International Journal of Computational Research in Science and Technology
Volume 1 Issue 1
Sep-Oct 2025, PP 39-45

A Survey on Electronic Health Data Analysis Techniques and Features for Machine Learning

Neeraj Mishra
Dept. of Computer Science & Engineering
SAM Engineering College Bhopal, MP, India
nm121052@gmail.com

Abstract. The increasing adoption of electronic health records (EHRs) and digital medical imaging systems has created unprecedented opportunities to apply machine learning (ML) in healthcare. This paper presents a comprehensive review of the features of electronic healthcare data—spanning structured tabular data, unstructured clinical notes, and imaging modalities—and the ML techniques used to extract clinical value from them. Work has list various learning approaches, including supervised, unsupervised, and ensemble methods. As most of medical data are in images hence image features like co-occurrence matrices, wavelet transformations, edge detection, etc. were brief. This review aims to bridge the gap between technical advances in machine learning and their practical implications for modern healthcare delivery.

Index Terms: Digital Image Processing, Electronic Health Data, Machine Learning, Feature Extraction.

I. Introduction

The explosion of digital data in healthcare systems, primarily through Electronic Health Records (EHRs), has transformed how medical information is collected, processed, and analyzed. These records encompass a broad spectrum of patient data, including clinical diagnoses, laboratory results, imaging, prescriptions, and free-text physician notes. The richness and granularity of this data make it a fertile ground for extracting insights using machine learning (ML) techniques, which can significantly enhance clinical decision-making, disease prediction, and patient outcome assessment.

Traditional data analytics approaches in healthcare often relied on rule-based systems or statistical models, which, while interpretable, struggled with the complexity, high dimensionality, and heterogeneity of real-world clinical data. In contrast, ML algorithms can model non-linear relationships, handle missingness, and adaptively improve with exposure to new data. This capability is particularly useful for handling EHD features such as irregular time series, varying data quality, and unstructured textual content (Shickel et al. Xiao et al.) [1].

Several studies have demonstrated the potential of ML to predict health outcomes more accurately than conventional statistical techniques. For example, Wong et al. (2018) employed decision trees and logistic regression to extract outcome measures from

EHRs, proving that even simple ML models could significantly enhance epidemiological research and clinical monitoring (Wong et al.) [2]. Similarly, Bardhi and Zapirain (2021) applied ensemble ML models to predict cancer patient survivability, outperforming traditional classifiers in predictive accuracy and model robustness (Bardhi & Zapirain) [3].

Moreover, deep learning, a subset of ML, has gained traction in healthcare due to its ability to automatically extract hierarchical features from raw data. In their seminal work, Shickel et al. (2017) [1]reviewed the application of deep neural networks to EHR data, highlighting breakthroughs in phenotype discovery, temporal modeling, and clinical risk prediction (Shickel et al.). Deep learning models such as Recurrent Neural Networks (RNNs) and Transformers have proven especially effective for modeling sequential data like longitudinal patient histories.

However, the integration of ML into healthcare workflows is not without challenges. Data privacy, ethical concerns, and algorithmic fairness remain critical hurdles. Seh et al. (2022) [4] emphasize the need for secure ML implementations that safeguard sensitive patient information while enabling real-time analytics (Seh et al.). In addition, Xiao et al. (2018) identify the lack of standardized data formats and ground truth labels as significant barriers to the scalability of deep learning models in EHR contexts (Xiao et. al.) [5].

Despite these barriers, there is growing consensus that ML has the potential to drive a paradigm shift in healthcare, moving from reactive care to proactive, predictive, and personalized medicine. This paper aims to review the defining features of EHD, explore the key ML techniques applied to them, address the challenges, and summarize recent landmark studies to provide a comprehensive overview of this rapidly evolving domain.

II. Features of Image

In medical image analysis, feature extraction is critical to enabling machine learning (ML) models to interpret and classify clinical imaging data. While deep learning techniques are capable of automatically learning features from raw image pixels, traditional handcrafted features continue to play a significant role—particularly in radiomics, classical ML pipelines, and settings requiring interpretability. These features, drawn from color, texture, frequency, edge, and corner properties of medical images, help convert raw data into quantifiable, clinically relevant information.

Color Features are primarily used in modalities that retain natural color information, such as dermoscopy, histopathology, endoscopy, and retinal imaging. These features are quantified using methods such as color histograms and color moments, which describe the distribution and statistical properties (e.g., mean, standard deviation, skewness) of pixel intensities across RGB or HSV channels. In their study on diabetes recognition using E-healthcare data, Haq et al. (2020) [6], emphasized the importance of color analysis in classifying lesion types from clinical images. Similarly, color variation is a primary criterion in melanoma detection and tissue classification in histopathology (Kather et al., 2016) [7].

Texture Features, particularly those derived from the Gray-Level Co-occurrence Matrix (GLCM)—also referred to as Co-occurrence Matrix (CCM)—are among the most commonly used handcrafted descriptors in medical imaging. The GLCM encodes how often pairs of pixel intensities occur at specific spatial relationships, enabling the extraction of high-level statistical properties such as contrast, correlation, energy, and homogeneity. These features are especially effective in characterizing tissue heterogeneity in cancer, liver fibrosis, or brain tumors. Shickel et al. (2017) [1] reviewed the application of texture-based radiomics features across different modalities, noting their effectiveness in predicting tumor aggressiveness and patient outcomes.

Frequency-Domain Features, derived from transformations such as the Fourier Transform and the Discrete Wavelet Transform (DWT), help identify periodic structures and textural variations at multiple scales. The Fourier Transform decomposes images into their frequency components, making it suitable for global pattern detection, while DWT provides localized frequency information. These features have shown strong performance in applications like breast cancer detection, retinal disease classification, and tissue segmentation. For example, Dua et al. (2014) [8]highlighted the use of wavelet-based features in mammogram and ultrasound image classification tasks. Wavelet features have also been used as inputs to SVM classifiers and hybrid ML-DL systems for improved lesion detection accuracy.

Edge Features: identify regions in the image where pixel intensity changes sharply, often corresponding to anatomical boundaries, tumor margins, or vascular structures. Methods such as the Sobel, Prewitt, and particularly the Canny edge detector are commonly employed in clinical settings. The Canny algorithm, known for its robustness to noise, is frequently used in lung field segmentation from chest X-rays and in detecting organ contours in CT and MRI scans. Pollettini et al. (2012) [9], successfully applied edge-based segmentation in electronic patient records to assist with classification tasks, reinforcing the value of these features in automated diagnosis.

Corner Features: are particularly useful for tasks like image registration, landmark detection, and shape recognition. The Harris Corner Detector is widely used for identifying interest points where intensity changes in multiple directions. More advanced descriptors like SIFT (Scale-Invariant Feature Transform) and SURF (Speeded-Up Robust Features) are scale- and rotation-invariant, making them ideal for multi-modal image alignment (e.g., MRI with PET) or longitudinal tracking. These descriptors have been used in aligning neuroimaging data across patients and in tracking tumor evolution over time (Liu et al., 2014) [10].

III. Techniques of Machine Learning

Machine learning (ML) has emerged as a powerful tool to unlock insights from vast and diverse healthcare datasets, particularly those derived from Electronic Health Records (EHRs). The application of different ML techniques depends on the nature of the data, the desired prediction outcomes, and the clinical domain. This section highlights a range of ML methods implemented in various healthcare studies, each selected to represent a different methodological approach.

Support Vector Machines (SVM): Clinical Feature-Based Diabetes Prediction Haq et al. (2020) [6] presented an intelligent system for diabetes recognition using a combination of clinical data and SVM. The researchers applied a correlation-based feature selection technique before training the SVM, enabling the model to focus on the most relevant predictors such as glucose level, BMI, and blood pressure. The system achieved an impressive accuracy of over 90%, outperforming decision trees and Naive Bayes classifiers.

Random Forest: Classifying Electronic Patient Records
In an early but foundational study, Pollettini et al. (2012) [9] implemented Random Forest classifiers to automate the classification of electronic patient records. Using a relatively small dataset of 100 EHRs, the model accurately grouped patients based on predefined diagnostic categories. The Random Forest method was selected due to its ability to handle noisy features and imbalanced classes, common issues in clinical datasets.

Ensemble Methods: Cancer Survivability Prediction
Bardhi and Zapirain (2021) [3] explored cancer survivability using an ensemble of nine machine learning algorithms, including SVM, Logistic Regression, Random Forest, and AdaBoost. They used the StratifiedKFold cross-validation method and compared each model's performance individually and in combination. The ensemble approach yielded better F1 scores and overall predictive stability, making it suitable for survival prediction in oncology.

Deep Neural Networks (DNN): Longitudinal EHR Representation
Shickel et al. (2017) [1] conducted a survey of deep learning architectures applied to EHR data and highlighted how Recurrent Neural Networks (RNNs) and Deep Neural Networks (DNNs) can model time-series patient records. For example, RNNs have been used to predict ICU patient mortality by learning temporal patterns across vitals, medications, and lab results. These models learn automatically from raw data without feature engineering, making them ideal for longitudinal patient trajectories.

Gradient Boosting Machines: Cardiovascular Risk Prediction

In a clinical decision support setting, Li et al. (2022) [10] applied Gradient Boosting Machines (GBM) to predict cardiovascular disease (CVD) using EMR data from a regional healthcare system. The researchers created a population-specific Atherosclerotic

Cardiovascular Disease (ASCVD) risk calculator, improving accuracy compared to traditional calculators like Framingham. GBM handled the non-linearity and variable importance ranking in complex clinical datasets effectively.

Privacy-Aware Machine Learning: Secure EHR Modeling

Seh et al. (2022) [4] reviewed privacy-preserving machine learning techniques, such as federated learning and differential privacy, applied in healthcare. These methods enable institutions to build ML models collaboratively across sites without sharing raw patient data. Their work emphasizes the need for secure AI systems in the context of EHR integration and regulatory compliance.

Study (Year)	ML Technique	Benefits	Limitations
Garriga et al. (2022) [11]	Gradient Boosting (XGBoost)	Accurate prediction of mental health crises using temporal EHR data; real-time applicability	Limited generalizability outside UK healthcare system; lacks explainability
Hobensack et al. (2023) [12]	Ensemble of Decision Trees, Random Forest	Comprehensive scoping review for home healthcare; interpretable models	Focused more on review than empirical modeling; sparse labeled data
Nguyen et al. (2024) [13]	Wide & Deep Learning	Combines generalization with deep feature learning; high performance in diabetes onset prediction	Requires significant computational resources; interpretability issues
Rojas et al. (2021) [14]	Logistic Regression, Random Forest, SVM	Effective ICU readmission prediction using real-world EHR and MIMIC-III data	Overfitting in smaller subpopulations; needs better calibration
Cheng et al. (2025) [15]	Deep Learning (Feedforward NN)	Predicts 30-day mortality/readmission; scalable for real-time EMR applications	Black-box nature makes clinical adoption harder; no external validation

IV. Conclusion

Machine learning has shown immense potential in transforming healthcare by extracting meaningful patterns from electronic health data, including structured records and medical images. This review highlighted key ML techniques—ranging from classical algorithms to deep learning—and their use in processing diverse healthcare features such as color, texture, frequency, and edges. Recent studies demonstrate improved prediction and decision support, though challenges like data heterogeneity, privacy, and interpretability remain. Continued research should focus on building explainable, privacy-preserving models that integrate seamlessly into clinical practice.

References

1. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604.
2. Wong, J., Horwitz, M. M., Zhou, L., & Toh, S. (2018). Using machine learning to identify health outcomes from electronic health record data. Current Epidemiology Reports, 5(3), 263–269.
3. Bardhi, O., & Zapirain, B. G. (2021). Machine learning techniques applied to electronic healthcare records to predict cancer patient survivability. Computer Methods and Programs in Biomedicine, 200, 105897.
4. Seh, A. H., Al-Amri, J. F., Subahi, A. F., & Agrawal, A. (2022). An analysis of integrating machine learning in healthcare for ensuring confidentiality of the electronic records. Computer Modeling in Engineering & Sciences (CMES), 132(2), 731–753.
5. Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. Journal of the American Medical Informatics Association (JAMIA), 25(10), 1419–1428.
6. Haq, A. U., Li, J. P., Khan, J., Memon, M. H., Nazir, S., & Ahmad, S. (2020). Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. Sensors, 20(9), 2649.
7. Kather, J. N., Weis, C. A., Bianconi, F., et al. (2016). Multi-class texture analysis in colorectal cancer histology. Scientific Reports, 6, 27988.
8. Dua, S., Acharya, U. R., & Dua, P. (2014). Machine learning in healthcare informatics. Springer.
9. Pollettini, J. T., Panico, S. R. G., Daneluzzi, J. C., & Tinós, R. (2012). Using machine learning classifiers to assist healthcare-related decisions: classification of electronic patient records. Journal of Medical Systems, 36, 345–350.
10. Liu, S., Liu, S., Cai, W., et al. (2014). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Transactions on Biomedical Engineering, 62(4), 1132–1140.
11. Garriga, R., Mas, J., Abraha, S., Nolan, J., & Harrison, O. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*.
12. Hobensack, M., Song, J., Scharp, D., & Bowles, K. H. (2023). Machine learning applied to electronic health record data in home healthcare: A scoping review. *Computer Methods and Programs in Biomedicine*, *226*, 107127.
13. Nguyen, B. P., Pham, H. N., Tran, H., & Nghiem, N. (2024). Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. *Computer Methods and Programs in Biomedicine*.

14. Rojas, J. C., Carey, K. A., & Edelson, D. P. (2021). Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society*, *15*(11), 1328–1335.
15. Cheng, Y., Wang, F., Zhang, P., & Hu, J. (2025). Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the SIAM International Conference on Data Mining*.