoising autoencoder (RNN-DAE) for encoding EHR and utilize it for mortality prediction. They compared it with k-means, Principal Component Analysis (PCA), SDAs and Gaussian Mixture Model (GMM) [23].

Jun et al. [24] suggested that variational RNN (VRNN) can handle missing value imputation, representation learning, and in-hospital mortality prediction in a single stream. Tann et al. [25] highlighted that mortality prediction is also important for evaluating early treatments and detecting high-risk patients that eventually increase the performance of the healthcare sector. They proposed the Uncertainty-Aware Convolutional Recurrent Neural Network (UA-CRNN) method. Ye et al. [18] suggested that the knowledge-guided CNN model was effective for mortality prediction with AUC-0.97. Table 3 presents the summary of the ML models used for mortality prediction. It can be concluded from Table 3 that according to current trends for mortality prediction, researchers are relying on neural networks and deep learning techniques and suggest that RNN and its variants have shown promising results.

### C. Disease Prediction

EHR provides patient data that helps to recognize the characteristics of different health issues that can be used for early disease prediction [26]. In recent years, the integration of ML techniques into the healthcare field has revolutionized how diseases are diagnosed, treated, and predicted. Disease prediction using ML has emerged as a promising research area, facilitating early detection and personalized interventions that hold the potential to improve patient outcomes and reduce the burden on healthcare systems. Researchers discovered that ML approaches produced excellent results, with 94.87 accuracy for MS severity assessment and 83.33 accuracy for disease progression prediction [27]. Evangelia et al. [28] mentioned several ML approaches for clinical prediction, including NN, SVM, and random forests. Here we have presented a few popular disease prediction models developed over time with ML.

#### 1) Sclerosis Prediction

Moustafa et al. [29] trained LR, SVM, RF NN, gradient boosting, and extreme gradient boosting and a Cox proportional hazards model to predict upper limb disability progression in multiple sclerosis (MS). Li et al. [30] stated that by using ML for assessing treatment switches among patients with multiple sclerosis, these models could accurately predict treatment switches among patients with MS. RF and LR models were used to predict treatment switching among patients with multiple sclerosis. The authors found that the RF model outperformed LR in predicting treatment switching. Zhao [31] assessed ML's capacity to forecast the illness progression of MS patients. The researchers utilised SVM and logistic regression to evaluate the predictive value of clinical and MRI characteristics in assessing patients' Expanded Disability Status Scale (EDSS) status over a five-year period. Law et al. [32] stated that three decision tree (DT) based models had greater AUC than independent and ensemble SVM models and LR. Afzal et al. [33] used deep learning techniques such as CNN. The author proposed an improved CNN, which uses LeNet architecture. Sarowar et al. [34] proposed an optimized-CNN algorithm which is a hybridization of CNN with Particle Swarm Optimization (PSO) for Tuberous Sclerosis Complex (TSC) disease prediction. Marzullo et al. [35] also use CNN on Magnetic Resonance (MR) images to predict disability progression in multiple sclerosis patients.

#### 2) Cardiovascular Disease Prediction

Predicting future cardiovascular disease risk using EHRs is an improtant research area in the healthcare industry [36]. Researchers are using ML techniques for the prediction of cardiovascular diseases. Rustam et al. [37] compared DT, Adaptive Boosting (ADA), SVM, RF, Extra Trees Classifier (ETC), LR, Stochastic Gradient Descent Classifier (SGDC), and the proposed Stacked Generalization with Linear Voting (SGLV) model. The study reported the accuracy of the models using CNN features, with the SGLV model achieving the highest accuracy of 91.5%.

***Table 3.*** Mortality Prediction using ML techniques

| Author | Dataset | Proposed Algorithm | Algorithms Compared | Performance metrics |
|---|---|---|---|---|
| **Hoogendoorn et al. [19]** | MIMIC-II, V2.6 | XGB | LR, KNN | AUC |
| **Lei et al. [23]** | Private | RNN-DAE | PCA, GMM, k means, SDAs | AUC, F1 Score |
| **Jun et al. [24]** | MIMIC-III, PhysioNet | V-RNN | KNN, GRU, RITS | AUC, AUPRC |
| **Abedi et al. [20]** | Private | XGB | RF, LR | AUROC |

***Table 4.*** Disease Prediction using ML techniques

| Author | Research domain | Dataset | Algorithms compared | Performance Metric | Proposed technique and algorithm |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Jieni Li et al. [30] | Sclerosis | TriNetX | LR, RF | Precision, recall, F1-score, accuracy | RF |
| Zhao et al. [31] | Multiple Sclerosis | private | LR | Sensitivity, specificity, accuracy | SVM |
| Law et al. [32] | Multiple Sclerosis | private | ensemble-SVM, LR | AUC | DT |
| Marzullo et al. [35] | Multiple Sclerosis | Magnetic Resonance (MR) images | - | RMSE | CNN |
| Furqan Rustam et al. [37] | Cardiovascular | CHD, SHD, SAHDD. HFPD | DT, SVM, RF, ETC, LR, SDGC, CNN, SGVL | - | CNN with SGVL |
| An et al. [38] | Cardiovascular | | | | DeepRisk |
| Oswald et al. [41] | Cardiovascular | - | SVM, DT, KNN, XGBoost, GBDT, CatBoost, Light GBM | AUC | LSTM, XGBoost |
| Xiao et al. [46] | Cancer | private | - | Precision, recall, accuracy | deep learning-based multi-model ensemble method |
| Kourou et al. [44] | Cancer | SEER cancer database | ANN, SVM, DT, RF | Sensitivity, specificity, Accuracy, AUC | BNs |
| Qin et al. [48] | Kidney | CKD | KNN, RF, SVM, LOG, FNN | Sensitivity, specificity, Accuracy, Precision, Recall, F1 Score | RF |
| Amirgaliyev et al. [49] | Kidney | UCI | - | Sensitivity, specificity | SVM |

An et al. [38] discussed EHR issues and explained how difficult it is to choose accurate characteristics from longitudinal and diverse EHRs, as well as how difficult it is to obtain accurate and robust representations for patients. To address this issue, they proposed the DeepRisk model, which is based on the attention mechanism and deep NN. Experiment results by Lyu et al. [39] showed that RF built by Classification and Regression Trees (CART) performs better than KNN, DT, GB with accuracy of 93.44%. To identify cardiovascular diseases K-means algorithm was used to analyze the characteristics and XGBoost to form a better classifier. [40] Ensemble learning algorithms were further proposed to predict Cardiovascular disease [41]. Zang et al. [42] proposed the LSTM-XGBoost model for early prediction of Heart Disease

### 3) *Cancer Prognosis and Prediction*

ML can be used for the prediction of cancer therapy [43]. Various kinds of research have shown promising results in the cancer prediction and prognosis. Kourou et al. [44] listed a variety of techniques like ANNs, Bayesian Networks (BNs), SVMs, and DT have been widely applied in cancer research for the development of predictive models. The author also highlighted that semi-supervised learning was also widely applied for cancer prediction, and SVMs are a most widely used ML methods for cancer prediction/prognosis. IT was observed that the BN with CFS algorithm performed better with 91.7% accuracy.

Raoof et al. [45] emphasize the importance of early detection, prediction, and diagnosis of lung cancer and how ML can aid in this process. ML techniques applied for the analysis and prognosis of lung cancer are Naive Bayes, SVM, LR, and ANN. Xiao et al. [46] proposed a deep learning-based multi-model ensemble method that shows accuracy of 98.8% on Lung Adenocarcinoma(LUAD) data

### 4) *Kidney Disease Prediction*

ML in the Kidney Disease Diagnosis (MLKDD) area is presently under active investigation that aims to assist physicians with computer-aided systems. Saha et al. [47] proposed neural network that is optimize by Adam optimizer outperforms other Random Forest, Naive Bayes, Multilayer Perceptron, Logistic Regression by predicting accuracy of 97.3%

Qin et al. [48] compared KNN, RF, SVM, LOG, and FNN and concluded that RF performed best having 99.75% diagnosis accuracy. Amirgaliyev et al. [49] and Khan et al. [50] used the SVM model for chronic kidney disease prediction.

Ogunleye et al. [51] suggested the use of the XGBoost model. A summary of disease prediction is presented in Table 4.

## VI. Findings and Future Directions

In this section we have reported the answers to the research questions in the form of findings and future directions.

*RQ1. What machine learning methods are employed for Electronic Health Records (EHR) analysis?*

Researchers proposed various ML techniques, like transfer learning and DeepRisk, but most studies focus on supervised learning techniques. LR, RF and SVM are the most widely used Baseline models. Some studies presented a modified version of these algorithms to improve their performance, while others used them to compare with other proposed algorithms. We have also discussed that the trend is moving towards deep learning techniques, and many researchers are using neural networks like CNN and RNN. Studies have shown that variants of RNN and CNN are providing promising results.

*RQ2. How EHR data is preprocessed for ML models?*

Some researchers highlighted preprocessing methods such as removing stop words, punctuation, numbers and converting all words to lowercase [8,9] MIMIC-III dataset was preprocessed using Spark. Some research also utilized n-gram feature extraction methods in their approach [9]. tfidf and word2vec are used for feature extraction [5].

*RQ3. Which datasets are utilized for validation?*

We have summarized various datasets used in the field of healthcare, such as Medical Information Mart for Intensive Care III (MIMIC III), Cleveland Heart Disease (CHD), BioASQ, South African Heart Disease Dataset (SAHDD), Statlog Heart Disease (SHD), HFPD, TriNetX, SEER cancer database, UCI, CKD, PhysioNet, CodiEsp corpus. Some studies used private hospital data, OPD reports, and Magnetic Resonance (MR) images. It is observed that the most widely used dataset is the publicly available MIMIC-III dataset for code assignment and mortality prediction.

*RQ4. Which performance metrics are used for the validation of results?*

In ML, performance metrics are used to validate results and assess the quality and effectiveness of trained models such as F1 measure, Accuracy, Precision, Recall, Area under the curve (AUC), Root Mean Square Error (RMSE) are extensively utilized performance measures.

*RQ5: What are the potential future directions in applying ML approaches for EHR?*

Although substantial research has been carried out to improve the healthcare sector, and numerous studies have contributed to remarkable improvements in various aspects, some studies still highlighted areas that need further attention and exploration in the future. Following are potential paths for future development

- **Unsupervised learning:** Most studies used supervised learning techniques, but in many instances, it is difficult to find properly labeled data that necessitates use of unsupervised and semi-supervised learning techniques. Very few studies have used unsupervised and semi supervised learning techniques. These areas are yet to be explored to analyze results. In other words, there is still a potential gap that can be filled in the future [52].
- **Datasets:** Healthcare datasets are complex in nature. In many instances it is difficult to find a suitable dataset for research. Many researchers use private datasets which are not freely accessible. MIMIC-III is the most widely used dataset for research purposes in the healthcare field. Researchers suggested that other datasets should be explored to gain better insights; like UCI Machine Learning Repository, PhysioNet, SEER cancer statistics.
- **Investigating alternative methodologies:** Exploring use of different deep learning architecture such as Hierarchical Attention Network (HAN), Transformer. Different feature extraction methods like sentence2vec or paragraph2vec should be explored to shorten input sequence. Though some studies used transfer learning, many researchers recommended exploration of transfer learning in their model [53][54].

## VII. Conclusions

We explored a wide range of ML algorithms applied to EHR. Our goal was to gain insights into the effectiveness and potential applications of these algorithms in the healthcare industry. Our survey revealed that ML algorithms had shown promising results in Disease prediction, Mortality prediction, and ICD-9/10 coding. Notably, with the advancement of deep learning algorithms, researchers use neural networks like CNN and RNN variants to achieve higher results. Various studies navigate the complexities of modern medical practice by incorporating unsupervised learning to enhance diagnostics and patient care in the healthcare industry. This paradigm shift has unlocked remarkable capabilities of deep learning approaches in extracting intricate patterns and features from complex datasets, which helps in enhancing the accuracy and performance of systems.

## References

[1] T. Sarwar, S. Seifollahi, J. Chan, X. Zhang, V. Aksakalli, I. Hudson, K. Verspoor, and L. Cavedon, “The secondary use of electronic health records for data mining: Data characteristics and challenges,” ACM Comput. Surv., vol. 55, no. 2, jan 2022.

[2] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis,” IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018.

[3] X. Zhan, M. Humbert-Droz, M. Humbert-Droz, P. Mukherjee, P. Mukherjee, O. Gevaert, and O. Gevaert, “Structuring clinical text with ai: old vs. new natural language processing techniques evaluated on eight common cardiovascular diseases,” medRxiv, 2021.

[4] S. Yang, P. Varghese, E. Stephenson, K. Tu, and J. Gronsbell, “Machine learning approaches for electronic health records phenotyping: a methodical review,” Journal of the American Medical Informatics Association, vol. 30, no. 2, pp. 367–381, 11 2022.

[5] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes," Computer Methods and Programs in Biomedicine, vol. 177, pp. 141–153, 2019.

[6] M. Gupta, T.-L. T. Phan, H. T. Bunnell, and R. Beheshti, "Concurrent imputation and prediction on ehr data using bi-directional gans: Bi-gans for EHR imputation and prediction," in Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, ser. BCB '21. New York, NY, USA: Association for Computing Machinery, 2021.

[7] S. Narayan, M. Gagne, and R. Safavi-Naini, "Privacy preserving EHR system using attribute- based infrastructure," in Proceedings of the 2010 ACM Workshop on Cloud Computing Security Workshop, ser. CCSW '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 47–52.

[8] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang, "Automatic icd-9 coding via deep transfer learning," Neurocomputing, vol. 324, pp. 43–50, 2019, deep Learning for Biological/Clinical Data.

[9] Singaravelan, C.-H. Hsieh, Y.-K. Liao, and J.- L. Hsu, "Predicting icd-9 codes using self-report of patients," Applied Sciences, 2021.

[10] J. Ferrão, F. Janela, M. Oliveira, and H. Mar- tins, "Using structured ehr data and svm to support icd-9-cm coding," in 2013 IEEE International Conference on Healthcare Informatics, 2013, pp. 511–516.

[11] Y. Chen, H. Lu, and L. Li, "Automatic ICD - 10 coding algorithm using an improved longest common subsequence based on semantic similarity," PLOS ONE, 2017.

[12] J. R. A. Solares, J. R. A. Solares, F. Rai- mondi, F. Raimondi, Y. Zhu, F. Rahimian, F. Rahimian, D. Canoy, J. Tran, A. C. P. Gomes, A. H. Payberah, A. H. Payberah, M. Zot- toli, M. Zottoli, M. Nazarzadeh, N. Conrad,K. Rahimi, and G. Salimi-Khorshidi, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," Journal of Biomedical Informatics, 2020.

[13] E. Moons, A. Khanna, A. Akkasi, A. Akkasi, and M.-F. Moens, "A comparison of deep learning methods for ICD coding of clinical records," Applied Sciences, 2020.

[14] M. Li, Z. Fei, M. Zeng, F.-X. Wu, Y. Li, Y. Pan, [20] and J. Wang, "Automated icd-9 coding via a deep learning approach," IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 16, no. 4, p. 1193–1202, 2019.

[15] A. Kumar, S. Roy, and S. Bhattacharjee, [21] "A fast unsupervised assignment of icd codes with clinical notes through explanations," in Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, ser. SAC '22.New York, NY, USA: Association for Computing Machinery, 2022, p. 610–618.

[16] Y. Liu, S. Qin, Z. Zhang, and W. Shao, "Compound Density Networks for Risk Prediction using Electronic Health Records," in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, dec 6 2022.

[17] J. Calvert, Q. Mao, J. L. Hoffman, M. Jay, T. Desautels, H. Mohamadlou, U. Chetti- pally, and R. Das, "Using electronic health record collected clinical variables to predict medical intensive care unit mortality," Annals of Medicine amp; Surgery, vol. 11, pp. 52–57, 11 2016.

[18] J. Ye, L. Yao, J. Shen, R. Janarthanam, and Y. Luo, "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes," BMC Medical Informatics and Decision Making, vol. 20, no. S11, 12 2020.

[19] M. Hoogendoorn, A. el Hassouni, K. Mok, M. Ghassemi, and P. Szolovits, "Prediction using patient comparison vs. modeling: A case study for mortality prediction," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 8 2016.

[20] V. Abedi, V. Avula, S.-M. Razavi, S. Bavishi, D. Chaudhary, S. Shahjouei, M. Wang, C. J. Griessenauer, J. Li, and R. Zand, "Predicting short and long-term mortality after acute ischemic stroke using ehr," Journal of the Neurological Sciences, vol. 427, p. 117560, 2021.

[21] J. Douglas Nascimento and T. Escovedo, "Construction of mortality tables using LSTM neural networks," in Proceedings of the XVII Brazilian Symposium on Information Systems, ser. SBSI '21. New York, NY, USA: Association for Computing Machinery, 2021.

[22] S. Kumar, S. C. Yu, T. Kannampallil, Z. Abrams, A. Michelson, and P. R. O. Payne, "Self-explaining neural network with concept based explanations for icu mortality prediction," in Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, ser. BCB '22. New York, NY, USA: Association for Computing Machinery, 2022.

[23] L. Lei, Y. Zhou, J. Zhai, L. Zhang, Z. Fang, P. He, and J. Gao, "An Effective Patient Representation Learning for Time-series Prediction Tasks Based on EHRs," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 12 2018.

[24] E. Jun, A. W. Mulyadi, and H.- I. Suk, "Stochastic Imputation and Uncertainty-Aware Attention to EHR for Mortality Prediction," in 2019 International Joint Conference on Neural Networks 7 (IJCNN). IEEE, 7 2019.

[25] Q. Tan, A. J. Ma, M. Ye, B. Yang, H. Deng, V. W.-S. Wong, Y.-K. Tse, T. C.-F. Yip, G. L.- H. Wong, J. Y.-L. Ching, F. K.-L. Chan, and P. C. Yuen, "Ua-crnn: Uncertainty-aware convolutional recurrent neural network for mortality risk prediction," in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 109–118

[26] M. K. Lodhi, R. Ansari, Y. Yao, G. M. Keena, D. J. Wilkie, and A. A. Khokhar, "Predictive Modeling for Comfortable Death Outcome Using Electronic Health Records," in 2015 IEEE International Congress on Big Data. IEEE, 6 2015.

[27] D. Plati, E. Tripoliti, S. Zelilidou, K. Vlachos, S. Konitsiotis, and D. I. Fotiadis, "Multiple Sclerosis Severity Estimation and Progression Prediction Based on Machine Learning Techniques," in 2022 44th Annual International Conference of the IEEE Engineering in Medicine amp; Biology Society (EMBC). IEEE, 2022.

[28] E. Christodoulou, J. Ma, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. V. Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical

prediction models," Journal of Clinical Epidemiology, 2019.

[29] S. Mostafa, I. H. J. Song, A. A. Metwally, N. Strauli, N. Sewde, M. Friesenhahn, M. Usdin, and X. Jia, "Predicting upper limb disability progression in primary progressive multiple sclerosis using machine learning and statistical methods," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021.

[30] J. Li, Y. Huang, G. J. Hutton, and R. R. Aparasu, "Assessing treatment switch among patients with multiple sclerosis: A machine learning approach," Exploratory Research in Clinical and Social Pharmacy, vol. 11, p. 100307, 2023.

[31] Y. Zhao, B. C. Healy, D. Rotstein, C. R. G. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, and T. Chitnis, "Exploration of machine learning techniques in predicting multiple sclerosis disease course." PLOS ONE, 2017.

[32] M. T. Law, A. L. Traboulsee, D. K. Li, R. L. Carruthers, M. S. Freedman, S. H. Kolind, and R. Tam, "Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression," Multiple Sclerosis Journal - Experimental, Translational and Clinical, vol. 5, no. 4, p. 205521731988598, 10, 2019.

[33] H. M. R. Afzal, S. Luo, S. Ramadan, J. Lechner-Scott, and J. Li, "Automatic prediction of the conversion of clinically isolated syndrome to multiple sclerosis using deep learning," in Proceedings of the 2018 2nd International Conference on Video and Image Processing, ser. ICVIP '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 231–235.

[34] M. G. Sarowar, F. Qasim, and S. H. Ripon, "Tuberous sclerosis complex (TSC) disease prediction using optimized convolutional neural network," in Proceedings of the 7th International Conference on Computer and Communications Management, ser. ICCCM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 210–215.

[35] A. Marzullo, G. Kocevar, C. Stamile, F. Calimeri, G. Terracina, F. Durand-Dubief, and D. Sappey-Marinier, "Prediction of Multiple Sclerosis Patient Disability from Structural Connectivity using Convolutional Neural Networks," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 7 2019.

[36] Y. An, K. Tang, and J. Wang, "Timeaware multi-type data fusion representation learning framework for risk prediction of cardiovascular diseases," IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 19, no. 6, p. 3725–3734, oct 2021.

[37] F. Rustam, A. Ishaq, K. Munir, K. Munir, M. Almutairi, N. Aslam, I. Ashraf, and I. Ashraf, "Incorporating cnn features for optimizing performance of ensemble classifier for cardiovascular disease prediction," Diagnostics, 2022.

[38] Y. An, N. Huang, X. Chen, F. Wu, and J. Wang, "High-risk prediction of cardiovascular diseases via attention-based deep neural networks," IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 18, no. 3, p. 1093–1105, 2019.

[39] H. Lyu, "A machine learning-based approach for cardiovascular diseases prediction," in 2022 14th International Conference on Machine Learning and Computing (ICMLC), ser. ICMLC 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 59–66.

[40] Y. Wang, "Identification of cardiovascular diseases based on machine learning," in Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences, ser. ISAIMS '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 531–536.

[41] Oswald, G. Jaya Sathwika, and A. Bhattacharya, "Prediction of cardiovascular disease (CVD) using ensemble learning algorithms," in 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD), ser. CODS-COMAD 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 292–293

[42] X. Zang, J. Du, and Y. Song, "Early prediction of heart disease via LSTM-XGboost," in Proceedings of the 2023 9th International Conference on Computing and Artificial Intelligence, ser. ICCAI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 631–637.

[43] R. Rafique, S. R. Islam, and J. U. Kazi, "Machine learning in the prediction of cancer therapy," Computational and Structural Biotechnology Journal, vol. 19, pp. 4003–4017, 2021.

[44] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015.

[45] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, "Lung cancer prediction using machine learning: A comprehensive approach," in 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 108–115.

[46] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," Computer Methods and Programs in Biomedicine, vol. 153, pp. 1–9, 2018.

[47] A. Saha, A. Saha, and T. Mittra, "Performance measurements of machine learning approaches for prediction and diagnosis of chronic kidney disease (ckd)," in Proceedings of the 7th International Conference on Computer and Communications Management, ser. ICCCM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 200–204.

[48] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for 9 diagnosing chronic kidney disease," IEEE Access, vol. 8, pp. 20 991–21 002, 2020.

[49] Y. Amirgaliyev, S. Shamiluulu, and A. Serek, "Analysis of chronic kidney disease dataset by applying machine learning methods," in 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1–4.

[50] B. Khan, R. Naseem, F. Muhammad, G. Abbas, and S. Kim, "An empirical evaluation of machine learning techniques for chronic kidney disease prophecy," IEEE Access, vol. 8, pp. 55 012– 55 022, 2020.

[51] A. Ogunleye and Q.-G. Wang, “Xgboost model for chronic kidney disease diagnosis,” IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 17, no. 6, p. 2131–2140, 2020.

[52] A. Rajasekhar, R.K. Jatoth, A. Abraham, Design of intelligent PID/PID speed controller for chopper fed DC motor drive using opposition based artificial bee colony algorithm, Engineering Applications of Artificial Intelligence, 29: 13-32, 2014.

[53] H. Izakian, B. Ladani, K. Zamanifar and A. Abraham, A Particle Swarm Optimization Approach for Grid Job Scheduling, Third International Conference on Information Systems, Technology and Management, Communications in Computer and Information Science, Springer Verlag, ISBN 978-3-642-00404-9, pp. 100-109, 2009.

[54] M. Ali, M. Pant and A. Abraham, Simplex differential Evolution, Acta Polytechnica Hungarica, 6(5):95-115, 2009.