BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Anomaly-based threat detection in smart health using machine learning

Muntaha Tabassum[1†], Saba Mahmood[1*†], Amal Bukhari[2†], Bader Alshemaimri[3†], Ali Daud[4*†] and Fatima Khalique[5†]

Abstract

Background Anomaly detection is crucial in healthcare data due to challenges associated with the integration of smart technologies and healthcare. Anomaly in electronic health record can be associated with an insider trying to access and manipulate the data. This article focuses around the anomalies under different contexts.

Methodology This research has proposed methodology to secure Electronic Health Records (EHRs) within a complex environment. We have employed a systematic approach encompassing data preprocessing, labeling, modeling, and evaluation. Anomalies are not labelled thus a mechanism is required that predicts them with greater accuracy and less false positive results. This research utilized unsupervised machine learning algorithms that includes Isolation Forest and Local Outlier Factor clustering algorithms. By calculating anomaly scores and validating clustering through metrics like the Silhouette Score and Dunn Score, we enhanced the capacity to secure sensitive healthcare data evolving digital threats. Three variations of Isolation Forest (IForest)models (SVM, Decision Tree, and Random Forest) and three variations of Local Outlier Factor (LOF) models (SVM, Decision Tree, and Random Forest) are evaluated based on accuracy, sensitivity, specificity, and F1 Score.

Results Isolation Forest SVM achieves the highest accuracy of 99.21%, high sensitivity (99.75%) and specificity (99.32%), and a commendable F1 Score of 98.72%. The Isolation Forest Decision Tree also performs well with an accuracy of 98.92% and an F1 Score of 99.35%. However, the Isolation Forest Random Forest exhibits lower specificity (72.84%) than the other models.

Conclusion The experimental results reveal that Isolation Forest SVM emerges as the top performer showcasing the effectiveness of these models in anomaly detection tasks. The proposed methodology utilizing isolation forest and SVM produced better results by detecting anomalies with less false positives in this specific EHR of a hospital in North England. Furthermore the proposal is also able to identify new contextual anomalies that were not identified in the baseline methodology.

Keywords Healthcare, Anomaly detection, Insider threats, Electronic Health Records(EHRs), Machine learning

†Muntaha Tabassum, Saba Mahmood, Amal Bukhari, Bader Alshemaimri, Ali Daud and Fatima Khalique contributed equally to this work.

*Correspondence:
Saba Mahmood
smahmood.buic@bahria.edu.pk
Ali Daud
alimsdb@gmail.com
Full list of author information is available at the end of the article

Introduction

Smart cities and healthcare systems are crucial due to the rising global population, integrating advanced technologies to improve urban living, resource management, infrastructure efficiency, and sustainability. Smart Cities are transforming global cities to improve efficiency, sustainability, and safety [1].

Smart healthcare enhances citizens well-being and contributes to resilience. It uses interconnected devices, health data analytics, and real-time monitoring to identify health threats, optimize resource allocation, and provide efficient services [2]. Smart healthcare technology, addressing aging and chronic diseases, offers personalized care, reduces facility burden, and improves patient outcomes by enabling remote support.

Smart healthcare and smart cities are collaborating to promote holistic well-being through data-driven insights and predictive models. This synergy enables intelligent, health-conscious urban living, fostering medical advancements and a resilient urban ecosystem, as cities evolve into innovation hubs. EHRs revolutionize healthcare systems by providing instant access to vital medical information, enabling authorized users to examine and edit patient records from any location thereby enhancing efficiency as depicted in the Fig. 1.

Electronic Medical Records(EMR) save costs, standardize treatment, and improve disease monitoring, with 96% [3] of non-federal American healthcare providers adopting them for communication and information sharing in smart healthcare systems.

EHR deployment has significantly transformed healthcare governance, improving patient data quality and scope. It consolidates medical history, prescriptions, and therapies, enhancing communication, reducing errors, and promoting collaborative treatment.

The rapid implementation of EHRs has raised threats of confidentiality and security of healthcare data. Healthcare information system administrators should prioritize patient privacy and data security, implementing encryption and access restrictions.

Healthcare organizations are increasingly reliant on digital platforms and networked technology, leading to cyberattacks on patient data. Recent research [4] discusses the escalating security and privacy threats in the healthcare industry, tracing back to the HITECH Act of 2009 and highlighting the importance of transparency in

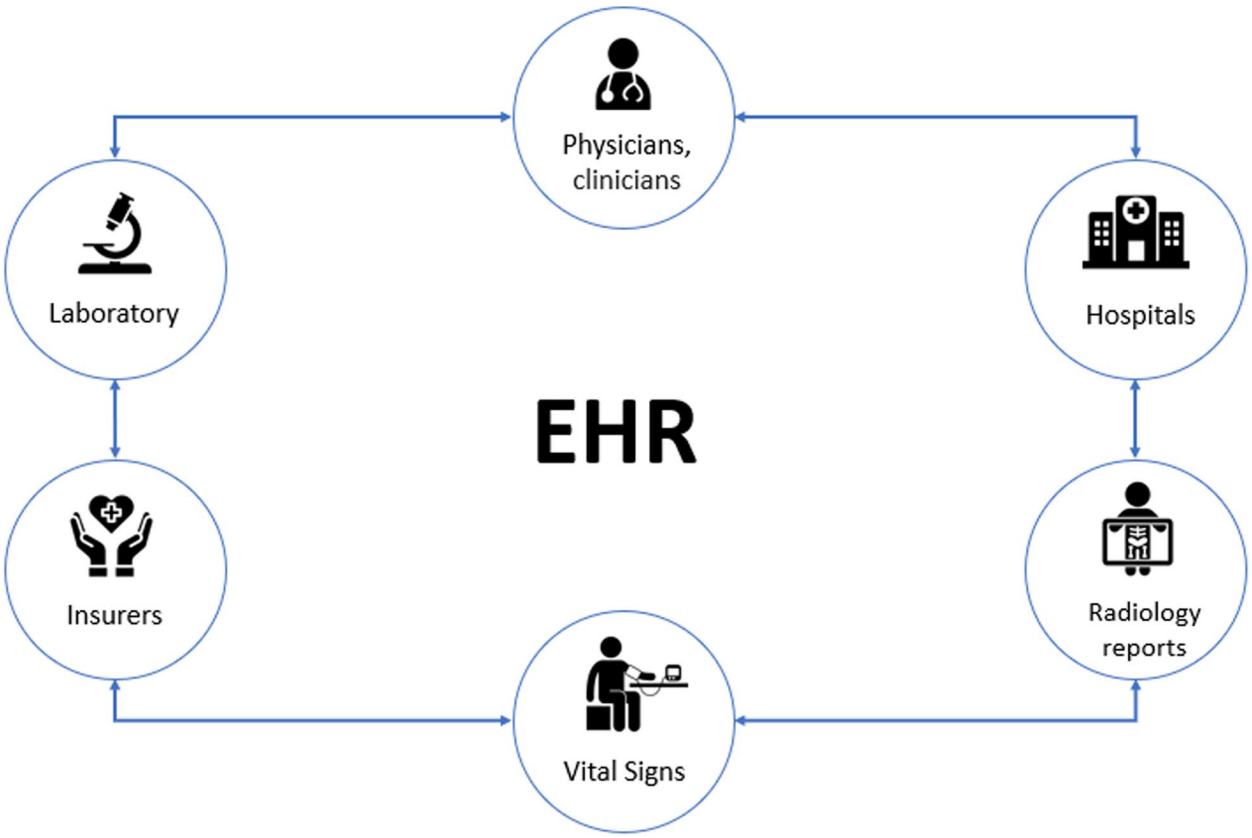

Fig. 1 Electronic health record

addressing vulnerabilities and empowering consumers. Cyber attacks can lead to identity theft, financial issues, and injuries. Ransomware encryption and disruption of vital functions pose threats. Advanced cybersecurity is crucial for healthcare infrastructure, and network resilience is essential to mitigate system-wide damage. El-Bakkouri and Mazri [5] report explores the impact of IoT based healthcare, highlighting its benefits and cyber security risks due to reliance on IT software and wireless networks.

Healthcare cyber risks require a comprehensive strategy involving machine learning for detection. A patient-centered design is crucial for data preservation, while machine learning models enhance detection accuracy and efficiency. This approach creates a safer digital healthcare environment for patients and stakeholders. Anomaly in the EHR if remains unnoticed can create problems for the patient and health practitioners. Anomaly due to security risk arises when an unauthorized personnel gets access to the EHR due to stolen or leaked passwords. It can be due to an insider having privileged access manipulating the data for certain gains.

Insider attack pose a significant threat to healthcare systems, causing data breaches and compromising patient outcomes. These threats can be malevolent or ignorant and can affect workers, contractors, and anyone with access. To prevent these threats, technical measures, staff training, and strict access restrictions are needed. Insider threats in healthcare involve authorized access to EHR systems leading to misuse of patient data and identity theft. The Fig. 2 shows different types of insider threats. An insider threat can be from different types of users that includes an insider employee of the hospital with malicious or criminal intentions, attacker that has stolen the credentials of a valid user or a negligent user who made mistakes in the hospital records. Machine learning algorithms can identify such dangerous user behavior patterns. An effective detection system can prevent patient data from threats.

Insider threat detection is a crucial process for organizations to protect themselves from potential internal threats. There are three approaches towards identification of threats that includes, signature based methods, rule based methods. Both of these techniques utilize predefined patterns and rules for the identification of any threat. Both methods require careful planning and constant updates to avoid false results.The third approach anomaly-based detection identifies deviations from normal system or user behavior but faces challenges distinguishing between malicious and legitimate changes. Continuous learning and adaptation are essential for effectiveness in dynamic environments.

Anomaly detection systems aid doctors by detecting anomalies in patient data, predicting health risks, and requiring advanced anomaly detection models for continuous surveillance [6].

Machine Learning-based insider threat detection systems are influenced by data quality and quantity. The challenge of machine learning models in anomaly based insider threat detection includes producing accurate results with low false positives. This study has proposed a unsupervised and supervised machine learning methodology in the identification of anomaly based insider threats. The data is unlabeled, thus we have proposed a

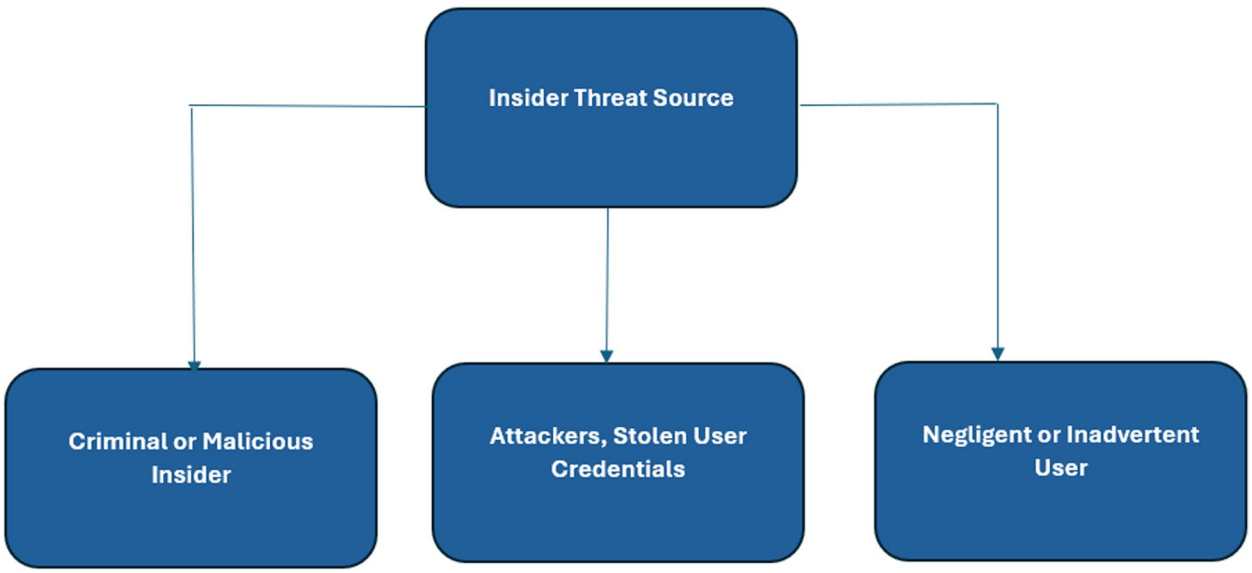

Fig. 2 Insider threat sources

combination of unsupervised and supervised model that can predict the anomalies effectively. The research is conducted on authentic EHR data from a healthcare organization. This study aims to enhance healthcare data security by utilizing machine learning techniques to strengthen Electronic Health Records against inside threats. It aims to improve patient privacy and data security, establishing robust mechanisms to protect sensitive patient information. Following are the key contributions in this research.

- A methodology that is based on unsupervised and supervised machine learning algorithms for the identification of contextual anomalies with better performance.
- The proposed methodology is able to identify some new contextual anomalies that remained unidentified in the previous approach.
- The research also produced insightful recommendation of application of cross correlation for feature selection on the dataset.

This paper consists of "Related work" section that provides discussion of the existing literature, setting the stage for the study. In "Methodology" section, we have discussed the proposed methodology followed by "Experiment and results" section related to Analysis and Results, that discussed performance evaluation and system parameters, accuracy, performance, and results. In the last "Comparison with baseline" section, the Conclusion includes the summary of the research and future work.

Related work

Accessing electronic health records (EHRs) is vital for effective diagnosis and treatment by offering a complete patient history. Anomaly detection through AI in smart healthcare can detect unusual patterns in data, alerting providers to potential health issues in real time, ultimately enhancing patient care and cutting costs.

Smart cities

A literature survey by [7] highlights the importance of smart cities in addressing the challenges of rapid urbanization. It analyzes origins, definitions, application domains, architectures, enabling technologies, and recent research, offering guidance for researchers in the field. Exploring smart cities, the article defines them as integrating ICT with traditional infrastructures [8]. It outlines seven goals and six research challenges, highlighting operational functioning, innovation, equity, and mobility. It elucidates the state of the art, presents scenarios, proposes project areas, anticipates paradigm shifts, and suggests key demonstrators to advance smart city science. Smart cities blend technology, governance, and society using IoT and AI to enhance various sectors. Research by [9] highlights the significance of small cities, exemplified by three cases in Finland, emphasizing public sector involvement in ecosystem-based development. Health ecosystem involves utilization of analytics techniques aided with expert opinion for decisions and planning. Recently researcher have analyzed the digital healthcare ecosystem [10, 11] for diseases like parkinsons,diabetes and etc.

Smart health

In healthcare technology, smart sensors, especially wearables, facilitate collaboration between health professionals and patients. Article [12], examines their impact on healthcare monitoring systems, highlighting machine learning's role in recognizing human activity. It presents a smart healthcare system using machine learning for activity recognition, addressing challenges, and demonstrating applications across various datasets, including mobile phones and wearables. Another research [13] analyzed patient data in IOMT for data privacy. Smart healthcare, driven by IoT, big data, cloud computing, and AI, revolutionizes medicine. Research explores foundational technologies, current applications, challenges, solutions, and the promising future of smart healthcare [14]. IoT and machine learning revolutionize healthcare through Implantable and Wearable Medical Devices (IWMDs), enabling continuous data collection and analysis. Machine learning deciphers patterns for predictive healthcare. Article [15, 16] explores challenges in smart healthcare design and implementation, aiming for a standardized framework and highlighting critical issues.

Smart health in smart cities

Advancements in 5G and IoT have led to smart cities, impacting areas like traffic systems and healthcare. This article discusses a smart ambulance navigation system, integrating patient monitors and tracking for rapid transmission of critical data to hospitals, enhancing emergency response and patient care efficiency [17, 18]. Healthcare is evolving with smart tech like AI, ML, IoT, and 6G, addressing challenges and improving patient and healthcare worker perspectives [19]. More validation is needed, but universal models stress the importance of regional adaptation for effectiveness [20]. Smart cities utilize ICT to improve living standards and tackle challenges. This article explores the role of Artifical intelligence and Machine Learning in shaping smart cities, from transportation to healthcare systems. It highlights research challenges and future directions [21].

Anomaly-based insider threats detection

Recently researcher are investing towards advancement of machine learning models to identify malicious activity carried out by an insider [22, 23]. Xio et al. [24] proposed a technique that utilizes Graph Neural Network for the detection of insider threats. It transform user behavior logs into contextual graphs, improving anomaly identification and achieving interpretability. Evaluation of the CERT dataset shows robustness and high accuracy in detecting insider threats, offering valuable insights for security analysts. A study highlighted use of datamining techniques [25] for privacy preservation.In another work [26] use of statistical methods is proposed for user dat privacy preservation. A recent study [27, 28] utilized LSTM and Graph Neural Networks to identify anomalous nodes in the heterogeneous networks. Another study presents a machine learning based layered architecture for the detection of malicious insider threats [29]. It addresses class imbalance through efficient sampling, with Nearmiss2 (NM-2) identified as optimal.This study has produced results with recall of 100% and Fscore of 78.72% with accuracy of 82.46%. The paper [30] gained first position in the CCF-BDCI competition by producing anomaly score on the CMU-CERT dataset. The technique is able to to detect anomalies under different contexts. A supervised machine learning approach was presented by [31] for the detection of anomalies in the electronic health record. The research produced results with accuracy of 97.6% and accuracy of 97.9% utilizing the SVM classification model. The focus on identifying anomalous behavioral patterns ensures the confidentiality and integrity of sensitive patient data.

Electronic Health Records (EHRs)

There is an emphasis on the global requirements, functional needs, and data security, for EHR development [32]. Examining the evolution of EHRs from 1992 to 2015 and forecasting 25 years, the study highlights a shift from academic to vendor systems. While technical advancements occurred, challenges in procedures, ethics, and politics emerged. Present EHRs struggle to meet healthcare demands, emphasizing unforeseen complexities. The paper anticipates international standards for interoperable applications, supporting precision medicine and a learning health system based on diverse data inputs [33]. Exploring EHRs showcases their potential to enhance patient care and streamline clinical research. Leveraging EHRs for observational studies and clinical trials addresses recruitment challenges and improves data collection, especially in cardiovascular research [34]. Success relies on analytic capabilities, security, privacy, and data quality validation. EHR are crucial for healthcare, integrating patient data for treatment development. Challenges like privacy and interoperability persist. This research emphasizes the Information Systems (IS) discipline's potential in EHR integration and analytics [35]. It identifies collaboration opportunities between IS scholars and healthcare disciplines for improved patient care and healthcare transformation. This research, based on qualitative methods including literature review and interviews with Electronic Patient Record(EPR) experts, develops a framework for EPR ethics, focusing on privacy, confidentiality, consent, data access and sharing, trust, and governance. The framework is validated through a national EPR implementation case study [36]. Using qualitative research with semi-structured interviews, the author explores primary healthcare professionals' perspectives on reminders in electronic patient records. The study finds mixed views, with some seeing reminders as beneficial for patient care and others as burdensome. Participants identify hindrances and enablers for the appropriate use of reminders in primary care [37]. Using collaborative filtering, this research identifies abnormal access to Electronic Health Records. The study is conducted on a dataset of 2 million EHRs and 4040 users at a academic medical center. The research has proposed a collaborative filtering algorithm to user access patterns, demonstrating 90.1% sensitivity and 96.5% specificity in detecting and preventing unintended EHR access [38]. This paper discusses the complexity of ensuring safety and robustness in eHealth networks and services [39]. It emphasizes the need for comprehensive tools, effective governance, and risk management programs to address growing complexity and evolving threats. Cooperation, consistent research, and development are highlighted for fostering a security and resilience culture in eHealth.

Anomaly detection in EHRs

Security and data breeches in Healthcare are well studied in a recent work, where they have explored various approaches to mitigate the issue of data breeches [40]. Recently a study analyzed security techniques for safeguarding electronic health records (EHRs) from unauthorized access, using a dataset of 52 articles published from 2010 to 2019 [41, 42]. Access control, authentication, and encryption are commonly employed methods, with blockchain technology suggested for improving EHR security. The study recommends employing multiple security techniques for a robust EHR security framework. This research develops a privacy-enhancing approach for patient profile management in collaborative eHealth. A system was trialed and tested with ten healthcare clinicians, showing increased privacy and security without disrupting access to patient information [43]. This research reviews mobile cloud computation in medicine, analyzing a dataset of 78 articles to assess benefits

and drawbacks. While it shows potential for improving healthcare delivery, concerns about data security and privacy persist [44]. The paper calls for more research to validate positive aspects and address challenges in mobile cloud computing for healthcare. This literature survey examines the state of secure and robust machine learning in healthcare, analyzing 115 articles. While there's a growing trend of using machine learning in healthcare, issues remain regarding data safety, model reliability, and privacy protection [45]. The report emphasizes the need for further investigation and innovative tools to address these challenges for secure and robust machine learning applications in healthcare. Authors [46] proposed a data mining approach to identify a fraudulent activity in the Health related insurance claims, thereby identifying potential frauds. The author proposes a system using blockchain technology to deter EMR(Electronic Medical Records) data corruption [47]. By connecting a medical blockchain platform with hospital information systems, the system protects EMR integrity and enhances information-sharing efficiency within hospitals and healthcare organizations. This research investigates using machine learning to secure EHR [48]. A systematic survey of literature and case studies shows that machine learning can improve EHR confidentiality by detecting and preventing unauthorized access and reporting security breaches. Cancer records [49] are evaluated for the detection of implausible information utilizing unsupervised algorithms. A recent research [50] utilized BERT to identify anomalies in EHR. They have treated EHR as a natural language and applied the proposed method to identify the anomalies in the sequence of events.

Research gap analysis

The discussion reveals that societies are moving towards digital and smart health, thereby improving patient services. With the increase in advancements in this domain threats are equally arising. Concerns regarding patient data privacy and security are increasing as well. EHR is accessible to the hospital staff and doctors, but as discussed anomalies appear in that record showing some intrusion. The abnormal pattern of data reflects this anomaly. Such anomaly can harm patient financial claims, patient treatment plan to name few. The literature [31] reveals that on this particular dataset of the UK hospital, the existing technique utilized clustering algorithms for labelling purpose and SVM for prediction. The labelling phase was further enhanced by using expert opinion for validation purposes. Expert opinion may not be always available and with evolution of different kinds anomalies the existing model showed limitations in accurate detection of anomalies in the absence of experts. Also, enhanced feature engineering and statistical analysis may reveal more features relevant for the detection of anomalies.

Methodology

This study examines the challenges in safeguarding sensitive EHR. Rapid technology integration necessitates fortifying EHRs against insider threats to ensure patient privacy and data integrity. In this research we have utilized a dataset from a hospital in North England. The data contains information related to patients under treatment in the hospital.

The Fig. 3 depicts various steps of the flow diagram and the Fig. 4 shows details of different phases of the methodology adopted. Firstly in the pre processing phase, data is cleaned and prepared for the analysis. Afterwards feature selection techniques including correlation analysis is carried out to include only the

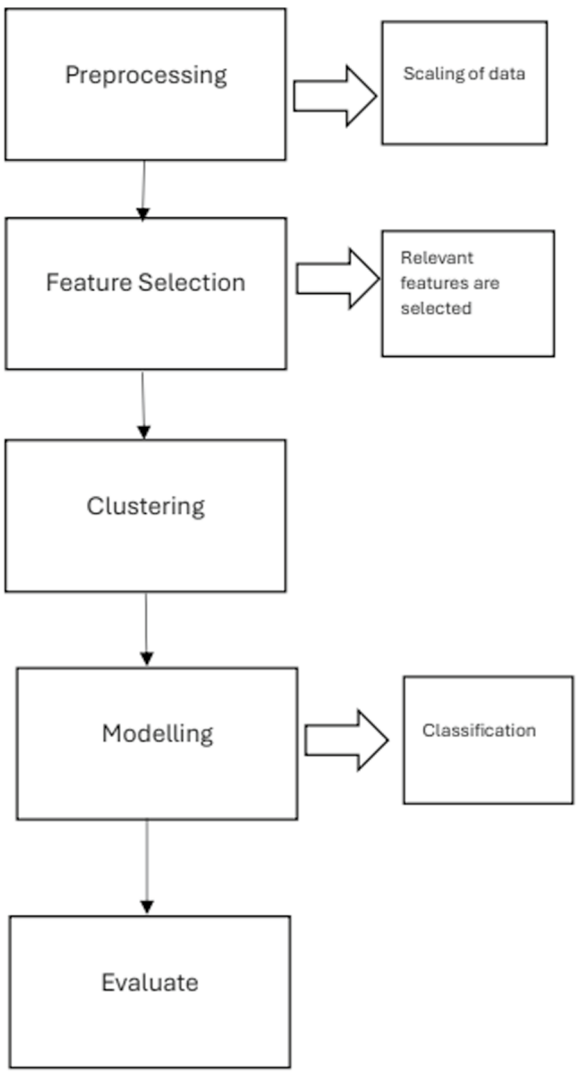

Fig. 3 Flow diagram of proposed methodology

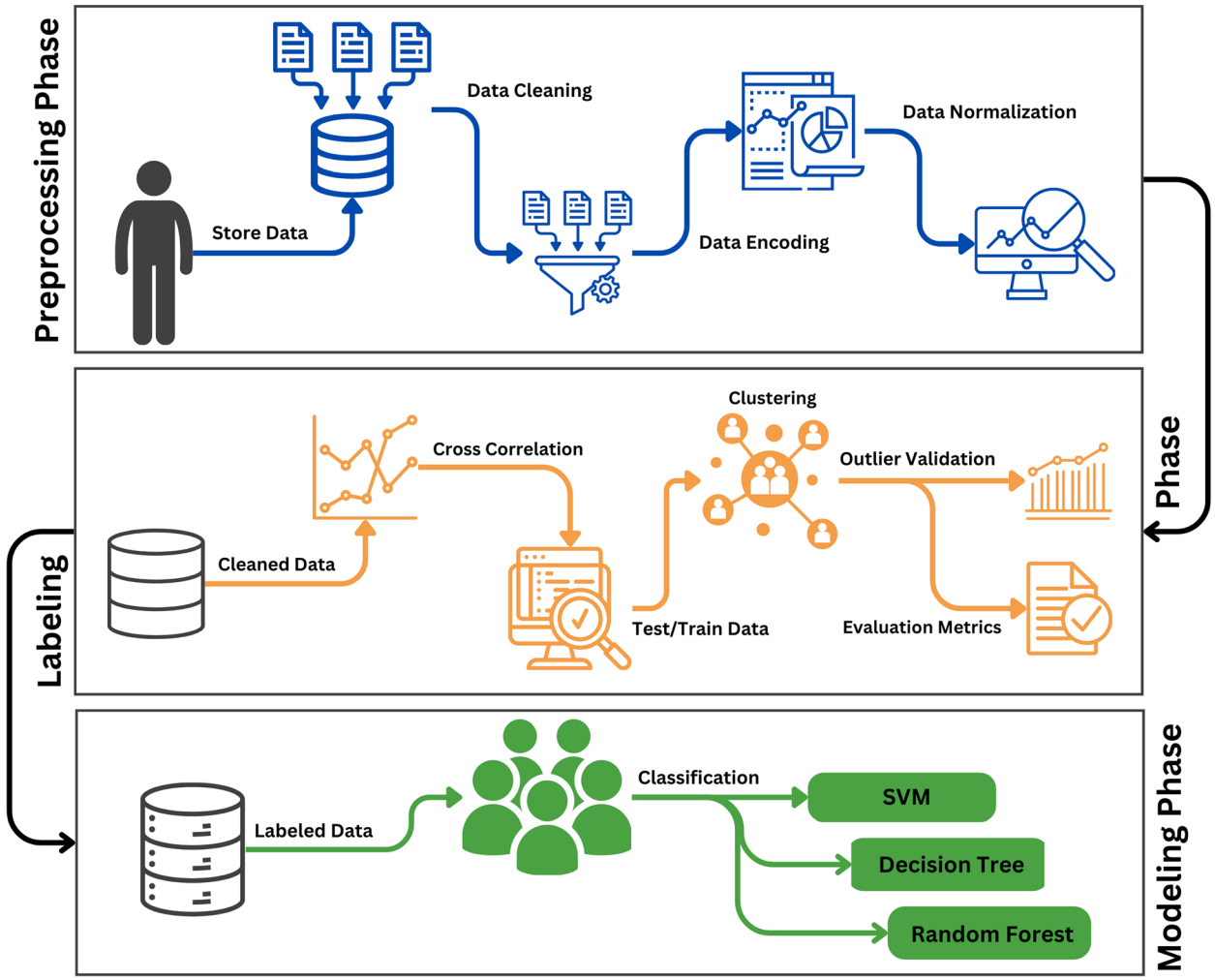

Fig. 4 Methodology

related features . As the data is unlabelled , we utilized two popular algorithms including the isolation forest and local outlier factor. The data is labelled and is validated by utilization of the anomaly scores and other metrics. Finally the model is trained utilizing the supervised algorithms. Thus the methodology is able to predict any unknown instance as anomalous or vice versa.

Dataset

This study has utilized a dataset of electronic health records (EHRs) [51]. The dataset consists of 1,007,727 entries from the audit logs. The dataset contains EHR of patient data, medical histories and other information as given in the Table 1. The data belongs to a hospital in North England.

Preprocessing phase

This phase consists of three main stages including data cleaning, missing values management and normalization. Numeric missing values are replaced with the mean value, while the nominal missing values are replaced with the mode. Categorical data is converted into numerical representation through One-hot encoding. Normalization is done using the min-max algorithm, adjusting numeric values to a range between 0 and 1 [52].

To ensure the data is suitably prepared for further analysis and modeling, these meticulous steps are implemented, resulting in outcomes that are more accurate and reliable.

$$Xnorm = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Table 1 Dataset

Date	Time	Device	User ID	Routine	Patient ID	Duration	Latest Adm date	Latest Dis date
2/28/2016	0:00	4Q7QF3J.1	U6199811	PHA.ORDS	P8290382	54	2/26/2016	2/27/2016
2/28/2016	0:02	27ZKF5J.1	U5053689	ASF	P1591062	13	7/22/2008	7/22/2008
2/28/2016	0:02	C0VLJ5J.2	U2151170	REC REC:(DRP) UK.OE	P3126528	77	2/15/2016	2/15/2016
2/28/2016	0:02	27ZKF5J.1	U5053689	ASF	P1591062	54	7/22/2008	7/22/2008
2/28/2016	0:04	C0VLJ5J.2	U2151170	REC UK.OE	P8672400	147	2/8/2016	2/8/2016
2/28/2016	0:04	BEDSIDE_09.1	U9786800	PHA.ORDS	P7076283	22	1/23/2002	1/23/2002
2/28/2016	0:04	27ZKF5J.1	U5053689	ASF VH	P2718689	39	9/28/2004	9/28/2004
2/28/2016	0:06	C0VLJ5J.2	U2151170	REC REC:(DRP) UK.OE	P8526192	165	1/8/2016	1/8/2016
2/28/2016	0:08	9P7QF3J.3	U4425924	NOTE	P5032341	75	1/25/2012	1/25/2012
2/28/2016	0:10	7ZTLJ5J.1	U8857044	PHA.ORDS	P8705655	42	3/4/2007	3/4/2007

The min-max algorithm, outlined in Eq. 1 in [53], is applied to numeric values to standardize them within the range of 0 to 1, facilitating easier comparison between different variables.

Labeling phase

Post-data refinement, a total of 90,385 distinct identifiers were discerned. Employing a cross-correlation examination, we performed feature selection. Subsequently, we partitioned the dataset into training and evaluation subsets, deploying assorted clustering methodologies. A comparative analysis of clustering efficacy between Isolation Forest and Local Outlier Factor (LOF) was conducted. Ultimately, the validity of outlier ratings derived from the clustering procedure was verified.

Cross correlation

Cross-correlation proves advantageous in scenarios where interrelations exist among dataset features, allowing exploitation of these associations. Consequently, a segment of the data remains untapped, resulting in reduced data volume and computational intricacy. The formula employed to ascertain the cross-correlation between two sequences, u(n) and h(n), is delineated as follows:

$$\text{Cross-correlation} = \sum_{n} u(n) \cdot h(n-k) \tag{2}$$

where: $u(n)$ represents the values of the first sequence at time n,

$h(n-k)$ represents the values of the second sequence shifted by k time units.

According to Eq. 2 cited in [54], the peak cross-correlation occurs when two sequences, represented as $u(n)$ and $h(n)$, are identical. This cross-correlation concept is widely applied across various domains, notably in network intrusion detection. Presented here is an overview of several methodologies leveraging cross-correlation for the identification of network intrusions. Researchers [55, 56] have utilized cross correlation as a feature selection technique for intrusion detection.

In another study [57], utilization of cross correlation demonstrated improvement in the detection accuracy of intrusion attacks in network traffic.

In the domain of classification and intrusion detection, false correlations among features can arise, posing challenges for intrusion detection systems (IDS). Additionally, redundant information across features may complicate the detection process. The inclusion of unnecessary features can prolong computation time and impact IDS accuracy. Achieving optimal classification accuracy hinges on selecting the most relevant subset of features that accurately classify training data. Cross Correlation aids by identifying redundant features and establishes features that can increase the performance of machine learning models.The cross correlation algorithm works according to the following steps.

1. Firstly an initial set of all features is established.
2. The correlation of feature to feature is calculated according to the Eq. 2.
3. The feature that produces maximum correlation is identified.
4. Iterate through step 2 and 3 until desired features are selected.

Features that have high value of correlation show they are similar, and carry similar information. Thus one of them can be selected. Low value indicates the features are unrelated and they carry different information. In the dataset in this study cross correlation revealed that an important feature of Discharge Date was not taken into account in the baseline approach thus loosing some valuable information.

Clustering-Isolation Forest (IForest)

The isolation forest constructs decision trees to isolate anomalies from normal instances, identifying anomalies as points requiring fewer splits across trees. According to standard 70% of the dataset was used for training purposes and the remaining 30% for testing the models. The clustering algorithm assigns labels to the dataset as anomaly or normal instances.

In summary, the training and testing process with the isolation forest offers a means to detect anomalies in datasets, contributing to the anomaly detection field in machine learning.

Training step

This is the stage where the IForest algorithm constructs an ensemble of isolation trees. It divides the training dataset recursively be further into a node where data point is isolated or until tree height is reached. The sub-sample size determines the tree height limit ψ: l = ceil(log2ψ), an average tree height level, with 2 being a good fit. The Algorithm 1 gives detailed steps of isolation forets elaborated below

Algorithm 1 IForest (Xi, n, w)

Inputs: Xi is input data, n is number of trees, w is sub-sampling size Output: a set of n isolation trees
1: Begin initialization of Forest
2: Height limit l = ceiling(log2w)
3: For i = 1 to n do
4: Xi' ← sample(Xi,w)
5: Forest ← Forest U Tree(Xi', 0, l)
6: End for
7: Return Forest

In Algorithm 1 from [58], two inputs of sub sampling size denoted as w and number of trees is denoted as n. This value can be adjusted according to the dataset, as it influences the algorithm's performance in anomaly detection. Steps 3 to 6 recursively runs until each data point is isolated or maximum limit l is achieved.

Testing step

To pinpoint points with high anomaly scores, we compute the average expected path length $E(h(x))$, where $h(x)$ is determined by the path length function (see Algorithm 2). Anomaly scores are then calculated using Eq. 1.

Algorithm 2 Path Length (i, iT, c)

Inputs: i instance, iTisolation tree, c current length
Output: Path Length of instance i
1: If (iT is an external node) then
2: Return $c + cost(iT.size)$
3: End if
4: $a \leftarrow iT.Normal$
5:$b \leftarrow iT.intercept$
6: If $(i-b).a <= 0$ then
7: Return Path Length(i, $iT.left$, $c+1$)
8: Return Path Length(i, $T.right$, $c+1$)
9: End if

According to Algorithm 2 in [58], i denotes the instance, iT denotes the isolation tree, and c denotes the current path length. When the current node iT is external, the path length of instance i is calculated as the sum of the current path length and the cost of the external node $iT.size$. Otherwise, the algorithm iteratively navigates the tree using the split feature and split value until reaching an external node.

Anomaly scores measure the level of uniqueness of every data point compared to the majority. Higher scores will indicate a higher probability of abnormalities. Usually, analysts set a threshold based on analysis or domain knowledge, automating outlier identification and improving anomaly detection, as mentioned in [59].

- Formulation: In the Isolation Forest algorithm, the anomaly score is calculated for each data point using a formula described in [60]. Anomaly scores evaluate a component's dispersion or isolation from the dataset's main population.

$$s(x,n) = 2^{-E(h(n))/c(n)} \tag{3}$$

Here in Eq. 3 from [61]:

- s(x, n) is the anomaly score for data point x in a dataset of size n.
- h(n) is the height of the decision tree for data point x, representing the number of splits or steps needed to isolate x.
- E(h(n)) is the average height of all decision trees in the forest.
- c(n) is a constant that represents the average path length of an unsuccessful search in a binary tree, and it is calculated as:

$$2H(n-1) - 4(n-1)/n \tag{4}$$

In Eq. 4 from [61]: H(i) is the harmonic number.

The harmonic number, H(n), is a mathematical construct that adds the reciprocals of the positive integers up to the value of n.

$$H(i) = 1 + 1/2 + 1/3 + 1/4 + ... + 1/n \quad (5)$$

The harmonic number increases gradually and finding a formula for large 'n' is difficult. In Eq. 5 from [61], It is denoted by 'H(n)', and its value increases logarithmic with 'n'.

Isolation Forests assigns anomaly scores to data points, with normalcy being indicated by a lower score. An anomaly score of 0.57 is applied to identifying and labeling the notable differences from the majority of the dataset, which is helpful for the detection of outliers.

Clustering-Local Outlier Factor (LOF)

The Local Outlier Factor (LOF) finds anomalies through a comparison of local densities and distances between data points. Data points having LOF score beyond 1 are assigned as local outliers. The following algorithm depicts the operation of the Local Outlier Factor (LOF):

Algorithm 3 LOF (k, m, D)

Input: k - number of nearest neighbors, m - number of outliers, D - dataset containing potential outliers
Output: Top m outliers
1: For $j = 1$ to len(D) do
2: Compute k nearest neighbor distances (k-dist(p))
3: Compute neighborhood ($N_k(p)$)
4: End for
5: Calculate reachability distance (reach-dist$_k(p, q)$) and local reachability density (lrd(p))
6: Calculate LOF(p)
7: Sort the LOF values of all points in descending order 8: Return the top m data objects with the largest LOF values, indicating outliers.

The above Algorithm 3 from [58], takes three inputs:

- k representing the number of nearest neighbors to consider, m indicating the number of outliers to identify, and D representing the dataset containing potential outliers.
- It iterates through each data point in the dataset D to compute the k nearest neighbor distances and determine the neighborhood for each point.
- For each data point, it calculates the reachability distance and local reachability density, which are then used to calculate the Local Outlier Factor (LOF) score.
- The LOF score is calculated for each data point, representing its degree of outlier within the dataset.
- After calculating the LOF scores for all data points, the algorithm sorts the LOF values in descending order.
- Finally, the algorithm returns the top m data objects with the largest LOF values, indicating the outliers in the dataset.

Local outlier factor score:

The Local Outlier Factor (LOF) score shows the extent to which each data point is unlike the rest of the dataset. A high LOF score implies a possibility of the data point being an outlier to a greater extent.

The formula for obtaining Local Outlier Factor (LOF) for a given data point is presented as.

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} lrd_k(o)}{lrd_k(p) \cdot |N_k(p)|} \quad (6)$$

Equation 6 from [54], assigns scores to data points in order to distinguish them as normal or anomalous. An average LOF score is equal to 0. 74 is attained. A LOF close to 1 is indicative of normality when local reachable density is equal to the average. On the contrary, a LOF value that is more than 1 indicates an anomaly when local density is lower than the average.

The LOF (Local Outlier Factor) Algorithm 3, is used to detect outliers in a dataset. It works by computing distances to nearest neighbors thereby calculating the reachable distance and the local reachable density of each data point. The LOF score calculates the estimation of outliers. The first data points with the greatest LOF values, which are the outliers of significance, are retrieved. The final LOF score is 0.74, which allows for the effective identification and prioritization of anomalies for further analysis.

Table 2 reveals the dataset statistics like unique IDs and the spread of anomalies. The Isolation Forest algorithm detects 397 anomalies and 89,988 normal data points among the 90,385 unique IDs. In the same way, the LOF algorithm detects 358 anomalies and 90,027 normal data points. This table summarizes dataset properties and cluster algorithms' efficiency in anomaly detection.

Evaluation metrics for labeling phase

Evaluation metrics are defined as numeric values used to measure the performance of the model and can be applied to different fields among which machine learning.

Table 2 Clustering

Clustering algo	Unique IDs	Anomaly	Normal
Isolation Forest	90,385	397	89,988
LOF	90,385	358	90,027

Silhouette score

The Silhouette Score is used to evaluate the clustering performance of Isolation Forest and LOF algorithms, assessing cluster separation and cohesion [62]. This metric measures how effectively the algorithms group data into meaningful clusters. Mathematically the score is presented as given in the equation.

$$\text{Silhouette Score} = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i - a_i}{\max(\mathrm{a_i}, \mathrm{b_i})} \tag{7}$$

Where: N variable present the total number of data points. a_i variable is the average distance from the i-th data point to other data points within the same cluster. b_i presents the minimum average distance from the i-th data point to data points in any other cluster, excluding its own.

Range for the silhouette score is between −1 to 1 where 1 shows that the data point is well clustered, while −1 is an indication that the point has been assigned to a wrong cluster. Score of 0 is an indication that the clusters are overlapping .

The silhouette score for the clustering algorithms applied to the EHR dataset of this study showed scores of 0.63 for the isolation forest and 0.41 for the LOF algorithm, showing that the isolation forest produced better clustering comparatively .

Dunn index

Dunn Index [63]is another evaluation metric that is employed to find the performance of clustering algorithms. Dunn Index is based on minimum inter cluster distance and maximum intra cluster distance.

The mathematical representation of Minimum Intercluster Distance $D_{\min}$ is:

$$D_{\min} = \min\{\mathrm{d}(\mathrm{x_i}, \mathrm{x_j}) \mid \mathrm{x_i} \in \mathrm{C_i}, \mathrm{x_j} \in \mathrm{C_j}, \mathrm{C_i} \neq \mathrm{C_j}\} \tag{8}$$

where $d(x_i, x_j)$ represents the distance between data points x_i and x_j, and C_i and C_j represent the clusters to which x_i and x_j belong, respectively. The computation of Maximum Intra-cluster Diameter $D_{\max}$ is done by:

$$D_{\max} = \max\{\mathrm{d}(\mathrm{x_i}, \mathrm{x_j}) \mid \mathrm{x_i}, \mathrm{x_j} \in \mathrm{C_k}\} \tag{9}$$

where $d(x_i, x_j)$ represents the distance between data points x_i and x_j within the same cluster C_k. Finally the Dunn Index is calculated utilizing following mathematical equation

$$\text{Dunn Index} = \frac{\text{Minimum Inter-cluster Distance}}{\text{Maximum Intra-cluster Diameter}} \tag{10}$$

The Dunn Score for the clustering algorithms applied in this study is 0.45 for the isolation forest and 0.38 for the LOF.

Modeling phase

The modelling phase consists of application of classification algorithms to the dataset. Classification models of support vector machines(SVM), random forest and decision tree are utilized. These models represent the patterns of binary classification where data points are grouped into pre-defined categories according to features to support applications such as medical diagnosis or fraud detection.

In anomaly detection, Decision Trees are used effectively in combination with anomaly scores from the Isolation Forest algorithm. These scores help to direct the tree's nodes in making decisions on instances as normal or anomalous. In this research, decision tree is utilized that is widely applied for anomaly detection the the domains of cyber-security and fraud detection. For training, we used anomaly scores as features for building the tree. They helped in creating decision boundaries and involved choosing features to reduce impurity or increase information gain. Each node was a choice according to anomaly scores and constructed branches through the feature space.

Our approach involves directing new instances through a Decision Tree based on decisions at each node wherein anomaly scores dictate the process. The tree makes instances traverse from root nodes to leaf nodes of the tree while assigning normal or anomalous labels. This approach is to utilize the anomaly scores for prediction which makes the Decision Tree valuable for anomaly detection because of its interpretation effectiveness.

Decision Trees also use anomaly scores for decision-making in anomaly detection but do not include a particular formula for nodes' feature splits. Decision Trees offer simplicity and interpretability, providing clear decision paths. Decision Trees may overfit with complex datasets and be sensitive to minor input changes [64].

In the context of anomaly detection, Random Forest takes advantage of the diversity of individual DTs. Each tree operates on a different subset of data and features which reduces the risk of over-fitting and enhances anomaly detection.

Random Forest investigated in our study uses ensemble learning with different Decision Trees. Random feature selection is used to mitigate overfitting, and each tree is trained on a different subset of the dataset to improve model stability. In Random Forest, each Decision Tree gives its output independently for predictions. For classification, the final prediction is based on a voting mechanism, choosing the class with the most votes. In regression tasks, predictions from each tree are averaged for the final prediction.

Pros of Random Forest include reduced overfitting, robustness against noise and outliers, and good

performance on large datasets. However, the model's complexity increases with the number of trees, and interpretability may be challenging.

Evaluation metrics for modelling phase

The effectiveness of the proposed scheme in achieving the stated objectives is assessed using the performance measures including,Accuracy, Precision, Recall and F1Score

$$Accuracy = TP + TN/TP + FP + FN + TN \quad (11)$$

$$Precision = TP/TP + FP \quad (12)$$

$$Recall = TP/TP + FN \quad (13)$$

$$F1 - score = 2 * (Recall * Precision)/(Recall + Precision) \quad (14)$$

Experiment and results

This section outlines the results obtained from the two algorithms Isolation Forest (IForest) and Local Outlier Factor (LOF). These algorithms were employed for anomaly detection task by labelling transaction as anomalous or vice versa. For the IForest algorithm, model training was conducted using the model.fit() function, with hyperparameters like contamination adjusted within a loop iterating from 0 to 50. We manually optimized parameters such as max_samples and n_estimators. The prediction was performed using the model. predict() function.

The Table 3 shows the evaluation results of the clustering algorithms namely isolation forest and the LOF. The evaluation metrics of silhouette score and Dunn Index show superior performance of the IF algorithm. It is important to mention that the LOF algorithm was employed on the same dataset by the baseline, we have proposed application of isolation forest with better results.

Scenario 1: Actions on patient record

It is observed that most users typically spend 60 seconds to 300 seconds or less, equivalent to 1 to 5 minutes, when performing actions on a patient. This is also supported by the evidence in the dataset where active users action duration do not go beyond 400 seconds.Such behavior aligns with an isolation forest score, indicating normality. However, anomalies are present where user is performing activities for long duration making up to 2 hours or 7000 seconds. Among these anomalies, the user stands out as a notable outlier. These extended durations are flagged as outliers in the isolation forest analysis.

Table 3 Clustering evaluation

Algorithms	Silhouette scores	Dunn index
Isolation Forest	0.63	0.45
Local Outlier Factor	0.41	0.38

Scenario 2: patient ID access

Regarding Patient ID, the dataset shows a prevalent pattern of access durations clustered around 1000 seconds, which translates to approximately 17 minutes. However, there are notable anomalies when Patient ID is accessed for significantly longer duration, such as 3700 seconds and beyond. The hospital has confirmed that typical duration for patient record access is within 1000 seconds reflecting a normalized isolation forest score. This is supported by the observation that a typical clinic session lasts 15 minutes. However there is no clear evidence that supports access of patient IDs for longer duration on different devices.

Scenario 3: device ID access

According to consultations with the hospital, the typical duration for Device ID access is approximately 400 seconds, with most users spending only a few seconds on a device. However, there is an instance where a user spends 1600 seconds and more on a device. Actions on a device involving patient-related tasks typically last 300 seconds or less. The evidences from the dataset also supports that device usage beyond 600 seconds is an abnormal behavior. However, there are exceptions when access duration reaches 1700 seconds. Therefore, the device access in the dataset is set at 400 seconds. likely resulting in an isolation forest score of 1.

Scenario 4: routene ID access

Upon analysis of Routine ID data, it is evident that the device is used for 400 seconds in the dataset. However, there are exceptions, such as routines accessed for up to 1700 seconds. The presence of extreme anomalies, like routines lasting 12,000 seconds, complicates the observation of Routine ID behavior.Typically a routine take 1000 seconds for completion. Therefore, the routine behavior in the dataset is set at 1000 seconds, aiding in the identification of anomalies using isolation forest analysis.

Scenario 5: patient discharge record access

In analyzing the record access date and patient's latest discharge date, the dataset reveals a consistent trend where patients are typically discharged within a reasonable time frame following their treatment as shown in the Table 4. Generally, patients spend 60 seconds to 300 seconds or less, equivalent to 5 minutes, on post-discharge actions. This pattern is evident across the majority of patient records, with the isolation forest score(Anomaly

Table 4 Statistics of dataset scenario 5

Date	Device	User ID	Routine	Patient ID	Duration	Latest Dis date
3/25/2016	341874J.1	U1029815	REC REC:(DRP)	P5110410	39	10/18/2007

Score) indicating normality. However, anomalies exist within the dataset, manifesting as significantly prolonged duration of record access post-discharge. For instance, some patients' records are accessed for duration as long as 7000 seconds almost 2 hours, OR less than 60 seconds, after their discharge, which deviates significantly from the norm. Notably, among these outliers, certain records, stand out as notable anomalies. These extended duration are identified as outliers in the isolation forest analysis, indicating abnormal post-discharge activities.

The following section elaborates on the anomaly detection criteria based on the previous model and the proposed model Fig. 5:

1. Date: The date when the patient record was accessed (e.g., March 25, 2016).
2. Device: The device used to access the patient record (e.g., 341874J.1).
3. User ID: The User accessing the patient record (e.g., U1029815).
4. Routine: The type of access or routine associated with the record access (e.g., REC REC:(DRP)).
5. Patient ID: The ID of the patient whose record was accessed (e.g., P5110410).
6. Duration: The duration of the record access in seconds (e.g., 39).
7. Latest Dis Date: The latest discharge date of the patient associated with the record (e.g., October 18, 2007).

Following contextual anomalies are identified in the baseline approach.

- If the duration of the record access falls between 60 and 300 seconds, it is considered normal.
- If the duration exceeds 300 seconds, it is flagged as an anomaly.

The proposed approach is able to identify following anomalies that remained undetected in the baseline approach.

- If the duration of the record access is less than 60 seconds and the access occurs after the patient's discharge date, it is considered an anomaly.
- If the duration of the record access is more than 300 seconds and the access occurs after the patient's discharge date, it is also considered an anomaly.

In the provided example, the record access duration is 39 seconds, below the 60-second threshold. The access also occurs after the patient's latest discharge date (October 18, 2007). Therefore, this record access is marked as an anomaly by the proposed model. This example illustrates the differences in anomaly detection criteria between the previous and proposed models, considering access duration and the timing of the patient's discharge date.

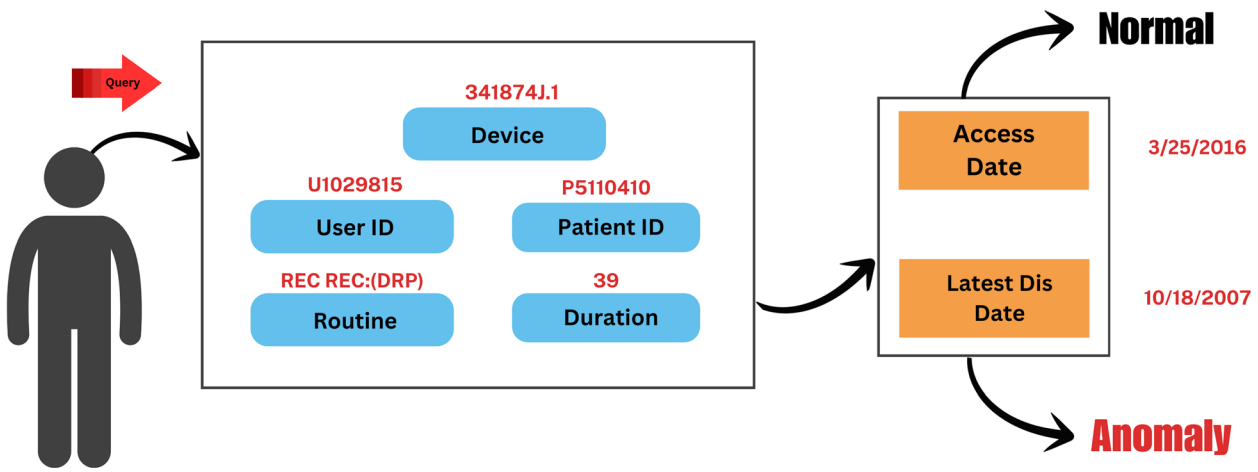

Fig. 5 Proposed model

Performance evaluation

In this research Isolation Forest, Local Outlier Factor, SVM, Decision Tree, and Random Forest algorithms are utilized to assess the effectiveness of the proposed methodology.The evaluation metrics showed better performance of the isolation forest algorithm. The Fig. 6 graphically presents the anomaly detection on the dataset.

The LOF algorithm is graphically presented in the Fig. 7. The algorithm compares the distance from the neighbors of the data point.

The second stage of the methodology involves application of classification algorithms. For this purpose we evaluated the performance of Support Vector Machines, Decision Tree and Random Forest. The results showed that SVM produced promising results with accuracy of 99.21%. Figure 8 dipicts the results when isolation forest was used for clustering followed the above discussed classifiers.

SVM also showed better results when LOF is used for clustering. The results show 98.21% accuracy as depicted in the Fig. 9

The Table 5 presents the results of isolation forest while utilizing different classifiers. The results show accuracy, specificity, sensitivity, F1 score and precision.

The Table 6 presents the results of LOF while utilizing different classifiers. The results show accuracy, specificity, sensitivity, F1 score and precision.

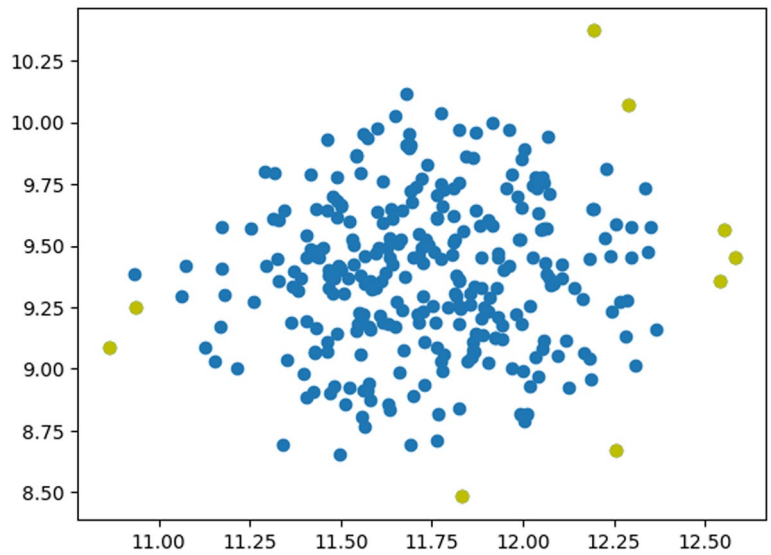

Fig. 6 Anomaly detection by IForest

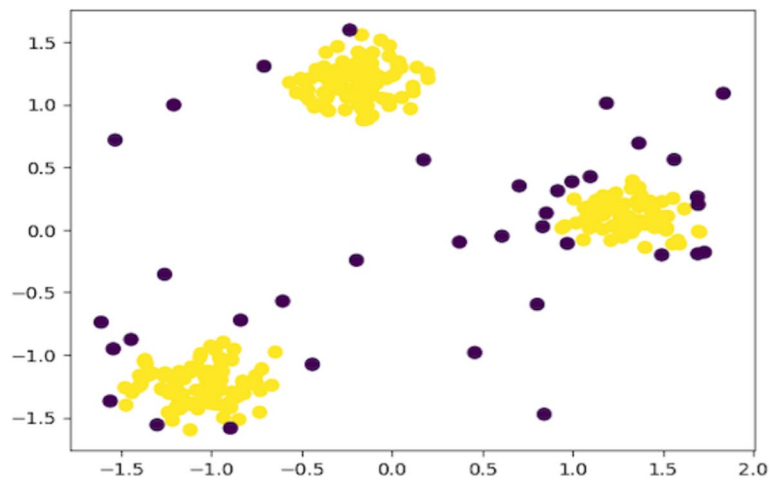

Fig. 7 Anomaly detection by LOF

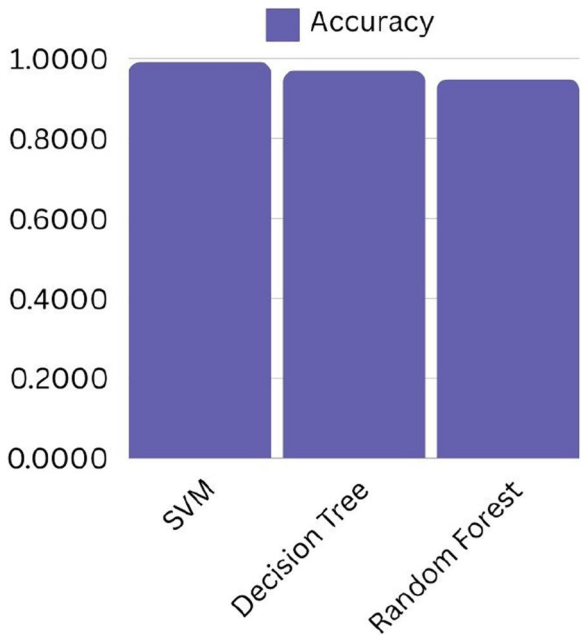

Fig. 8 IForest accuracy

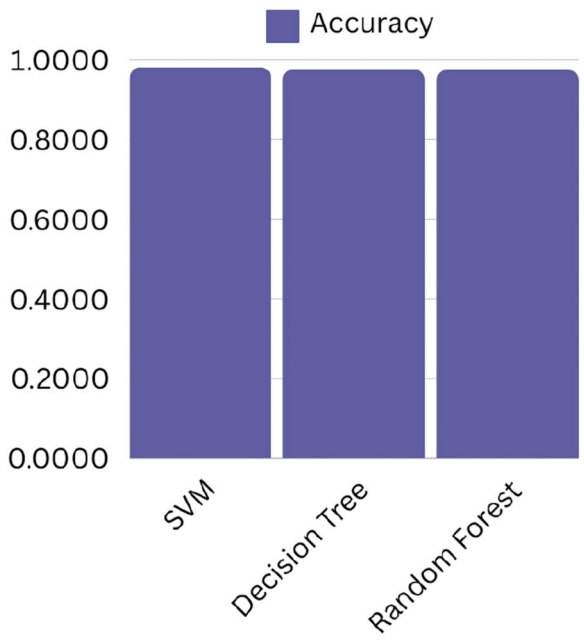

Fig. 9 LOF accuracy

Results are graphically presented in the Figs. 10 and 11. In terms of accuracy, IForest generally outperforms LOF across all models. SVM achieved the highest accuracy among IForest models at 99.21%, whereas LOF's highest accuracy, achieved by SVM as well, was slightly lower at 98.21%. The Decision Tree model in both IForest and LOF yielded comparable accuracies of 98.92% and 97.82% respectively. However, Random Forest showed a significant difference in accuracy between IForest (98.85%) and LOF (97.75%), indicating a better performance in the IForest framework.

Regarding sensitivity or the true positive rate, IForest models exhibited higher values than LOF models. The Decision Tree model in IForest achieved the highest sensitivity at 99.75%, followed closely by SVM at 98.23%. In contrast, LOF's Decision Tree model had a sensitivity of 97.52%, and its SVM model scored slightly higher at 99.45%. Random Forest models showed similar trends, with IForest achieving higher sensitivity compared to LOF.

Specificity, representing the true negative rate, also favored IForest models over LOF. SVM in IForest demonstrated the highest specificity at 99.32%, followed by the Decision Tree model at 98.97%. LOF's SVM model had a slightly lower specificity of 98.34% while its Decision Tree model scored 97.92%, Figure 4.6. Random Forest models, however, showed lower specificitiy across both IForest and LOF, with IForest's Random Forest achieving 72.84% compared to LOF's 69.84%.

The F1 Score, which balances precision and recall, showcased a similar trend as sensitivity and specificity. IForest models generally had higher F1 Scores compared to LOF models, indicating better overall performance in terms of both precision and recall.

Recall, representing the ability of the classifier to find all positive instances, mirrored sensitivity and showed higher values for IForest models compared to LOF models. Overall, these comparisons suggest that IForest performs better than LOF considering accuracy, sensitivity, specificity, F1 Score, and recall thereby, making it a referring it as the suitable choice for anomaly detection.

Table 5 IForest results

Models	Accuracy	Sensitivity	Specificity	F1 score	Precision	Kappa
SVM	0.9921	0.9975	0.9932	0.9872	0.9823	0.6823
Decision Tree	0.9892	0.9823	0.9897	0.9935	0.9975	0.6631
Random Forest	0.9885	0.9728	0.7284	0.8321	0.9728	0.5921

Table 6 LOF results

Models	Accuracy	Sensitivity	Specificity	F1 score	Precision	Kappa
SVM	0.9821	0.9945	0.9834	0.9765	0.9752	0.6752
Decision Tree	0.9782	0.9752	0.9792	0.9864	0.9945	0.6612
Random Forest	0.9775	0.9618	0.6984	0.7923	0.9618	0.5810

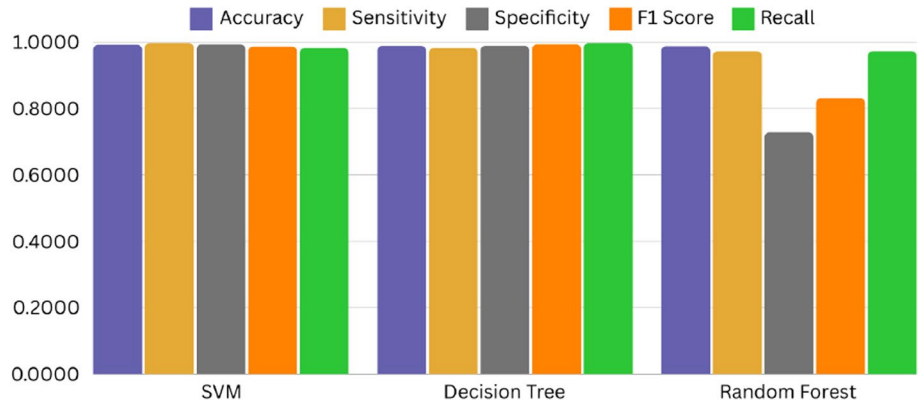

Fig. 10 IForest performance

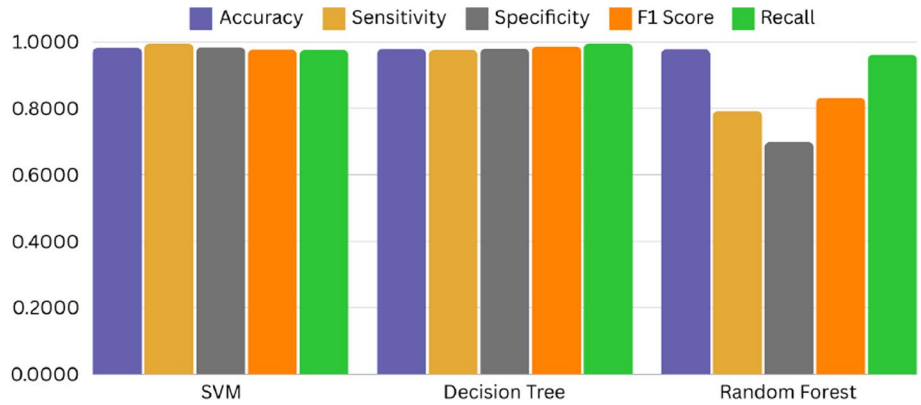

Fig. 11 LOF performance

Comparison with baseline

A comparison table summarizing the results of the baseline classifiers (LOF) and the classifiers trained using Isolation Forest and LOF is shown in the Table 7.

Comparison of the classification performance of SVM, Decision Tree, and Random Forest models between baseline results and the proposed. In the baseline scenario, SVM demonstrated an accuracy of 98.96%, sensitivity of 97.96%, and specificity of 98.97%. Meanwhile, the Decision Tree model showed an accuracy of 98.783%, high sensitivity at 99.98%, and specificity of 98.86%. However, the Random Forest model exhibited a slightly lower accuracy of 98.783%, sensitivity of 97.436%, and relatively lower specificity at 39.2523%. In our experimental results, SVM displayed improved accuracy at 99.21%, sensitivity at 99.75%, and maintained a high specificity of 99.32%, with an F1 Score of 98.72%. Similarly, the Decision Tree model maintained its accuracy at 98.92%, sensitivity at 98.23%, and specificity at 98.97%, with a commendable F1 Score of 99.35%. However, the Random Forest model's performance slightly decreased compared to the baseline, with an accuracy of 98.85%, sensitivity of 97.28%, specificity of 72.84%, and an F1 Score of 83.21%. Overall, the experimental results suggest enhancements in SVM and Decision Tree models' classification parameters, while Random Forest's performance showed a slight decline.

Table 7 Comparison with baseline

Model	Accuracy	Sensitivity	Specificity	F1 score
Baseline (LOF) SVM	0.9896	0.9796	0.9897	0.9846
Baseline (LOF) Decision Tree	0.98783	0.9998	0.9886	0.9938
Baseline (LOF) Random Forest	0.98783	0.97436	0.392523	0.9811
Isolation Forest SVM	0.9921	0.9975	0.9932	0.9872
Isolation Forest Decision Tree	0.9892	0.9823	0.9897	0.9935
Isolation Forest Random Forest	0.9885	0.9728	0.7284	0.8321
LOF SVM	0.9821	0.9945	0.9834	0.9765
LOF Decision Tree	0.9782	0.9752	0.9792	0.9864
LOF Random Forest	0.9775	0.9618	0.6984	0.7923

Conclusion

The findings from this study on utilizing technology for accessing and analyzing patient's health information for smart healthcare systems and smart cities concluded that security and privacy are critical in advanced systems. This paper aimed to apply machine learning techniques to defend EHRs; specifically, Isolation Forest and Local Outlier Factor (LOF) algorithms were used to identify unwanted changes collectively. The Isolation Forest is a tree-based algorithm for detecting anomalies, and the isolation score revolves around isolation depth, at the same time LOF utilizes scoring methods based on densities compared to proximal neighbors.

In order to test the performance, we have utilized the Silhouette Score and Dunn Index that quantify the cohesiveness of clusters and the distance between clusters. Isolation Forest, achieved significantly high accuracy of (99. 21%) in contrast with LOF (98. 21%). Furthermore, the sensitivity & specificity bore high values indicating the robustness of the algorithm. Classification algorithms including Support Vector Machines (SVMs), Decision Trees, and Random Forests were analyzed for the performance. The results of the performance metrics of accuracy, precision and F1 Score revealed that the combination of isolation forest and SVM classifier produced better results for this dataset. They were also able to identify newer contextual anomalies that were not addressed in the previous work. We identified abnormal pattern of data from pattern of increased user actions, time spent while performing an action , date of access of certain data.

This paper analyzed and proposed a methodology for the identification of the contextual anomalies on the specific dataset. The proposed methodology produces improved results with inclusion of anomaly based models compared to previous work on the same dataset, thereby limiting the inclusion of experts in validating the results. The anomaly detection system can be deployed in hospital to monitor an unusual pattern in the EHR. According to the specific dataset utilized in this study the system can aid to identify if any unusual duration is spent on accessing patient record, device, and the action being conducted referred as routine id. Furthermore, the proposed methodology also identified discharge date a candidate for unusual pattern identification, whereby access of patient data beyond discharge date for abnormal duration is flagged as an anomaly. The system can help the hospital administration for any potential insider threat.

Limitations and future work

Medical practices and procedures evolve over time and sometimes context considered as anomaly might not be something of concern. EHR data is very sensitive thus availability and access to large volumes of data for anomaly detection is legally constraining. The evaluations for anomaly detection can be further enhanced with the inclusion of other features related to billing, pharmacy thereby leading to a more comprehensive conclusion when analysing the anomalies in relationship In future application of newer models and data mining techniques can also be explored for its effectiveness on unseen data.

Acknowledgements
Not applicable.

Institutional review board statement
Not applicable for studies not involving humans or animals.

Informed consent statement
Not applicable for studies not involving humans.

Authors' contributions
All authors contributed equally to this work. Muntaha and Saba wrote main manuscript and developed the idea including Dataset Acquisition. Amal , Bader reviewed the Methodology , Ali Reviewed and validated the results of the manuscript. Fatima provided consultancy for the improvement and comparative analysis of the work. All authors reviewed the work.

Funding
This research is funded by Rabdan Academy , Abu Dhabi, United Arab Emirates.

Data availability
The simulation files/data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors are agreed to publish this work.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Computer Science, Bahria University, Islamabad, Pakistan. [2]Department of Information Systems and Technology, Collage of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia. [3]Software Engineering Department, College of Computing and Information Sciences, King Saud University, Riyadh, Saudi Arabia. [4]Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates. [5]Centre of Excellence in Artificial Intelligence COE-AI, Bahria University, Islamabad, Pakistan.

Received: 18 July 2024 Accepted: 11 November 2024
Published online: 19 November 2024

References

1. Ristvej J, Lacinák M, Ondrejka R. On smart city and safe city concepts. Mob Netw Appl. 2020;25:836–45.
2. Galvão YM, Castro L, Ferreira J, Neto FBdL, Fagundes RAdA, Fernandes BJ. Anomaly Detection in Smart Houses for Healthcare: Recent Advances, and Future Perspectives. SN Comput Sci. 2024;5(1):136.
3. Heekin AM, Kontor J, Sax HC, Keller MS, Wellington A, Weingarten S. Choosing Wisely clinical decision support adherence and associated inpatient outcomes. Am J Manage Care. 2018;24(8):361.
4. Hoffman SAE. Cybersecurity Threats in Healthcare Organizations:: Exposing Vulnerabilities in the Healthcare Information Infrastructure. World Libr. 2020;24(1).
5. El-Bakkouri N, Mazri T. Security Threats in Smart Healthcare. Int Arch Photogramm Remote Sens Spat Inf Sci. 2020;44:209–14.
6. Kavitha M, Srinivas P, Kalyampudi PL, Srinivasulu S, et al. Machine learning techniques for anomaly detection in smart healthcare. In: 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE; 2021. pp. 1350–1356.
7. Yin C, Xiong Z, Chen H, Wang J, Cooper D, David B. A literature survey on smart cities. Sci China Inf Sci. 2015;58(10):1–18.
8. Batty M, Axhausen KW, Giannotti F, Pozdnoukhov A, Bazzani A, Wachowicz M, et al. Smart cities of the future. Eur Phys J Spec Top. 2012;214:481–518.
9. Ruohomaa H, Salminen V, Kunttu I. Towards a smart city concept in small cities. Technol Innov Manag Rev. 2019;9:5–14.
10. Ruokolainen J, Nätti S, Juutinen M, Puustinen J, Holm A, Vehkaoja A, et al. Digital healthcare platform ecosystem design: A case study of an ecosystem for Parkinson's disease patients. Technovation. 2023;120:102551.
11. Herman H, Grobbelaar SS, Pistorius C. The design and development of technology platforms in a developing country healthcare context from an ecosystem perspective. BMC Med Inform Dec Making. 2020;20:1–24.
12. Newaz AI, Sikder AK, Rahman MA, Uluagac AS. Healthguard: A machine learning-based security framework for smart healthcare systems. In: 2019 sixth international conference on social networks analysis, management and security (SNAMS). IEEE; 2019. pp. 389–96.
13. Masood I, Wang Y, Daud A, Aljohani NR, Dawood H. Towards smart healthcare: patient data privacy and security in sensor-cloud infrastructure. Wirel Commun Mob Comput. 2018;2018(1):2143897.
14. Tian S, Yang W, Le Grange JM, Wang P, Huang W, Ye Z. Smart healthcare: making medical care more intelligent. Glob Health J. 2019;3(3):62–5.
15. Yin H, Akmandor AO, Mosenia A, Jha NK, et al. Smart healthcare. Found Trends® Electron Des Autom. 2018;12(4):401–66.
16. Alharbey R, Kim JI, Daud A, Song M, Alshdadi AA, Hayat MK. Indexing important drugs from medical literature. Scientometrics. 2022;127(5):2661–81.
17. Poongodi M, Sharma A, Hamdi M, Maode M, Chilamkurti N. Smart healthcare in smart cities: wireless patient monitoring system using IoT. J Supercomput. 2021;77:12230–55.
18. Tian YJ, Felber NA, Pageau F, Schwab DR, Wangmo T. Benefits and barriers associated with the use of smart home health technologies in the care of older persons: a systematic review. BMC Geriatr. 2024;24(1):152.
19. Abbas T, Haider AK, Kanwal K, Daud A, Irfan M, Bukhari A, et al. IoMT-Based Healthcare Systems: A Review. Comput Syst Sci Eng. 2024;48(4):871–95.
20. Kamruzzaman M. New opportunities, challenges, and applications of edge-AI for connected healthcare in smart cities. In: 2021 IEEE Globecom Workshops (GC Wkshps). IEEE; 2021. pp. 1–6.
21. Ullah Z, Al-Turjman F, Mostarda L, Gagliardi R. Applications of artificial intelligence and machine learning in smart cities. Comput Commun. 2020;154:313–23.
22. Alzaabi FR, Mehmood A. A review of recent advances, challenges, and opportunities in malicious insider threat detection using machine learning methods. IEEE Access. 2024;12:30907–27.
23. Tn N, Pramod D. Insider intrusion detection techniques: A state-of-the-art review. J Comput Inf Syst. 2024;64(1):106–23.
24. Xiao J, Yang L, Zhong F, Wang X, Chen H, Li D. Robust anomaly-based insider threat detection using graph neural network. IEEE Trans Netw Serv Manag. 2022;20(3):3717–33.
25. Kumar GS, Premalatha K. STIF: Intuitionistic fuzzy Gaussian membership function with statistical transformation weight of evidence and information value for private information preservation. Distrib Parallel Databases. 2023;41(3):233–66.
26. Kumar GS, Premalatha K, Maheshwari GU, Kanna PR, Vijaya G, Nivaashini M. Differential privacy scheme using Laplace mechanism and statistical method computation in deep neural network for privacy preservation. Eng Appl Artif Intell. 2024;128:107399.
27. Hayat MK, Daud A, Banjar A, Alharbey R, Bukhari A. A deep co-evolution architecture for anomaly detection in dynamic networks. Multimed Tools Appl. 2024;83(14):40489–508.
28. Hayat MK, Daud A. Anomaly detection in heterogeneous bibliographic information networks using co-evolution pattern mining. Scientometrics. 2017;113(1):149–75.
29. Asha S, Shanmugapriya D, Padmavathi G. Malicious insider threat detection using variation of sampling methods for anomaly detection in cloud environment. Comput Electr Eng. 2023;105:108519.
30. Wang E, Li Q, Zhao S, Han X. Anomaly-Based Insider Threat Detection via Hierarchical Information Fusion. In: International Conference on Artificial Neural Networks. Springer; 2023. pp. 13–25.
31. Hurst W, Tekinerdogan B, Alskaif T, Boddy A, Shone N. Securing electronic health records against insider-threats: a supervised machine learning approach. Smart Health. 2022;26:100354.
32. Hoerbst A, Ammenwerth E. Electronic health records. Methods Inf Med. 2010;49(04):320–36.
33. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform. 2016;25(S 01):S48–61.
34. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017;106:1–9.
35. Kohli R, Tan SSL. Electronic health records. MIS Q. 2016;40(3):553–74.
36. Jacquemard T, Doherty CP, Fitzsimons MB. The anatomy of electronic patient record ethics: a framework to guide design, development, implementation, and use. BMC Med Ethics. 2021;22(1):1–14.
37. Cecil E, Dewa L, Ma R, Majeed A, Aylin P. RF20 Primary health care professionals views of reminders in electronic patient records. J Epidemiol Community Health. 2019;73(Suppl 1):A64.
38. Menon AK, Jiang X, Kim J, Vaidya J, Ohno-Machado L. Detecting inappropriate access to electronic health records using collaborative filtering. Mach Learn. 2014;95:87–101.
39. Liveri D, Sarri A, Skouloudi C. Security and resilience in eHealth infrastructures and services. Secur Chall Risks. 2015.
40. Nemec Zlatolas L, Welzer T, Lhotska L. Data breaches in healthcare: security mechanisms for attack mitigation. Clust Comput. 2024:1–16.
41. Kruse CS, Smith B, Vanderlinden H, Nealand A. Security techniques for the electronic health records. J Med Syst. 2017;41:1–9.
42. Feroze A, Daud A, Amjad T, Hayat MK. Group anomaly detection: Past notions, present insights, and future prospects. SN Comput Sci. 2021;2:1–27.
43. Sánchez-Guerrero R, Mendoza FA, Diaz-Sanchez D, Cabarcos PA, López AM. Collaborative ehealth meets security: Privacy-enhancing patient profile management. IEEE J Biomed Health Inform. 2017;21(6):1741–9.
44. Wang X, Jin Z. An overview of mobile cloud computing for pervasive healthcare. IEEE Access. 2019;7:66774–91.
45. Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: A survey. IEEE Rev Biomed Eng. 2020;14:156–80.
46. Hamid Z, Khalique F, Mahmood S, Daud A, Bukhari A, Alshemaimri B. Healthcare insurance fraud detection using data mining. BMC Med Inform Decis Mak. 2024;24(1):112.
47. Hang L, Choi E, Kim DH. A novel EMR integrity management based on a medical blockchain platform in hospital. Electronics. 2019;8(4):467.

48. Seh AH, Al-Amri JF, Subahi AF, Agrawal A, Pathak N, Kumar R, et al. An analysis of integrating machine learning in healthcare for ensuring confidentiality of the electronic records. Comput Model Eng Sci. 2021;130(3):1387–422.
49. Röchner P, Rothlauf F. Unsupervised anomaly detection of implausible electronic health records: a real-world evaluation in cancer registries. BMC Med Res Methodol. 2023;23(1):125.
50. Niu H, Omitaomu OA, Langston MA, Olama M, Ozmen O, Klasky HB, et al. EHR-BERT: A BERT-based model for effective anomaly detection in electronic health records. J Biomed Inform. 2024;150:104605.
51. Hurst W. Electronic Patient Record Dataset - UK Hospital. DANS Data Station Life Sciences; 2017. https://doi.org/10.17026/dans-znf-sh4q.
52. Liu Z, et al. A method of SVM with normalization in intrusion detection. Procedia Environ Sci. 2011;11:256–62.
53. Alanazi R, Aljuhani A. Anomaly Detection for Industrial Internet of Things Cyberattacks. Comput Syst Sci Eng. 2023;44(3).
54. Farahani G. Feature selection based on cross-correlation for the intrusion detection system. Secur Commun Netw. 2020;2020:1–17.
55. Amiri F, Yousefi MR, Lucas C, Shakery A, Yazdani N. Mutual information-based feature selection for intrusion detection systems. J Netw Comput Appl. 2011;34(4):1184–99.
56. Zhang X, Zhu Z, Fan P. Intrusion detection based on cross-correlation of system call sequences. In: 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05). IEEE; 2005. pp. 7–pp.
57. Zhang Y, Yang Q, Lambotharan S, Kyriakopoulos K, Ghafir I, AsSadhan B. Anomaly-based network intrusion detection using SVM. In: 2019 11th International conference on wireless communications and signal processing (WCSP). IEEE; 2019. pp. 1–6.
58. Fadul AMA. Anomaly Detection based on Isolation Forest and Local Outlier Factor. Africa University; 2023.
59. Kaushal A, Shukla M. Comparative analysis to highlight pros and cons of data mining techniques-clustering, neural network and decision tree. Int J Comput Sci Inf Technol. 2014;5(1):651–6.
60. Ali J, Khan R, Ahmad N, Maqsood I. Random forests and decision trees. Int J Comput Sci Issues (IJCSI). 2012;9(5):272.
61. Mensi A, Bicego M. A novel anomaly score for isolation forests. In: Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part I 20. Springer; 2019. pp. 152–163.
62. Jones PJ, James MK, Davies MJ, Khunti K, Catt M, Yates T, et al. FilterK: a new outlier detection method for k-means clustering of physical activity. J Biomed Inform. 2020;104:103397.
63. Bezdek JC, Pal NR. Cluster validation with generalized Dunn's indices. In: Proceedings 1995 second New Zealand international two-stream conference on artificial neural networks and expert systems. IEEE; 1995. pp. 190–193.
64. Kumar VP, Sowmya I. A review on pros and cons of machine learning algorithms. J Eng Sci. 2021;12(10):272–6.

Publisher's Note