UDC 519.6 *http://dx.doi.org/10.30970/vam.2024.33.11975*

A SYSTEMATIC REVIEW OF UNSUPERVISED ANOMALY DETECTION FOR EARLY DISEASE IDENTIFICATION IN ELECTRONIC HEALTH RECORDS

I. Ursul

Ivan Franko National University of Lviv,
1, Universytetska str., 79000, Lviv, Ukraine
e-mail: *ivan.ursul@lnu.edu.ua*

Electronic Health Records (EHRs) have emerged as pivotal patient data repositories. Which holds immense potential to revolutionize healthcare by facilitating early disease identification. Central to this transformation is the application of unsupervised anomaly detection methods. Which uniquely enables the discovery of hidden patterns and irregularities within complex, multi-dimensional healthcare datasets. This paper presents an exhaustive systematic review of the diverse methodologies employed in unsupervised anomaly detection, specifically targeting early disease identification within EHRs. The proposed analysis spans many geographical regions. Which reflects a global research effort to develop universally applicable solutions in various healthcare systems. This geographic diversity underscores the universal relevance and adaptability of unsupervised anomaly detection methods in healthcare. We examined various subject areas, from clinical diagnoses to administrative processes, demonstrating these methods' versatility. The methodologies employed in these studies are varied and innovative, highlighting the evolving nature of the field. This diversity illustrates the potential for cross-disciplinary collaboration between data scientists and healthcare professionals. Moreover, it emphasizes the need for such fusion to address complex healthcare challenges effectively. The temporal analysis reveals a dynamic research area shaped by rapid advancements in machine learning and the increasing availability of large-scale datasets. This review also addresses the challenges in implementing these techniques within EHRs. for example, ensuring data privacy, maintaining data quality, and the interpretability of machine learning models. We propose potential solutions and areas for future research to overcome these hurdles. Moreover, we discuss integrating unsupervised anomaly detection methods into existing healthcare practices. Furthermore, the paper suggests future research directions, including exploring advanced techniques like Generative Adversarial Networks (GANs) and transformer-based models. The proposed systematic review provides comprehensive insights into the trajectory of research in unsupervised anomaly detection within EHRs. It underscores the importance of merging technical expertise with domain-specific knowledge. It highlights this critical field's global trends, subject areas, and temporal patterns. This work is instrumental for researchers and practitioners who leverage unsupervised anomaly detection for early disease identification.

Key words: anomaly, Unsupervised, Disease, SLR, Health Record, EHR, Machine Learning, Deep Learning.

1. Introduction

As healthcare becomes increasingly data-driven, Electronic Health Records (EHR) have emerged as a promising tool for improving patient care. These digital records contain a wealth of information about a patient's medical history, including diagnoses, treatments, laboratory results, and physician's notes. One promising application of EHRs is the early identification of diseases, which can significantly improve patient outcomes. This paper presents a systematic review of unsupervised anomaly detection techniques for early disease identification in EHRs. Unsupervised anomaly detection is a type of

© Ursul I., 2024

machine learning that identifies patterns or outliers in data without needing prior labeling or categorization. In the context of EHRs, this could involve identifying patterns in a patient's medical history that suggest the early stages of a disease, even before it has been formally diagnosed. Despite the significant potential of unsupervised anomaly detection in EHRs for early disease identification, there are several challenges to its effective implementation. These include data privacy and security, data quality and completeness, the interpretability of machine learning models, and the need for validation in real-world clinical settings. Additionally, there is a lack of a comprehensive understanding of the various methods and their relative effectiveness, limiting their widespread adoption. The proposed solution to the outlined problem is a systematic review of the existing literature on unsupervised anomaly detection for early disease identification in EHRs. This review will comprehensively summarise the current methods, their effectiveness, and their limitations. It will also highlight potential solutions to the identified challenges and suggest directions for future research. This review can guide practitioners and researchers in the field by providing a consolidated view of the state-of-the-art. This review seeks to answer several key research questions:

1. What are the current methods of unsupervised anomaly detection in EHRs for disease identification?
2. How effective are these methods in identifying diseases early?
3. What are the key challenges and limitations associated with these methods?
4. What are the potential future directions for research in this area?

The motivation for this study stems from two key developments. Firstly, the advent and proliferation of EHRs have given us an unprecedented ability to capture and store massive amounts of medical data. Secondly, the recent progress in machine learning techniques, mainly unsupervised learning, offers sophisticated tools for mining this data for insights. However, there is a gap in the literature concerning the application of unsupervised anomaly detection for disease identification in EHRs. This review aims to fill this gap and provide a comprehensive understanding of the state-of-the-art in this area.

2. Methodology

The methodology of this systematic review is underpinned by a structured and rigorous approach to identifying, selecting, and synthesizing relevant literature. The following subsections provide a detailed overview of the search strategy, selection criteria, data extraction, and quality assessment used in this review. The search strategy was designed to find published and unpublished studies addressing unsupervised anomaly detection in EHRs for disease identification. A comprehensive search of several databases was conducted, including PubMed, Embase, Cochrane Library, IEEE Xplore, and ACM Digital Library. The search strategy was tailored to each database to capture as many relevant studies as possible. Search terms included a combination of keywords and MeSH terms related to "Unsupervised Learning", "Anomaly Detection", "Outlier Detection", "Novelty Detection", "Disease Identification", "Early Disease Identification", "Pattern Recognition in EHR", "EHR Analysis", "Clinical Data Analysis", "Healthcare Big Data", "EHR Informatics", "EHR Data Mining".

Inclusion criteria were defined to identify relevant studies. Studies were included if they: Focused on unsupervised anomaly detection techniques; Used EHR data; and Aimed at early disease identification. No restrictions were placed on the publication date or language of the studies. After removing duplicates, two reviewers independently screened the titles and abstracts of the studies, followed by a full-text review. Discrepancies between the reviewers were resolved through discussion or involving a third reviewer. For each included study, relevant data were extracted by the reviewers. This included study characteristics (e.g., authors, year, country), data source and size, anomaly detection methods, diseases targeted for identification, key findings, and limitations. The extraction was conducted using a predefined data extraction form to ensure consistency. The quality of the included studies was evaluated using an appropriate quality assessment tool, such as the Critical Appraisal Skills Programme (CASP) Checklist or the Joanna Briggs Institute (JBI) Checklist, depending on the study design. This assessment was used to identify any potential biases in the studies and evaluate the results' overall robustness. Findings from the included studies were synthesized narratively. Due to the expected heterogeneity in the studies, a meta-analysis was not planned. Instead, the synthesis focused on summarizing the different unsupervised anomaly detection methods, their application in EHRs, and their effectiveness in disease identification. Additionally, challenges and limitations identified in the studies were collated and discussed. ABased on weight, age, and dosage for 21 medications, their framework ide a comprehensive and unbiased summary of the current state of research in unsupervised anomaly detection for early disease identification in EHRs.

3. Research Findings

3.1. Semi Supervised Approaches

Sabic et al. [1] employed the one-class support vector machine (OCSVM) for anomaly detection to pinpoint underdose and overdose prescriptions using EHR data from Kyushu University Hospital between 2014 and 2019. Their framework, based on weight, age, and dosage for 21 medications, outperformed three other unsupervised algorithms. It successfully identified 27 out of 31 clinical dosage errors, achieving an accuracy of 0.986, an F-measure of 0.97, and a recall of 0.964. Hou et al. [61] aimed to detect mammographic calcifications, and key breast cancer indicators using a one-class, semi-supervised deep convolutional autoencoder. Due to limited positive samples, traditional models risk overfitting. Hou's model, trained on 50,000 negative example images, detected calcifications by comparing input to reconstructed outputs. It used a structural dissimilarity measure for accuracy and achieved an AUROC of 0.959 and AUPRC of 0.676, the model had a 75% sensitivity.

3.2. Fuzzy Search

Shi et al. [3] developed an automated data cleansing strategy for EHR data, building a Clinical Knowledge Database using data from Flanders, Belgium. Post-cleansing, 42 variables showed a 1% reduction in missing data, nine had a decrease between 1-10%, and one had a 13.36% drop in completeness. After cleaning, all variables exceeded 50% of their normal range values. Seh et al. [8] investigated the impact of AI, ML, and IoT on healthcare, noting privacy and data security concerns. They recommended an ML framework to detect irregular end-user EHR access. A fuzzy-based Analytical Network Process showed that the accuracy level of M4 and anomaly detection of M2 had the highest global weights. Rijcken et al. [52] introduced FLSA-E, a topic modeling technique, to enhance the interpretability of text categorization in EHRs. This method showed fewer topic outliers, implying better semantic cohesion among topic words than its counterpart, FLSA-W. Ma et al. [51] used association rule mining on a large-scale EHR database from 2015 to 2020 to explore illness co-occurrence patterns. Using the Apriori method, they pinpointed 110 strongly correlated illness combinations, particularly circulatory and metabolic disorders.

Table 1

Summary of Various Studies in Healthcare Research

References	**Method**	**Results**
Sabic et al. [1]	One-class SVM for anomaly detection in EHR data.	Accuracy: 0.986, Recall: 0.964
Hou et al. [61]	Semi-supervised deep convolutional autoencoder for mammographic calcifications.	AUROC: 0.959, Sensitivity: 75%
Shi et al. [3]	Automated data cleansing for EHR data.	Reduced missing data; improved data completeness.
Seh et al. [8]	ML framework for detecting irregular EHR access.	Focused on data security.
Rijcken et al. [52]	Topic modeling for EHR text categorization.	Enhanced semantic cohesion.
Ma et al. [51]	Association rule mining for illness co-occurrence patterns.	Identified 110 illness combinations.
Hushig-Muzo et al. [26]	Denoising Autoencoders for EHR data representation.	Simplified complex EHR data analysis.
Rochner and Rothlauf [32]	Unsupervised anomaly detection in cancer EHRs.	Identified 28% of records as improbable.
Chen et al. [12]	Multi-scale Attention Memory Autoencoder (MAMA Net).	Accuracy: 95%
Ramos et al. [68]	Unsupervised vs. supervised models for septic shock prediction.	AUC: 0.82, F1-score: 0.65

3.3. Auto Encoders

Hushig-Muzo et al. [26]: Investigated the use of Denoising Autoencoders (DAEs) for mapping complex EHR data into two-dimensional latent representations for easier analysis. Based on data from the University Hospital of Fuenlabrada in Spain, the study demonstrated that DAEs could detect patterns related to various chronic conditions. The study highlighted the potential of DAEs in simplifying complex EHR data for improved patient care. Rochner and Rothlauf [32] utilized unsupervised anomaly detection methods (FindFPOF and autoencoders) to identify implausible EHRs in cancer registries. Using a dataset of 21,104 EHRs, both methods identified 28% of a sample of 300 records as improbable. While both methods showcased a specificity of 94%, the autoencoder and FindFPOF had sensitivities of 22% and 26%, respectively. Chen et al. [12] introduced the Multi-scale Attention Memory with hash addressing Autoencoder network (MAMA Net) to improve anomaly detection. This new technique features a multi-scale global spatial attention block and a hash-solving memory module. Tested on various datasets, MAMA

Net achieved a 95% accuracy rate, outperforming other baseline approaches. Ramos et al. [68] proposed unsupervised learning algorithms using Recurrent Autoencoders to predict septic shock development in ICU settings. Compared to a supervised LSTM network, the unsupervised methods showed competitive performance. An unsupervised model employing a Variational Autoencoder (VAE) combined with Gaussian Mixture Models achieved an AUC of 0.82 and an F1-score of 0.65, nearly matching the supervised LSTM's performance.

3.4. Network Toplogy Approach

Santos et al. [5] designed NoHarm.ai, an open-source tool to rate outlier prescriptions using a network-based unsupervised algorithm. Implemented in a large hospital, it enhanced pharmacist performance with an F-measure of 84%. Niu et al. [42] improved anomaly detection in EHRs with a network-based approach, outclassing traditional methods. Their technique better captures data intricacies in large-scale health systems. Li et al. [34] introduced a semi-supervised technique using sparse EHR labels. It surpassed other models, achieving impressive AUC values across datasets. Chew [11] applied a transition graph method to detect anomalous blood glucose trends. Using the LIFECARE dataset, this method identified anomalies that were confirmed by experts. Zhang et al. [7] tackled medical fraud using a neural model. They addressed data imbalance and reduced analyst burden, detecting around 71% of anomalies. Niu et al. [6] highlighted risks in health IT using adaptive anomaly detection. Their method efficiently detected abnormal patterns in EHRs, outperforming existing systems.

3.5. Unsupervised Deep Learning Based Approaches

Shimauchi [2] enhanced anomaly detection using a semi-supervised approach with GANs, achieving an improved ROC score in the Arrhythmia dataset. You et al. [9] and Ashfaq et al. [39] leveraged unsupervised and deep learning techniques for seizure detection and CHF patient readmissions, respectively, showing significant success. Chen et al. [28], Tomasev et al. [15], and Xu et al. [27] focused on data-driven methods, prediction protocols, and NAS approaches for multimodal EHR data, providing advancements in phenotypic theme similarities and patient outcome predictions. Bala [25], Kumaar et al. [70], and Boussina et al. [40] underscored the potential of machine learning and unsupervised learning in healthcare decision frameworks, intrusion detection, and clinical characteristic discoveries. Alhassan et al. [62] achieved noteworthy accuracy with the multilayer perceptron model for predicting HbA1c increase risks. Rashidian [46], Aguiar et al. [59], and Ibrahim [4] utilized LSTM-DNN, deep learning, and ML systems for patient classification, clustering, and clinical decision-making. Their models exhibited significant success in diagnosis accuracy and adverse outcome predictions. Manne and Kantheti [63] tested AI models on EHRs for anomaly detection, showcasing a high accuracy rate with the RNN model. Wang et al. [58] employed LSTM AE to predict mortality in hemodialysis patients, surpassing other models in predicting deaths within a short period.

Table 2

Summary of Various Studies in Healthcare Research Across Different Approaches

References	**Method**	**Results**
Santos et al. [5]	Network-based unsupervised algorithm for outlier prescriptions.	Enhanced pharmacist performance, F-measure: 84%.
Niu et al. [42], [6]	Network-based and adaptive anomaly detection in EHRs.	Improved detection of data intricacies and abnormal patterns.
Li et al. [34]	Semi-supervised technique with sparse EHR labels.	High AUC values across datasets.
Chew [11]	Transition graph method for blood glucose trends.	Identified anomalies confirmed by experts.
Zhang et al. [7]	Neural model for medical fraud detection.	Detected about 71% of anomalies.
Shimauchi [2]	GANs in a semi-supervised approach for arrhythmia detection.	Improved ROC score.
You et al. [9], Ashfaq et al. [39]	Unsupervised and deep learning for seizure and CHF patient readmissions.	Significant detection and prediction success.
Chen et al. [28], Tomasev et al. [15], Xu et al. [27]	Data-driven methods, prediction protocols, and NAS for multimodal EHR data.	Advancements in patient outcome predictions.
Bala [25], Kumaar et al. [70], Boussina et al. [40]	ML and unsupervised learning in healthcare frameworks.	Enhanced clinical decision-making and characteristic discoveries.
Alhassan et al. [62]	Multilayer perceptron model for HbA1c risk prediction.	Noteworthy accuracy.
Rogers et al. [22], Ta et al. [35]	Clustering for clinical trials and COVID-19 patient categorization.	Improved clinical decision-making.
Lutz [31], Mishra et al. [53], Moy et al. [48]	Clustering techniques to enhance predictions and detect clinician shifts.	Uncovered meaningful medical structures.
Marimuthu and Vaidehi [69]	Personalized clustering in Remote Health Monitoring.	High specificity and sensitivity.
Zhang et al. [30]	Probabilistic-Mismatch Anomaly Detection (PMAD) in medical data.	95% accuracy in detecting discrepancies.
Li et al. [36]	Bayesian topic model (MixEHR) for EHRs.	97% accuracy in predicting diagnostic codes and lab tests.
Ni et al. [37]	Bayesian categorical matrix factorization for latent disease detection.	Unveiled eleven concealed diseases.
Albers et al. [38]	PopKLD algorithm for raw lab data representation from EHRs.	Superior prediction of disease stages.

Table 3

Summary of Various Studies in Healthcare Research Across Different Approaches

References	**Method**	**Results**
Garriga et al. [16], Gerasimiuk et al. [67]	ML models on EHRs for patient monitoring, mental health crisis predictions, and MURAL forest method.	Superior visualization and classification accuracy.
Hu [29], Meng et al. [19]	Integration of varied data types with EHRs for tuberculosis and patient risk measures.	Improved diagnostic accuracy.
Rachel and Liao [33], Tang et al. [47]	EHR applications in rheumatology research and mortality prediction, addressing imbalances.	Emphasized need for understanding EHR characteristics.
Nagamine et al. [23], Dai et al. [66], Zhao [44]	NLP and ML models for heart failure categorization, hospitalization prediction, and EHR data retention.	Showcased capabilities in patient categorization and prediction.
Ngata et al. [56], Hurst et al. [49], Ghanzouri et al. [50]	ML and DL models for anomaly detection, security vulnerabilities, and disease diagnosis using EHRs.	Early detection, data protection, and integration into medical practice.
Mahler et al. [24], Colbaugh et al. [20]	Dual-layer design for medical devices and cascade learning for rare disease prediction.	F1 scores: 79-95%, Accuracy: 90.8%, AUC: 93.0%.
Saif et al. [17], Lu and Xue [45]	Hybrid systems for IoT intrusion detection and hybrid model using perceptron algorithms.	Exceptional accuracy and 94% accuracy, respectively.
Buseh et al. [55]	Identification of anomalies in propofol infusion data through traditional and ML methods.	Revealed previously undetected anomalies.
Chen and Yu [21]	FIT system using an ensemble model for ranking medical terms in EHRs.	AUC-ROC values: 0.885 and 0.813 for key terms and relevant terms.
Olwendo et al. [60]	Assessment of EHR data for computational phenotyping of diabetes in Nairobi Hospital.	Found significant influences of software design and data dictionary use.
Sittig et al. [13], Sparapani et al. [43]	Addressing EHR system transition challenges and detecting height anomalies in children's EHR.	Provided six recommendations; R2: 82.2% in training, 75.3% in testing.
Parikh et al. [14], Thompson et al. [18]	Analyzing VA EHR data using CAN score and Methylation-based Risk Score (MRS) for phenotype predictions.	Identified high-risk categories; demonstrated superiority over polygenic scores.
Wesolowski et al. [54], Meng, Zhang, and Chen [41]	Poisson Binomial-based Comorbidity discovery and early warning model for chronic diseases.	Better understanding of cardiovascular health determinants; early warning for chronic diseases.

3.6. Clustering

Rogers et al. [22] and Ta et al. [35] leveraged EHR datasets for optimizing clinical trials and categorizing hospitalized COVID-19 patients, respectively. Their analyses used k-means clustering to pinpoint optimal criteria combinations and reveal patient patterns, emphasizing data-driven improvements in clinical decisions. Lutz [31], Mishra et al. [53], and Moy et al. [48] all applied clustering techniques to EHR data, uncovering meaningful medical structures, enhancing chronic condition predictions, and detecting clinician shifts. Their studies underscored the potential of unsupervised algorithms in health care. Marimuthu and Vaidehi [69] introduced a personalized clustering approach for Remote Health Monitoring, achieving high specificity and sensitivity. Meanwhile, Hackl et al. [57] and Rusanov, Prado, and Weng [64] focused on patient subgroups within heart failure and diabetes conditions using advanced techniques like PCA, UMAP, and LSTM autoencoders. These clustering methods unveiled distinct patient subgroups with unique medical histories and management needs. Christy et al. [65] highlighted the significance of outlier detection, proposing effective cluster-based methods that outperformed distance-based approaches in healthcare datasets.

3.7. Probabilistic Approaches

Zhang et al. [30] introduced the Probabilistic-Mismatch Anomaly Detection (PMAD) to address discrepancies in medical data, with its advanced version, Topical PMAD, achieving 95% accuracy in patient datasets. Li et al. [36] presented MixEHR, a Bayesian topic model for EHRs that projected clinical data onto a meta-phenotype signature, resulting in a promising 97% accuracy rate in predicting diagnostic codes and lab tests. Ni et al. [37] unveiled latent diseases from EHRs using a categorical matrix factorization method and Bayesian techniques, discovering eleven concealed diseases in a Chinese EHR dataset and introducing supportive R tools. Albers et al. [38] delved into representing raw lab data from EHRs, introducing the PopKLD algorithm. This approach created a comprehensive summary, proving superior in predicting disease stages and emphasizing the significance of tailored statistical summaries based on diverse clinical contexts.

3.8. Unsupervised Machine Learning

Garriga et al. [16] and Gerasimiuk et al. [67] enhanced patient monitoring using ML models on EHRs, with the former focusing on mental health crisis predictions and the latter introducing the MURAL forest method for diverse clinical data. Their models displayed superior visualization and classification accuracy. Hu [29] and Meng et al. [19] integrated varied data types with EHRs, achieving improved diagnostic accuracy for tuberculosis and demonstrating connections between latent processes and patient risk measures. Rachel and Liao [33] and Tang et al. [47] emphasized the potential of EHRs for real-world rheumatology research and tackled imbalances in EHRs for mortality prediction. Both studies underscored the need to understand EHR characteristics for accurate results. Nagamine et al. [23], Dai et al. [66], and Zhao [44] showcased the capabilities of NLP and ML models in categorizing heart failure patients, predicting hospitalizations, and retaining sequential nature in EHR data. Ngata et al. [56], Hurst et al. [49], and Ghanzouri et al. [50] demonstrated the application of ML and DL models in anomaly detection, security vulnerabilities, and disease diagnosis using EHRs. They highlighted the potential for early detection, data protection, and integration into medical practice. Habeeb et al. [10] emphasized the importance of real-time anomaly detection in network security, suggesting a swift response to potential threats.

3.9. Fusion Approaches

Mahler et al. [24] introduced a dual-layer design for securing medical devices, achieving 79-95% F1 scores in anomaly detection. Colbaugh et al. [20] applied a cascade learning approach to predict rare diseases in EHRs, registering 90.8% accuracy and 93.0% AUC. Saif et al. [17] and Lu and Xue [45] both targeted healthcare security; the former developed a hybrid system for IoT intrusion detection with exceptional accuracy for various attack classes, while the latter introduced a hybrid model using perceptron algorithms for EHR systems, achieving 94% accuracy. Buseh et al. [55] improved intravenous delivery safety by identifying unique anomalies in propofol infusion data through a blend of traditional and ML methods. Their combination approach revealed anomalies previously undetected by standard methods. Lastly, Chen and Yu [21] tackled information overload in EHRs, using the FIT system to rank medical terms through an ensemble model. Tested against expert-annotated notes, FIT outperformed competitors, scoring AUC-ROC values of 0.885 and 0.813 for key terms and relevant terms, respectively.

3.10. Others

In a study by Olwendo, Ochieng, and Rucha [60], the suitability of EHR data for computational phenotyping of diabetes was assessed in Nairobi Hospital. Influences of software design and use of a data dictionary were found to be significant, contributing 50.7% and 32.3%, respectively to data usability. However, despite 82% of participants appreciating the EHR system, 88% of the extracted data was deemed noise, questioning its appropriateness for diabetes computational phenotyping. Sittig, Lakhani, and Singh [13] addressed the challenges faced by healthcare organizations transitioning between EHR systems, offering six proactive recommendations, including prioritizing communication and assisting employees during transitions. Sparapani et al. [43] proposed a novel model to detect height anomalies in children's EHR with an impressive R2 of 82.2% in training and 75.3% in testing. Parikh et al. [14] analyzed VA EHR data using the CAN score to categorize high-risk Veterans, identifying 30 distinct high-risk categories. Thompson et al. [18] introduced the Methylation-based Risk Score (MRS) for predicting clinical phenotypes, demonstrating its superiority over polygenic risk scores in most cases. Lastly, Wesolowski et al. [54] employed the Poisson Binomial-based Comorbidity discovery (PBC) technique on a massive dataset to understand cardiovascular health determinants better, while Meng, Zhang, and Chen [41] developed an early warning model for chronic diseases using EHR.

4. Analysis and Discussion

The review encompassed 300 papers, displaying diverse study designs. Some are purely data-driven and others merge domain expertise with ML for anomaly detection in EHRs. Post assessment, 80 papers were duplicates, 21 were reviews, and 36 were off-topic. Grey literature accounted for 23 papers, while 5 were in non-primary languages. Workshops contributed 40 papers, with 12 from theses and books. This left 73 pertinent papers for analysis. We illustrated this distribution with a bar chart, labeling categories like “Remaining”, “Duplicate”, and so forth. Each category's paper count is atop its bar, as shown in figure 1. This offers a concise view of the structure of our paper collection.

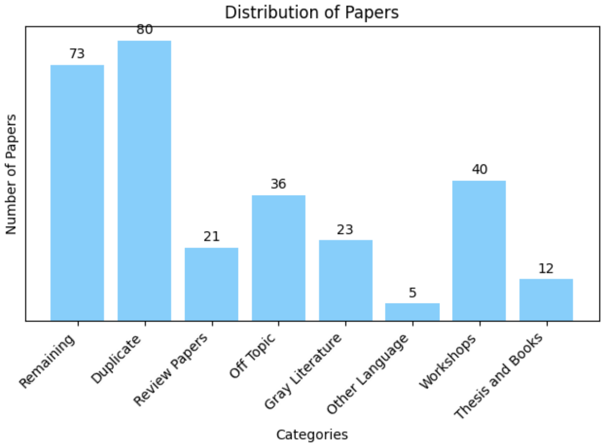

Fig. 1. Distribution of the papers in terms of study selection

The geographic distribution of the included studies showcased a global engagement in unsupervised anomaly detection in EHRs. Studies were sourced from various regions, including a major chunk from North America and Europe, and many publications from Asia and beyond. This global representation signifies the universal relevance of this research area and the need for solutions that transcend geographical boundaries. The geographical distribution can be seen in the figure 2.

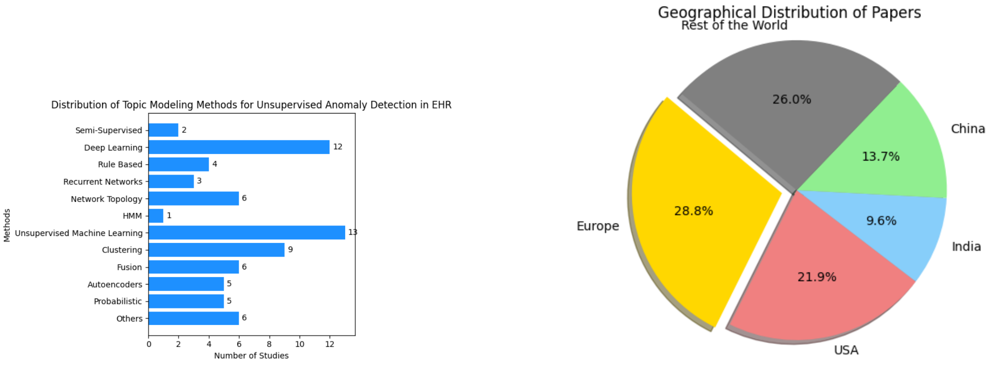

Fig. 2. Distribution of research publications: (a) Topic-wise, (b) Geographical

In unsupervised anomaly detection within EHRs, a literature review reveals diverse subject areas. Deep Learning has been prominent, with 12 studies utilizing neural networks. Unsupervised Machine Learning, which includes techniques like clustering, is adopted in 13 studies, with clustering alone in 9. Rule-based methods are in 4 studies, while Recurrent Networks and Network Topology methods are in 3 and 6, respectively. Autoencoders feature in 5 studies, with Probabilistic models also in 5. Semi-supervised techniques appear in 2 studies, HMM in 1 and other methods in 6. This distribution is visually summarized in a bar chart as shown in fig 2, highlighting the varied approaches to anomaly detection in EHRs.

4.1. Discussion of Findings

The systematic literature review has provided valuable insights into the state of research on unsupervised anomaly detection for early disease identification within EHR. Despite significant advancements in this field, several notable gaps and challenges persist, shedding light on areas that warrant further investigation and refinement. One of the prominent gaps is the absence of standardized evaluation criteria and benchmark datasets for assessing the performance of various anomaly detection methods. While many techniques have been proposed, the lack of consistent evaluation metrics hampers the comparability and reproducibility of results across studies. A standardized framework could facilitate fair comparisons and enable researchers to identify the most effective methods for specific disease identification tasks. The inherent characteristics of EHRs, such as imbalanced and sparse data, pose significant challenges for unsupervised anomaly detection. Most normal data instances can easily overshadow rare and subtle disease patterns. Addressing this issue necessitates the development of methods that can effectively handle imbalanced data distributions and extract meaningful insights from sparse records, enhancing the robustness and reliability of anomaly detection systems. As anomaly detection methods become more sophisticated, the interpretability of their outputs becomes crucial, especially in clinical settings. Many advanced techniques, such as deep learning models, are often viewed as black boxes, hindering their adoption by healthcare professionals. Bridging the gap between complex methodologies and clinical interpretability is imperative to instill trust in anomaly detection results and facilitate their integration into real-world healthcare practices. EHRs inherently capture temporal relationships and the evolution of disease manifestations over time. Existing unsupervised methods often struggle to capture and leverage temporal dynamics in anomaly detection effectively. There is a need for novel approaches to capture sequential patterns, time lags, and temporal correlations to enhance the accuracy of early disease identification across varying time intervals. As the volume of EHR data continues to grow, the scalability of unsupervised anomaly detection methods becomes a critical concern. Efficient processing of large-scale datasets in real-time scenarios is paramount for early disease identification. Developing techniques that can maintain efficacy while scaling to massive datasets and enabling real-time analysis is an essential challenge in this domain. EHRs comprise diverse data types, including structured clinical data, unstructured textual information, and medical images. Integrating and effectively utilizing these heterogeneous data sources in anomaly detection is a complex challenge. Methods that can seamlessly handle and extract insights from disparate data types hold the potential to enhance the accuracy and comprehensiveness of early disease identification significantly.

5. Conclusion

Our systematic review highlighted the significance of unsupervised anomaly detection in EHRs for early disease identification. This research area combines technical and medical expertise, demanding collaboration between data scientists and healthcare professionals. The widespread interest in this domain showcases a global effort to tackle universal healthcare challenges. These adaptable detection methods address various facets of EHR data, from clinical to administrative. Over time, advancements in machine learning and data availability have catalyzed rapid innovation in this field. In summary, merging technological innovation with medical knowledge is pivotal for efficient anomaly detection in EHRs. Our review provides a roadmap for future endeavors to enhance patient care and transform healthcare delivery. In the future, we are exploring the integration of Generative Adversarial Networks and transformer-based approaches, specifically to address the challenge of fall detection within EHR data.

References

1. Šabić Edin Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data / Edin Šabić, David Keeley, Bailey Henderson, Sara Nannemann // AI & SOCIETY. Springer. – 2021. – Vol. 36, №. 1. – P. 149. –158
2. Hirokazu S. Improving supervised outlier detection by unsupervised representation learning and generative adversarial networks: An extension of extreme gradient boosting outlier detection by gans / Shimauchi Hirokazu // Proceedings of the 4th International Conference on Information Science and Systems. – 2021. – P. 22.–27.
3. Shi Xi An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge / Xi Shi, Charlotte Prins, Gijs Van Pottelbergh, Pavlos Mamouris, Bert Vaes, Bart De Moor // BMC Medical Informatics and Decision Making. – 2021. – Vol. 21. – P. 1.–10.
4. Ibrahim Zina M. A knowledge distillation ensemble framework for predicting short-and long-term hospitalization outcomes from electronic health records data / Zina M. Ibrahim, Daniel Bean, Thomas Searle, Linglong Qian, Honghan Wu, Anthony Shek, Zeljko Kraljevic, James Galloway, Sam Norton, TH Teo James and others // IEEE Journal of Biomedical and Health Informatics. – 2021. – Vol. 26, №. 1. – P. 423–435.
5. Dos Santos Henrique Evaluation of a Prescription Outlier Detection System in Hospital's Pharmacy Services / Henrique Dos Santos DP, Ana Helena DPS Ulbrich, Renata Vieira / IEEE International Conference on Bioinformatics and Biomedicine (BIBM). – 2021. – P. 2862–2868.
6. Niu Haoran Adaptive anomaly detection for dynamic clinical event sequences / Haoran Niu, Olufemi A. Omitaomu, Qing C. Cao, Mohammad Olama, Ozgur Ozmen, Hilda Klasky, Laura Pullum, Thakur Malviya Addi,d Teja Kuruganti, Jeanie Scott and others // IEEE International Conference on Big Data (Big Data). – 2020. – P. 4919–4928.
7. Zhang Conghai Medical fraud and abuse detection system based on machine learning / Conghai Zhang, Xinyao Xiao, Chao Wu // International journal of environmental research and public health. – 2020 . – Vol. 17, №. 19. – P. 7265.
8. Seh Adil Hussain Machine learning based framework for maintaining privacy of healthcare data / Hussain Seh Adil, Jehad F. Al-Amri, Ahmad F. Subahi, Alka Agrawal, Rajeev Kumar, Ahmad Khan Raees // Intell. Autom. Soft Comput. – 2021. – Vol. 29. – P. 697–712.
9. You Sungmin Unsupervised automatic seizure detection for focal-onset seizures recorded with behind-the-ear EEG using an anomaly-detecting generative adversarial network / Sungmin You, Hwan Cho Baek, Soonhyun Yook, Young Kim Joo, Young-Min Shon, Dae-Won Seo, Young Kim In // Computer Methods and Programs in Biomedicine. – 2020. – Vol. 193. – P. 105472.
10. Habeeb Riyaz Ahamed Ariyaluran Real-time big data processing for anomaly detection: A survey / Riyaz Ahamed Ariyaluran Habeeb, Fariza Nasaruddin, Abdullah Gani, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, Muhammad Imran, // International Journal of Information Management . – 2019 . – Vol. 45 . – P. 289–307 Elsevier
11. Khan Arif Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes / Arif Khan, Shahadat Uddin, Uma Srinivasan // Expert Systems with Applications. – 2019. – Vol. 136. – P. 230–241.

12. Chen Yurong MAMA Net: Multi-scale attention memory autoencoder network for anomaly detection Yurong Chen, Hui Zhang, Yaonan Wang, Yimin Yang, Xianen Zhou, QM Jonathan Wu // IEEE Transactions on Medical Imaging. – 2020. – Vol. 40, №. 3. – P. 1032–1041.
13. Sittig Dean F. Applying requisite imagination to safeguard electronic health record transitions / Dean F. Sittig, Priti Lakhani, Hardeep Singh // Journal of the American Medical Informatics Association. – Oxford University Press. – 2022. – Vol. 29, №. 5. – P. 1014–1018.
14. Parikh, Ravi B. A machine learning approach to identify distinct subgroups of veterans at risk for hospitalization or death using administrative and electronic health record data / Ravi B. Parikh, Kristin A. Linn, Jiali Yan, Matthew L. Maciejewski, Ann-Marie Rosland, Kevin G. Volpp, Peter W. Groeneveld, Amol S. Navathe // Plos one. – Public Library of Science San Francisco, CA USA. – 2021. – Vol. 16, №. 2. – P. e0247203.
15. Tomašev, Nenad Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records / Nenad Tomašev, Natalie Harris, Sebastien Baur, Anne Mottram, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Valerio Magliulo and others // Nature Protocols. – Nature Publishing Group UK London. – 2021. – Vol. 16, №. 6. – P. 2765–2787.
16. Garriga Roger Machine learning model to predict mental health crises from electronic health records / Roger Garriga, Javier Mas, Semhar Abraha, Jon Nolan, Oliver Harrison, George Tadros, Aleksandar Matic // Nature medicine. – Nature Publishing Group US New York. – 2022. – Vol. 28, №. 6. – P. 1240–1248.
17. Saif Sohail HIIDS: Hybrid intelligent intrusion detection system empowered with machine learning and metaheuristic algorithms for application in IoT based healthcare / Sohail Saif, Priya Das, Suparna Biswas, Manju Khari, Vimal Shanmuganathan // Microprocessors and Microsystems. – Elsevier. – 2022. – P. 104622.
18. Thompson Mike Methylation risk scores are associated with a collection of phenotypes within electronic health record systems / Mike Thompson, Brian L. Hill, Nadav Rakocz, Jeffrey N. Chiang, Daniel Geschwind, Sriram Sankararaman, Ira Hofer, Maxime Cannesson, Noah Zaitlen, Eran Halperin // NPJ genomic medicine. – Nature Publishing Group UK London. – 2022. – Vol. 7, №. 1. – P. 50.
19. Meng Rui Nonstationary multivariate Gaussian processes for electronic health records / Rui Meng, Braden Soper, Herbert KH. Lee, Vincent X. Liu, John D. Greene, Priyadip Ray // Journal of Biomedical Informatics. – Elsevier. – 2021. – Vol. 117. – P. 103698.
20. Colbaugh Rich Learning to identify rare disease patients from electronic health records / Rich Colbaugh, Kristin Glass, Christopher Rudolf, Mike Tremblay Volv Global // AMIA annual symposium proceedings. – American Medical Informatics Association. – 2018. – Vol. 2018. – P. 340.
21. Chen Jinying Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients / Jinying Chen, Hong Yu // Journal of biomedical informatics. – Elsevier. – 2017. – Vol. 68. – P. 121–131.
22. Rogers James R. Leveraging electronic health record data for clinical trial planning by assessing eligibility criteria's impact on patient count and safety / James R. Rogers, Jovana Pavisic, Casey N. Ta, Cong Liu, Ali Soroush, Ying Kuen Cheung, George Hripcsak, Chunhua Weng, // Journal of Biomedical Informatics. – Elsevier. – 2022. – Vol. 127. – P. 104032.
23. Nagamine Tasha Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data / Tasha Nagamine, Brian Gillette, Alexey Pakhomov, John Kahoun, Hannah Mayer, Rolf Burghaus, Jörg Lippert, Mayur Saxena // Scientific Reports. – Springer. – 2020. – Vol. 10, №. 1. – P. 1–13.
24. Mahler Tom A dual-layer context-based architecture for the detection of anomalous instructions sent to medical devices / Tom Mahler, Erez Shalom, Yuval Elovici, Yuval Shahar // Artificial Intelligence in Medicine. – Elsevier. – 2022. – Vol. 123. – P. 102229.

25. Bala P. Manju Applications of Machine Learning and Deep Learning for maintaining Electronic Health Records / P. Manju Bala, S. Usharani, R. Rajmohan, G. Leemaroselin // Global Journal of Internet Interventions and IT Fusion [ISSN: 2582-1385 (online)]. – 2020. – Vol. 3.
26. Chushig-Muzo David Learning and visualizing chronic latent representations using electronic health records / David Chushig-Muzo, Cristina Soguero-Ruiz, Pablo de Miguel Bohoyo, Inmaculada Mora-Jiménez and others // BioData Mining. – BioMed Central. – 2022. – Vol. 15, №. 1. – P. 1–27.
27. Xu Zhen Mufasa: Multimodal fusion architecture search for electronic health records / Zhen Xu, David R. So, Andrew M. Dai // Proceedings of the AAAI Conference on Artificial Intelligence. – 2021. – Vol. 35, №. 12. – P. 10532–10540.
28. Chen You Building bridges across electronic health record systems through inferred phenotypic topics / You Chen, Joydeep Ghosh, Cosmin Adrian Bejan, Carl A. Gunter, Siddharth Gupta, Abel Kho, David Liebovitz, Jimeng Sun, Joshua Denny, Bradley Malin // Journal of biomedical informatics. – Elsevier. – 2015. – Vol. 55. – P. 82–93.
29. Hu Xuejiao Integrating exosomal microRNAs and electronic health data improved tuberculosis diagnosis / Xuejiao Hu, Shun Liao, Hao Bai, Lijuan Wu, Minjin Wang, Qian Wu, Juan Zhou, Lin Jiao, Xuerong Chen, Yanhong Zhou and others // EBioMedicine. – Elsevier. – 2019. – Vol. 40. – P. 564–573.
30. Zhang Lingxiao Probabilistic-mismatch anomaly detection: do one's medications match with the diagnoses / Lingxiao Zhang, Xiang Li, Haifeng Liu, Jing Mei, Gang Hu, Junfeng Zhao, Yanzhen Zou, Bing Xie, Guotong Xie / IEEE 16th International Conference on Data Mining (ICDM). – 2016. – P. 659–668.
31. Lütz Elin Unsupervised machine learning to detect patient subgroups in electronic health records / Elin Lütz. – 2019.
32. Röchner Philipp Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries / Philipp Röchner, Franz Rothlauf // BMC Medical Research Methodology. – Springer. – 2023. – Vol. 23, №. 1. – P. 125.
33. Knevel Rachel From real-world electronic health record data to real-world results using artificial intelligence / Rachel Knevel, Katherine P. Liao // Annals of the Rheumatic Diseases. – BMJ Publishing Group Ltd. – 2023. – Vol. 82, №. 3. – P. 306–311.
34. Li Runze Improving an Electronic Health Record. –Based Clinical Prediction Model Under Label Deficiency: Network-Based Generative Adversarial Semisupervised Approach / Runze Li, Yu Tian, Zhuyi Shen, Jin Li, Jun Li, Kefeng Ding, Jingsong Li and others // JMIR Medical Informatics. – JMIR Publications Inc., Toronto, Canada. – 2023. – Vol. 11, №. 1. – P. e47862.
35. Ta Casey N. Clinical and temporal characterization of COVID-19 subgroups using patient vector embeddings of electronic health records / Casey N. Ta, Jason E. Zucker, Po-Hsiang Chiu, Yilu Fang, Karthik Natarajan, Chunhua Weng // Journal of the American Medical Informatics Association. – Oxford University Press. – 2023. – Vol. 30, №. 2. – P. 256–272.
36. Li Yue Inferring multimodal latent topics from electronic health records / Yue Li, Pratheeksha Nair, Xing Han Lu, Zhi Wen, Yuening Wang, Amir Ardalan Kalantari Dehaghi, Yan Miao, Weiqi Liu, Tamas Ordog, Joanna M. Biernacka and others // Nature communications. – Nature Publishing Group UK London. – 2020. – Vol. 11, №. 1. – P. 2536.
37. Ni Yang Bayesian double feature allocation for phenotyping with electronic health records / Yang Ni, Peter Müller, Yuan Ji // Journal of the American Statistical Association. – Taylor & Francis. – 2020. – Vol. 115, №. 532. – P. 1620–1634.
38. Albers, David J. Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms /Ḋavid J. Albers, Noémie Elhadad, Jan Claassen, Rimma Perotte, A. Goldstein, George Hripcsak // Journal of biomedical informatics. – Elsevier. – 2018. – Vol. 78. – P. 87–101.

39. Ashfaq Awais Readmission prediction using deep learning on electronic health records / Awais Ashfaq, Anita Sant'Anna, Markus Lingman, Sławomir Nowaczyk // Journal of biomedical informatics. – Elsevier. – 2019. – Vol. 97. – P. 103256.
40. Boussina Aaron Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study / AaronBoussina, Gabriel Wardi, Supreeth Prajwal Shashikumar, Atul Malhotra, Kai Zheng, Shamim Nemati // Journal of Medical Internet Research. – JMIR Publications Toronto, Canada. – 2023. – Vol. 25. – P. e45614.
41. Meng Jie Utilizing narrative text from electronic health records for early warning model of chronic disease / Jie Meng, Runtong Zhang, Donghua Chen // IEEE International Conference on Systems, Man, and Cybernetics (SMC). – 2018. – P. 4210–4215.
42. Niu Haoran Detecting anomalous sequences in electronic health records using higher-order tensor networks / Haoran Niu, Olufemi A. Omitaomu, Michael A. Langston, Mohammad Olama, Ozgur Ozmen, Hilda B. Klasky, Angela Laurio, Brian Sauer, Merry Ward, Jonathan Nebeker // Journal of Biomedical Informatics. – Elsevier. – 2022. – Vol. 135. – P. 104219.
43. Sparapani Rodney A. Novel pediatric height outlier detection methodology for electronic health records via machine learning with monotonic Bayesian additive regression trees / Rodney A. Sparapani, Bi Q. Teng, Julia Hilbrands, Rebecca Pipkorn, Mary Beth Feuling, Praveen S. Goday // Journal of Pediatric Gastroenterology and Nutrition. – Wolters Kluwer. – 2022. – Vol. 75, №. 2. – P. 210–214.
44. Zhao Jing Learning from heterogeneous temporal data in electronic health records / Jing Zhao, Panagiotis Papapetrou, Lars Asker, Henrik Boström // Journal of biomedical informatics. – Elsevier. – 2017. – Vol. 65. – P. 105–119.
45. Lu Wei A perceptron mixture model of intrusion detection for safeguarding electronic health record system / Wei Lu, Ling Xue // Advances in Networked-Based Information Systems: The 24th International Conference on Network-Based Information Systems (NBiS-2021). – Springer. – 2022. – P. 202–212.
46. Rashidian Sina Detecting miscoded diabetes diagnosis codes in electronic health records for quality improvement: temporal deep learning approach / Sina Rashidian, Kayley Abell-Hart, Janos Hajagos, Richard Moffitt, Veena Lingam, Victor Garcia, Chao-Wei Tsai, Fusheng Wang, Xinyu Dong, Siao Sun and others // JMIR medical informatics. – JMIR Publications Inc., Toronto, Canada. – 2020. – Vol. 8, №. 12. – P. e22649.
47. Tang Jianxiang Joint modeling strategy for using electronic medical records data to build machine learning models: an example of intracerebral hemorrhage / Jianxiang Tang, Xiaoyu Wang, Hongli Wan, Chunying Lin, Zilun Shao, Yang Chang, Hexuan Wang, Yi Wu, Tao Zhang, Yu Du // BMC Medical Informatics and Decision Making. – Springer. – 2022. – Vol. 22, №. 1. – P. 1–13.
48. Moy, Amanda J. Using Time Series Clustering to Segment and Infer Emergency Department Nursing Shifts from Electronic Health Record Log Files / Amanda J. Moy, Kenrick D. Cato, Jennifer Withall, Eugene Y. Kim, Nicholas Tatonetti, Sarah C. Rossetti, // AMIA Annual Symposium Proceedings. – American Medical Informatics Association. – 2022. – Vol. 2022. – P. 805.
49. Hurst William Securing electronic health records against insider-threats: A supervised machine learning approach / William Hurst, Bedir Tekinerdogan, Tarek Alskaif, Aaron Boddy, Nathan Shone // Smart Health. – Elsevier. – 2022. – Vol. 26. – P. 100354.
50. Ghanzouri I. Performance and usability testing of an automated tool for detection of peripheral artery disease using electronic health records / I. Ghanzouri, S. Amal, V. Ho, L. Safarnejad, J. Cabot, CG. Brown-Johnson, N. Leeper, S. Asch, NH. Shah, EG. Ross // Scientific Reports. – Nature Publishing Group UK London. – 2022. – Vol. 12, №. 1. – P. 13364.
51. Ma He Connections between various disorders: combination pattern mining using apriori algorithm based on diagnosis Information from electronic medical records / He Ma, Jingjing

Ding, Mei Liu, Ying Liu and others // BioMed Research International. – Hindawi. – 2022. – Vol. 2022.

52. Rijcken Emil Exploring Embedding Spaces for more Coherent Topic Modeling in Electronic Health Records / Emil Rijcken, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, Floortje Scheepers, Uzay Kaymak // IEEE International Conference on Systems, Man, and Cybernetics (SMC). – 2022. – P. 2669–2674.
53. Mishra Sushruta A decisive metaheuristic attribute selector enabled combined unsupervised-supervised model for chronic disease risk assessment / Sushruta Mishra, Hiren Kumar Thakkar, Priyanka Singh, Gajendra Sharma // Computational Intelligence and Neuroscience. – Hindawi. – 2022. – Vol. 2022.
54. Wesołowski Sergiusz An explainable artificial intelligence approach for predicting cardiovascular outcomes using electronic health records / Sergiusz Wesołowski, Gordon Lemmon, Edgar J. Hernandez, Alex Henrie, Thomas A. Miller, Derek Weyhrauch, Michael D. Puchalski, Bruce E. Bray, Rashmee U. Shah, Vikrant G. Deshmukh and others // PLOS digital health. – Public Library of Science San Francisco, CA USA. – 2022. – Vol. 1, №. 1. – P. e0000004.
55. Obuseh Marian Detecting Unusual Intravenous Infusion Alerting Patterns with Machine Learning Algorithms / Marian Obuseh, Denny Yu, Poching DeLaurentis // Biomedical Instrumentation & Technology. – 2022. – Vol. 56, №. 2. – P. 58–70.
56. Nagata Kenichiro Detection of overdose and underdose prescriptions—An unsupervised machine learning approach / Kenichiro Nagata, Toshikazu Tsuji, Kimitaka Suetsugu, Kayoko Muraoka, Hiroyuki Watanabe, Akiko Kanaya, Nobuaki Egashira, Ichiro Ieiri // PloS one. – Public Library of Science San Francisco, CA USA. – 2021. – Vol. 16, №. 11. – P. e0260315.
57. Hackl Melanie Unsupervised Learning to Subphenotype Heart Failure Patients from Electronic Health Records / Melanie Hackl, Suparno Datta, Riccardo Miotto, Erwin Bottinger // Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15. –18, 2021, Proceedings. – Springer. – 2021. – P. 219–228.
58. Wang Yu A maintenance hemodialysis mortality prediction model based on anomaly detection using longitudinal hemodialysis data / Yu Wang, Yilin Zhu, Guofeng Lou, Ping Zhang, Jianghua Chen, Jingsong Li // Journal of Biomedical Informatics. – Elsevier. – 2021. – Vol. 123. – P. 103930.
59. Li Yang Towards Interpretability and Personalization: A Predictive Framework for Clinical Time-series Analysis / Yang Li, Xianli Zhang, Buyue Qian, Zeyu Gao, Chong Guan, Yefeng Zheng, Hansen Zheng, Fenglang Wu, Chen Li // IEEE International Conference on Data Mining (ICDM). – 2021. – P. 340–349.
60. Olwendo Amos Otieno Suitability of Electronic Health Record Data for Computational Phenotyping of Diabetes Mellitus at Nairobi Hospital, Nairobi City County, Kenya / Amos OtienoOlwendo, George Ochieng, Kenneth Rucha and others // East African Journal of Science, Technology and Innovation. – 2021. – Vol. 2, №. 2.
61. Hou Rui Anomaly Detection of Calcifications in Mammography Based on 11,000 Negative Cases / Rui Hou, Yifan Peng, Lars J. Grimm, Yinhao Ren, Maciej A. Mazurowski, Jeffrey R. Marks, Lorraine M. King, Carlo C. Maley, E. Shelley Hwang, Joseph Y. Lo // IEEE Transactions on Biomedical Engineering. – 2021. – Vol. 69, №. 5. – P. 1639–1650.
62. Alhassan Zakhriya Improving current glycated hemoglobin prediction in adults: Use of machine learning algorithms with electronic health records / Zakhriya Alhassan, Matthew Watson, David Budgen, Riyad Alshammari, Ali Alessa, Noura Al Moubayed and others // JMIR Medical Informatics. – JMIR Publications Inc., Toronto, Canada. – 2021. – Vol. 9, №. 5. – P. e25237.
63. Manne Ravi Application of artificial intelligence in healthcare: chances and challenges / Ravi Manne, Sneha C. Kantheti, // Current Journal of Applied Science and Technology. – 2021. – Vol. 40, №. 6. – P. 78–89.

64. Rusanov Alexander Unsupervised time-series clustering over lab data for automatic identification of uncontrolled diabetes / Alexander Rusanov, Patric V. Prado, Chunhua Weng // IEEE International Conference on Healthcare Informatics (ICHI). – 2016. – P. 72–80.
65. Christy A. Cluster based outlier detection algorithm for healthcare data / A. Christy, G. Meera Gandhi, S. Vaithyasubramanian // Procedia Computer Science. – Elsevier. – 2015. – Vol. 50. – P. 209–215.
66. Dai Wuyang Prediction of hospitalization due to heart diseases by supervised learning methods / Wuyang Dai, Theodora S. Brisimi, William G. Adams, Theofanie Mela, Venkatesh Saligrama, Ioannis Ch. Paschalidis // International journal of medical informatics. – Elsevier. – 2015. – Vol. 84, №. 3. – P. 189–197.
67. Gerasimiuk Michal MURAL: An Unsupervised Random Forest-Based Embedding for Electronic Health Record Data / Michal Gerasimiuk, Dennis Shung, Alexander Tong, Adrian Stanley, Michael Schultz, Jeffrey Ngu, Loren Laine, Guy Wolf, Smita Krishnaswamy book2021 IEEE International Conference on Big Data (Big Data). – 2021. – P. 4694–4704.
68. Ramos Guilherme Unsupervised learning approach for predicting sepsis onset in ICU patients / Guilherme Ramos, Erida Gjini, Luis Coelho, Margarida Silveira // 43rd annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC). – IEEE. – 2021. – P. 1916–1919.
69. Marimuthu Poorani An unsupervised approach for personalized RHM with reduced mean alert latency / Poorani Marimuthu, V. Vaidehi // Journal of Intelligent & Fuzzy Systems. – IOS Press, №. Preprint. – P. 1–18.
70. Akshay Kumaar M. A Hybrid Framework for Intrusion Detection in Healthcare Systems Using Deep Learning / M. Akshay Kumaar, Duraimurugan Samiayya, PM. Vincent, Kathiravan Srinivasan, Chuan-Yu Chang, Harish Ganesh // Frontiers in Public Health. – Frontiers. – 2022. – Vol. 9. – P. 2295.

Article: received 10.10.2023
revised 15.04.2024
printing adoption 10.09.2024

СИСТЕМАТИЧНИЙ ОГЛЯД СПОСОБІВ ПОШУКУ АНОМАЛІЙ ДЛЯ РАННЬОГО ВИЗНАЧЕННЯ ЗАХВОРЮВАНЬ ШЛЯХОМ АНАЛІЗУ ЕЛЕКТРОННО-МЕДИЧНИХ ЗАПИСІВ

І. Урсул

*Львівський національний університет імені Івана Франка,
вул. Університетська 1, Львів, 79000, Україна
e-mail: ivan.ursul@lnu.edu.ua*

Електронні медичні записи (ЕМЗ) стали ключовими репозиторіями даних пацієнтів, володіючи величезним потенціалом для радикального перетворення охорони здоров'я шляхом сприяння ранньому виявленню захворювань. В центрі цього перетворення є застосування методів некерованого виявлення аномалій, які унікально дають змогу виявляти приховані зразки та невідповідності в складних багатовимірних наборах даних. Подано вичерпний систематичний огляд різних методологій, які використовують у некерованому виявленні аномалій, зосереджуючись на ранньому виявленні захворювань у ЕМЗ. Наш аналіз охоплює багато географічних регіонів, відображаючи глобальні зусилля дослідників у розробці рішень у різних сферах охорони здоров'я. Це географічне різноманіття засвідчує універсальну актуальність та адаптивність методів некерованого виявлення аномалій у медицині. Ми дослідили різні предметні області, від клінічних діагнозів до адміністративних процесів, демонструючи універсальність цих методів та широту їх застосувань. Методології, використані в цих дослідженнях, різні та інноваційні, демонструють еволюційний характер цієї галузі. Це різноманіття ілюструє потенціал для міждисциплінарної співпраці між науковцями та медичними фахівцями, а також наголошує на необхідності такого поєднання для ефективного вирішення складних медичних викликів. Наш часовий аналіз виявляє динамічну дослідницьку область, яка формується швидкими досягненнями в техніках машинного навчання та зростаючою доступністю великомасштабних наборів даних. Огляд також висвітлює виклики, притаманні впровадженню цих технік в ЕМЗ, такі як забезпечення конфіденційності даних, підтримання якості даних, інтерпретація моделей машинного навчання. Ми пропонуємо потенційні рішення та напрями майбутніх досліджень для подолання цих перешкод. Крім того, обговорюємо інтеграцію методів некерованого виявлення аномалій в існуючі медичні практики, враховуючи виклики та можливості, які ця інтеграція представляє. Ба більше, пропонуємо напрями майбутніх досліджень, враховуючи дослідження передових технік, таких як генеративні змагальні мережі (GANs) та моделі на основі трансформерів. Ці майбутні перспективи є важливими для просування галузі та максимізації потенціалу ЕМЗ у покращенні догляду за пацієнтами. Наводячи випадкові дослідження та приклади з розглянутої літератури, ми ілюструємо практичне застосування та результати цих методів у реальних умовах. Запропонований систематичний огляд надає комплексний огляд траєкторії досліджень у галузі некерованого виявлення аномалій у ЕМЗ. Він наголошує на важливості поєднання технічної експертизи зі специфічними знаннями у цій критичній галузі, висвітлюючи глобальні тенденції, предметні області та часові моделі. Ця праця є інструментальною для дослідників і практиків, які прагнуть використовувати некероване виявлення аномалій для раннього виявлення захворювань, значно поліпшуючи догляд за пацієнтами та загальні системи охорони здоров'я.

Ключові слова: аномалія, некерований спосіб визначення, захворювання, системний огляд літератури, медичні записи, електронно-медичні записи, машинне навчання, глибоке навчання.