Development of a Blockchain Platform for Protection and Security of Medical Data and Patient Identification in Kazakhstan

Alisher Batkuldin *, Leila Rzayeva *‡, Baktiyar Toksanbay *, Pavel Kim *, Zhaksylyk Kozhakhmet *, Abilkair Imanberdi *

* Research and Innovation Center "CyberTech", Astana IT University, 010000

(a.batkuldin@astanait.edu.kz, l.rzayeva@astanait.edu.kz, 212486@astanait.edu.kz, 212397@astanait.edu.kz, zh.kozhakhmet@astanait.edu.kz, a.imanberdiyev@astanait.edu.kz)

‡Corresponding Author; Leila Rzayeva, Research and Innovation Center "CyberTech", Astana IT University, 010000, l.rzayeva@astanait.edu.kz

Received: 07.08.2025 Accepted:04.12.2025

Abstract- Data breaches in healthcare are increasingly caused by insider threats, yet traditional logging systems remain vulnerable to tampering. This paper proposes a hybrid security framework for medical data in Kazakhstan, combining blockchain for immutable log storage with machine learning for automated anomaly detection. We introduce a dual-backend architecture where a private Ethereum network secures access logs, while a separate analysis module utilizes supervised learning algorithms to identify suspicious behaviour patterns in real-time. To validate the system, we generated a region-specific synthetic dataset compliant with Kazakhstani national identification standards (IIN). Experimental results demonstrate that the system effectively secures audit trails against modification and detects anomalous access patterns with high accuracy. The proposed solution addresses the critical gap between data integrity and proactive threat detection, offering a scalable architecture compliant with data privacy regulations.

Keywords: Personal data protection, blockchain, anomaly detection, machine learning, keyless signature infrastructure.

1. Introduction

Data leaks are a significant concern globally, including in Kazakhstan. Data leakage can be caused by a range of factors, including hacker attacks, system misconfiguration, and human error. The research of data breaches in Kazakhstan found that 24 cases of data compromise occurred. As shown in Fig.1, only 27.3% of these leaks were caused by external intruders. The remaining 72.7% of data breaches were caused by internal attackers, underscoring the need to improve data security measures within organizations. Internal threats often go undetected due to insufficient monitoring and lack of stringent access controls, which highlights the importance of enhancing internal security protocols and implementing more robust insider threat detection mechanisms.

The healthcare sector is especially vulnerable to data breaches due to the sensitive nature of stored information that can be targeted by attackers. According to a report by IBM Security on the cost of data breaches in 2022 and 2023 [1], the healthcare sector had the highest cost among all sectors in both years, as shown in Fig.2. The financial impact of these breaches extends beyond immediate recovery and remediation costs to include long-term reputational damage and regulatory fines. Furthermore, the compromised data often includes highly sensitive personal and medical information, which can be exploited for identity theft, insurance fraud, and other malicious activities, thereby amplifying the potential damage.

Regarding breaches, Jiang J. and Bai G. [2] presented statistics on data breaches in the health sector from 2009 to 2017 on Fig.3. During this period, over 1,800 data leaks were reported in US hospitals. Among these incidents, 53% were internal. Out of these, 32.5% were stolen by unknown individuals, 10.5% were due to employee errors in data dissemination, and 9.0% were stolen by employees. The reliability of storing patient data, even in the hands of medical staff, is called into question by the results of this study.

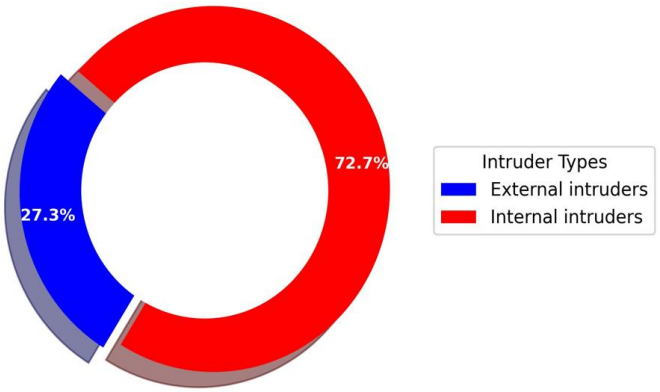

Fig. 1. Types of intruders in data breaches in Kazakhstan.

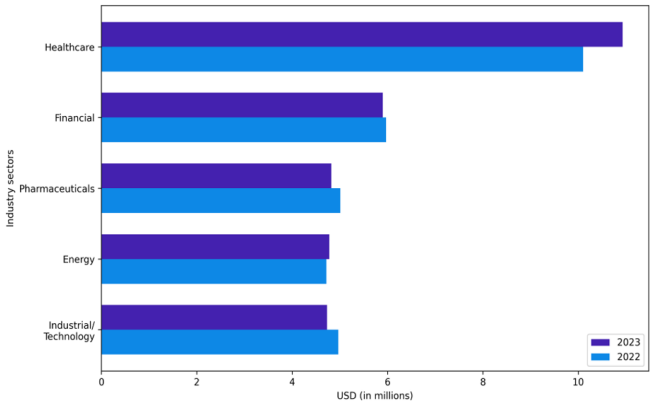

Fig. 2. Cost of data breaches by industrial sector in 2022 and 2023.

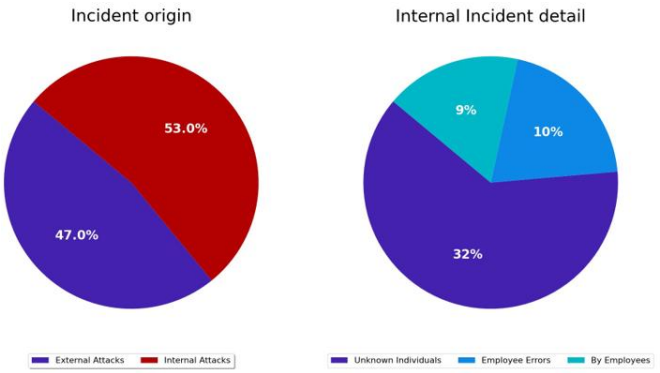

Fig. 3. Incident origins in US hospitals from 2009 to 2017.

Such a high proportion of unknown individuals calls into question the methods of detecting the intruders. This data emphasizes the critical need for robust internal security measures and comprehensive employee training programs to mitigate the risk of internal breaches. The significant proportion of breaches attributed to unknown individuals indicates a gap in the effectiveness of current security systems in identifying and tracking unauthorized access attempts.

As the above statistics show, the challenges of detecting and investigating data breaches are also significant. For instance, E-Estonia's report on the implementation of blockchain in the state revealed that this process took an average of seven months [3]. However, after the implementation of blockchain, the process became instantaneous. Therefore, it is worth giving attention to Blockchain as one of the most possible solutions, because Blockchain is also known for its ability to ensure data integrity, so that data stored on the blockchain cannot be edited or deleted. Therefore, a blockchain-based solution methodology may be the most effective approach to quickly identify and address corruption- related problems. Blockchain's inherent properties of transparency, immutability, and decentralization offer significant advantages in the detection and prevention of data breaches, making it a promising technology for enhancing the security of sensitive information in the healthcare sector and beyond. Additionally, integrating blockchain with existing security protocols could provide a more comprehensive and resilient defense against both internal and external threats.

Blockchain's ability to provide an immutable ledger of transactions can help in tracing the origins of a breach, ensuring that all actions are recorded in a tamper-proof manner. This feature is particularly beneficial in forensic investigations, where establishing a clear chain of events is crucial. Moreover, blockchain can facilitate the implementation of smart contracts to automate and enforce security policies, reducing the likelihood of human error and ensuring consistent application of security measures.

The ongoing evolution of cyber threats necessitates continuous advancements in security technologies and methodologies. Organizations must remain vigilant and proactive in adopting cutting-edge solutions such as blockchain to safeguard their data assets effectively. In summary, addressing the multifaceted challenges of data breaches requires a holistic approach that combines technological innovation, stringent security policies, and comprehensive employee training. Implementing blockchain technology can significantly enhance the security posture of organizations by providing a robust framework for data protection, ensuring transparency and accountability, and enabling rapid detection and response to security incidents.

Furthermore, collaboration between industries and regulatory bodies is essential to establish standardized practices and guidelines for the secure implementation of blockchain technology. As blockchain continues to evolve, its integration with other emerging technologies such as artificial intelligence and machine learning could further enhance its capabilities, providing more sophisticated tools for threat detection and prevention. By leveraging the strengths of blockchain and other advanced technologies, organizations can build a resilient and secure infrastructure that effectively mitigates the risks associated with data breaches.

In conclusion, the implementation of blockchain technology presents a promising avenue for improving data security across various sectors. Its unique attributes make it a powerful tool for protecting sensitive information, particularly in environments where data integrity and transparency are paramount. As cyber threats continue to evolve, embracing innovative solutions like blockchain will be crucial in maintaining the security and trustworthiness of data systems. A novel dual-backend system architecture that physically and

logically separates routine medical data operations from security analysis and log monitoring, enhancing security oversight was proposed in this study. Introducing a method for generating a synthetic medical dataset tailored to the Kazakhstani context, including regional naming conventions and ID number formats, provides a valuable resource where real data is inaccessible. An end-to-end, fully containerized (Dockerized) implementation of the system, demonstrating its practical feasibility and deployment pathway was presented in this paper.

The main contributions of this study are:

1. A hybrid architecture combining off-chain medical data storage with on-chain immutable logging to ensure GDPR/local law compliance.

2. A custom synthetic dataset generator tailored to the specific data formats of the Republic of Kazakhstan.

3. A comparative analysis of ML algorithms for detecting insider threats within blockchain-secured logs.

2. Literature Review

In the evolving landscape of healthcare data management, the integration of blockchain technology has emerged as a pivotal advancement, promising enhanced security, privacy, and efficiency. This literature review delves into the seminal works that have shaped the current understanding and application of blockchain within the realm of healthcare, emphasizing key innovations, the synthesis of blockchain with artificial intelligence (AI), and the practical implications of these technologies.

In the evolving landscape of healthcare data management, the integration of blockchain technology has emerged as a pivotal advancement, promising enhanced security, privacy, and efficiency. At the forefront of blockchain application in secure digital signatures is the Keyless Signature Infrastructure (KSI) technology utilized in Estonia [3]. This infrastructure leverages blockchain to ensure the integrity and authenticity of digital signatures without relying on traditional public-key infrastructure, thereby enhancing security and scalability. Building on this foundation, recent studies have explored various applications and innovations in healthcare leveraging blockchain technology. For instance, Naik et al. [4] developed SecureHealth, a blockchain-based healthcare application designed to improve data security and confidentiality. The application uses digital signatures and immutable blockchain records to ensure the integrity of patient data, aligning with the principles of KSI.

Further advancing the integration of blockchain in healthcare, Egala et al. [5] identified key determinants for the application of blockchain technology in primary healthcare delivery. Their study utilized an integrated best-worst approach to evaluate these determinants, highlighting the practical deployment challenges and benefits of blockchain in enhancing healthcare services. A systematic review by Merlo et al. [6] examined the exploitation of blockchain technology in the healthcare sector. The review discussed various applications of blockchain, such as improving data security, interoperability, and patient privacy. This comprehensive analysis underscores the potential of blockchain to revolutionize healthcare by providing a secure and transparent method for managing health data.

The editorial by Jurdak et al. [7] focused on the development of sustainable and secure healthcare systems using blockchain technology, emphasizing the importance of secure digital signatures and data integrity in healthcare applications. Collectively, these studies illustrate the progressive adaptation and implementation of blockchain technology in healthcare, addressing both theoretical and practical challenges. The integration of blockchain with KSI and other advanced technologies continues to enhance the security and efficiency of healthcare data management systems, paving the way for more reliable and patient-centered healthcare solutions.

The exploration of blockchain scalability and adaptability is exemplified by the development of the Alphabill platform, as presented by Buldas et al. [8]. This innovative blockchain technology addresses both unlimited scalability and unrestricted adaptability through the introduction of KSI Cash, a sharded blockchain technology with a new electronic money scheme, the bill scheme. The performance tests conducted in collaboration with the European Central Bank demonstrated the system's capability to operate with 100 million wallets and 15,000 transactions per second under realistic conditions, highlighting its efficiency and minimal carbon footprint of 0.0001g CO_2 per transaction. Furthermore, the system showcased the ability to handle up to 2 million payment orders per second in a laboratory setting, reinforcing the claim of unlimited scalability. The Alphabill platform also introduces a universal tokenization architecture, enabling seamless asset transfer and exchange, and includes innovative features such as a dust collection solution and a scalable multi-asset atomic swap solution. This research underscores the potential of advanced blockchain technologies to meet the growing demands for scalable, secure, and efficient digital financial systems, positioning Alphabill as a significant milestone in the evolution of blockchain applications in various domains, including healthcare [8].

The integration of keyless signature infrastructure (KSI) with blockchain technology for securing e-health records is explored by Nagasubramanian et al. [9]. Their study addresses critical issues in healthcare data management, such as confidentiality, integrity, and scalability of electronic health records (EHR). The proposed system utilizes KSI to ensure digital signatures' secrecy, enhancing authentication, and employs blockchain to maintain data integrity. The framework leverages Fast Healthcare Interoperability Resources (FHIR) standards, managed by Health Level Seven International (HL7), to facilitate interoperability. By comparing parameters like average time, size, and cost of data storage and retrieval, the study shows that the proposed system reduces response times by approximately 50% and decreases storage costs by about 20% compared to conventional data storage techniques. Additionally, the system's ability to handle data availability issues and ensure continuous accessibility, even in cases of hardware failure or user error, is highlighted. The use of blockchain technology in this framework ensures that health

data is stored in a decentralized manner, enhancing security and transparency. The adoption of Merkle tree structures for log maintenance and timestamping further bolsters data integrity and traceability. This comprehensive approach not only addresses the current limitations of centralized storage systems but also paves the way for more secure and efficient healthcare data management practices. The proposed KSI-Blockchain (KSIBC) framework represents a significant advancement in securing sensitive healthcare information, offering a scalable and cost-effective solution that enhances both data integrity and confidentiality in the cloud environment. This work demonstrates the potential of combining KSI and blockchain to create a secure, efficient, and scalable solution for managing health records in the cloud, thereby addressing significant challenges in data security and privacy in healthcare [9].

The potential of Keyless Signature Infrastructure (KSI) to enhance authentication management in private blockchains is thoroughly examined by Gyeong-Jin Ra and Im-Yeong Lee [10]. This study addresses the inefficiencies and security risks associated with traditional Public Key Infrastructure (PKI) and Accountable Key Infrastructure (AKI) by proposing a KSI-based authentication framework tailored for private blockchains. Unlike PKI, which relies on a central trusted third party and is prone to single points of failure, the KSI approach utilizes hash chains and Merkle trees to ensure data integrity and non-repudiation without relying on asymmetric key pairs. This method significantly improves the robustness and reliability of the blockchain network by decentralizing the authentication process, thus enhancing resistance to man-in-the-middle attacks and other common security threats. The proposed system demonstrates efficiency by reducing computational overhead and increasing transaction speeds, which is crucial for private blockchains where rapid block generation and verification are essential. Through comprehensive performance analysis, Ra and Lee highlight the system's ability to maintain high levels of security and efficiency, making it a viable solution for secure, scalable, and efficient blockchain applications in various fields, including healthcare [10].

The secure and auditable logging infrastructure based on a permissioned blockchain is examined by Putz et al. [11]. This study addresses the challenges of maintaining log integrity and auditability in information systems, which are crucial for legal admissibility in forensic investigations. The proposed system uses a blockchain to store non-repudiable proofs of existence for all generated log records, ensuring that no modifications occur during processing. Unlike traditional methods, this approach does not depend on trusted third parties or specialized hardware, significantly reducing costs and implementation complexities. The infrastructure employs a permissioned blockchain, providing higher throughput and lower latency compared to public blockchains. The prototype developed within the DINGfest project demonstrates the feasibility of this system, achieving transaction rates of up to 3,500 per second. The system's design incorporates both on-chain and off-chain storage, enhancing scalability and privacy by storing log data locally while using the blockchain for integrity proofs. This hybrid approach ensures that log records are both tamper-resistant and confidential, addressing key security and performance requirements. Putz et al.'s work highlights the potential of blockchain technology to provide robust and efficient solutions for secure logging, paving the way for more reliable and transparent audit processes in various sectors, including healthcare [11].

In addressing the vulnerabilities of conventional log management systems, Huang, Li, and Zhang developed a blockchain-based logging system that employs a consensus mechanism to ensure data integrity [12]. This innovative approach diverges from J¨amthagen and Hell's focus, emphasizing blockchain's versatility in enhancing log security across healthcare data management dimensions. Salah et al. offer a broader perspective by examining the synergistic relationship between blockchain and AI in healthcare [13], contrasting with the specific focus of Huang, Li, and Zhang. Their review identifies the urgent need for research to overcome challenges in data privacy and scalability, advocating for an interdisciplinary approach that melds blockchain and AI into a comprehensive healthcare data security framework.

Pourmajidi and Miranskyy's Logchain project [14] and the SecNet framework by Wang et al. [15] represent distinctive efforts to augment data security and system intelligence through the integration of blockchain and AI. Logchain emphasizes log storage immutability, while SecNet harnesses AI for dynamic security protocol generation, collectively broadening the research scope in blockchain and AI for healthcare. The exploration of blockchain's potential in healthcare continues with Barbaria et al.'s study on leveraging patient information sharing through blockchain-based distributed networks [16]. This work delves into the deployment of blockchain technology within hospital information systems to fortify data integrity, availability, and privacy, incorporating patient consent into data-sharing controls and addressing critical privacy concerns effectively. Like this, the paper by Jennath, Anoop, and Asharaf's explores the potential for creating trusted blockchain-based AI models in eHealth [17]. They propose a transparent platform for consent-based data sharing. The blockchain's audit trail allows the data owner to understand how the data is exposed.

The integration of blockchain and AI in healthcare is further exemplified by Bhavya et al.'s investigation into strengthening healthcare systems against data breaches and unauthorized access [18]. Their framework employs blockchain's immutable nature for data integrity and AI-driven algorithms to ensure reliable and precise healthcare data processing, highlighting the innovative approaches towards enhancing data security and improving decision-making processes within healthcare systems. Pourmajidi and Miranskyy's Logchain initiative proposes a blockchain-assisted log storage system [14], introducing a novel approach to journal management that ensures the immutability and integrity of log data. This method addresses the challenges of traditional log management systems, offering a tamper-proof and transparent solution for storing critical logs, and signifies a leap towards enhancing the security and reliability of healthcare data management.

The transformative potential of blockchain in the healthcare industry is further highlighted by Palani et al. [19],

who propose an Ethereum blockchain-based solution aimed at revolutionizing healthcare data management. Their research underscores the critical issue of data breaches, presenting a decentralized system that ensures the secure storage of health records. This approach effectively mitigates risks associated with centralized data repositories and unauthorized data manipulation, granting patients full control over their data and promoting transparency in the healthcare journey.

In addressing the challenges of log management in the context of blockchain applications, Klinkmu¨ller et al.'s introduction of the Ethereum Logging Framework (ELF) [20] provides a cost-efficient method for logging and data extraction. This framework facilitates the embedding of compact logging code into smart contracts, enhancing data throughput at reduced costs. Although its application is limited to Ethereum-based platforms, ELF's contribution underscores the need for efficient data management solutions in blockchain applications, highlighting the ongoing efforts to address scalability and efficiency. The security of traditional log systems is critically examined by Jiansen Huang, Hui Li, and Jiyang Zhang [12], who leverage blockchain technology to enhance log system security through a voting-based consensus algorithm. This method ensures log integrity and security, marking a significant improvement over centralized log storage solution. The blockchain-based approach presents a scalable, tamper- proof solution, although it may face challenges related to blockchain's inherent limitations, such as consensus time and increased latency.

Salah et al.'s comprehensive review on the convergence of AI and blockchain illuminates the potential for these technologies to automate payments in cryptocurrency and provide secure, decentralized data access [13]. This work categorizes blockchain applications in AI, identifying key areas where blockchain supports AI's decision-making capabilities, and highlights the nascent stage of research at this intersection, emphasizing the need for empirical studies to validate proposed applications and address scalability and data privacy challenges. Tagde et al.'s investigation into the integration of blockchain and AI technologies in e- Health underscores the transformative potential of these technologies to ensure secure, transparent access to patient records and facilitate enhanced diagnostics and care [21]. This approach promises to democratize access to healthcare and improve outcomes, though it also faces implementation challenges and scalability issues.

Thomas F. Heston's case study on Estonia's use of blockchain to secure health records showcases an innovative approach to using blockchain technology to enhance health record security and privacy [22]. This model's success in Estonia highlights the practical application of blockchain in healthcare, although scalability and adaptation to different healthcare systems remain potential challenges. Abbas et al. delve into the role of blockchain in enhancing healthcare data management [23], identifying the technology's capacity to manage patient information and pharmaceutical supply chains securely. Their comprehensive view of blockchain's benefits, alongside strategic recommendations for overcoming adoption challenges, emphasizes the need for robust risk management frameworks to navigate technological adoption hurdles effectively. Kim et al.'s development of DynamiChain [24], a blockchain-based system with a dynamic consent mechanism, addresses the challenge of securing medical data while complying with privacy regulations. This innovative system, which adapts to varying regulatory and personal privacy requirements, highlights the complexity of implementing blockchain solutions in healthcare and the need for flexibility in addressing global regulations.

This literature review examines the research on the integration of blockchain and AI technologies into healthcare data management. The reviewed studies collectively demonstrate the potential of these technologies to revolutionize healthcare systems by enhancing security, privacy, and data management. Using the example of blockchain-based logging techniques, the ongoing synthesis of blockchain and other advanced AI technologies has the potential to address the complex challenges of data security in healthcare. This could lead to safer, more efficient, and more patient-centred healthcare systems worldwide. In summary, existing literature confirms the potential of blockchain for log integrity and ML for anomaly detection. However, few studies present a complete, deployable system architecture, validate it with a regionally-specific dataset, and discuss the practicalities of a full-stack implementation. This work aims to bridge that gap by moving from concept to a concrete, reproducible system.

3. Methodology

3.1. System Architecture and Workflow

The system architecture encapsulates the interaction between the database, back-end, and blockchain components, as illustrated in Fig.4. The process initiates with data operations performed within the healthcare facility's front-end interface. Each interaction, whether data retrieval, modification, or other forms of manipulation, triggers the back-end to log the transaction details, including timestamps, patient and physician identifiers, and the type of interaction. These logs are then securely recorded onto the blockchain, ensuring an immutable and tamper-proof record of all activities.

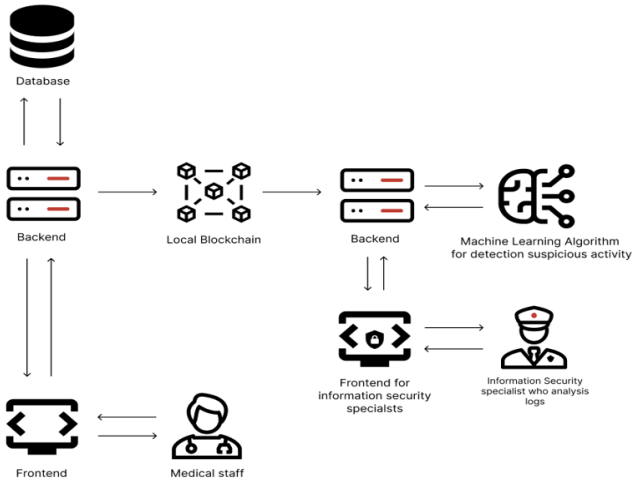

Fig. 4. Proposed project schema.

The primary back-end system interfaces directly with the local blockchain, capturing and storing every transaction initiated by the medical staff. This back-end is responsible for handling data integrity and ensuring that all interactions are logged accurately and systematically. The local blockchain serves as a decentralized ledger that provides a secure and transparent method for recording these transactions, effectively mitigating the risk of data tampering or unauthorized modifications.

Simultaneously, a dedicated platform for information security professionals is maintained. This platform includes a secondary back-end system designed to process the logged data further. Once the data is recorded on the blockchain, it is processed by this secondary back-end system, which then feeds the data to the front-end interface specifically designed for security analysis. This system enables information security specialists to monitor and analyze the logs in real- time, providing a comprehensive view of all activities within the healthcare facility.

The integrated machine learning model plays a crucial role in this architecture. The secondary back-end system processes the data logs and inputs them into the machine learning algorithm designed to detect suspicious activity. This model leverages advanced machine learning techniques to analyze patterns and identify potential data breaches or anomalies. By comparing the real-time data against predefined patterns and historical data, the model can flag any activities that deviate from the norm, classifying them as normal, suspicious, or anomalous.

The results of this analysis are then communicated back to the information security specialists through the dedicated front-end platform. This platform provides detailed insights and alerts, allowing security professionals to take immediate action in response to any identified threats. The real-time monitoring capabilities facilitate proactive data breach prevention, ensuring that any potential issues are addressed promptly. Additionally, the system includes robust error handling mechanisms to manage and report any exceptions that occur during the data processing and analysis phases, ensuring that the integrity of the monitoring process is maintained.

Overall, this comprehensive architecture integrates advanced technologies, including blockchain and machine learning, to provide a robust solution for healthcare data management. By ensuring secure logging of all interactions and employing sophisticated anomaly detection techniques, the system enhances the overall security posture of the healthcare facility, ensuring the integrity and confidentiality of sensitive patient data.

3.2. Database and Back-end Implementation

In the absence of a publicly available medical dataset that could emulate real- world interactions with medical data deeply enough,the data had to be collected by artificially creating them.. The most popular male and female first and last names in Kazakhstan for the year 2023 were taken, from which it was possible to generate randomized full names.

Table 1. Medical dataset features summary.

Feature	Description
Name	Randomly generated full names based on popular names in Kazakhstan, 2023
Gender	Gender of the individual
Date of Birth	Date of birth, used to derive other features such as IIN
City	Randomly generated city address in Kazakhstan
Place of Work	Randomly generated company names
IIN	Emulated Individual Identification Number derived from date of birth and a checksum
Diagnosis	Randomly generated medical diagnoses
Phone Number	Randomly generated sequence of digits representing a phone number
Email	Randomly generated email addresses

As illustrated in the Table 1, the dataset also includes gender, date of birth, address in the form of city, places of work in the form of randomly generated companies, an emulation of the IIN based on a formula derived from the date of birth and a randomly generated checksum, diagnoses, phone numbers in the form of randomly generated sequences of digits, and email addresses. This dataset contributes to a comprehensive representation of medical scenarios, thus emulating the database of a medical organization.

The back-end architecture plays an important role in enhancing security measures within healthcare establishments. A fundamental aspect of the server infrastructure is its capacity to emulate the practical application of the platform intended for healthcare organizations. This emulation encompasses the registration and management of patient and physician profiles, crucial to the digital health system's operations. Through this simulation, healthcare professionals are empowered to manage and consult patient information, thereby underpinning a realistic framework for evaluating and refining security and anomaly detection strategies. Central to the approach for ensuring data integrity is the immutable logging of data interactions on the blockchain. The back-end system is instrumental in recording each data interaction, encompassing access, modification, or deletion, thereby establishing a verifiable and secure audit trail. Moreover, this infrastructure is responsible for the retrieval of these logs and their subsequent dissemination to information security analysts for examination. This evaluation is essential for the identification and mitigation of potential security vulnerabilities or anomalies. Additionally, the server infrastructure is tasked with the reformation of these logs into a structure conducive to machine learning analysis. This reformatting process aggregates logs into sessions, capturing essential details such

as request frequency, unique patient engagements, and associated IP addresses, while also integrating an anomaly detection attribute. This systematic approach ensures the generation of a dataset optimized for the development and testing of machine learning models aimed at enhancing data security protocols.

3.3. Blockchain Integration

Central to the data integrity strategy is the implementation of smart contracts, coded in Solidity. The adoption of blockchain for log storage leverages its ability to protect data from tampering and unauthorized changes. Blockchain's immutable ledger ensures the permanence of recorded data, establishing a verifiable and unalterable log repository. Was employed Ganache for blockchain deployment, enabling an Ethereum-like blockchain simulation in a local setup. This method eliminates the need to deploy contracts on the public network, facilitating rapid prototyping and controlled testing. Smart contract development was conducted within the Hardhat environment, optimizing contract coding and debugging.

The smart contract architecture supports essential functions such as addLog, getLog, and getLogs, streamlining the precise logging and retrieval of data interactions. Log entries meticulously detail each interaction, capturing the action (GET, PUT, DELETE), timestamp, doctor's ID, patient's ID, and doctor's IP address. This detailed log structure not only improves audit trails but also strengthens the overall security framework of the system.

Figure 5 illustrates the back-end interfaces with the blockchain through a client-server model, employing Remote Procedure Calls (RPC) to facilitate communication. In this model, the back-end serves as the client, initiating RPC requests to the blockchain network. These requests, analogous to function calls in conventional programming paradigms, are transmitted over the network. Upon receipt, the blockchain processes the requests and returns responses, thereby enabling interaction between the back-end system and the blockchain infrastructure through a structured protocol of network requests and responses.

The integration of blockchain technology with the medical data platform is facilitated through two primary mechanisms, both of which involve RPC requests initiated by the back-end system to interact with the blockchain network.

The first mechanism is activated following a doctor's interaction with patient data, be it through reading or updating records. Subsequent to such an interaction, the back-end system dispatches an RPC request to the blockchain, instructing it to record a log of the said action.

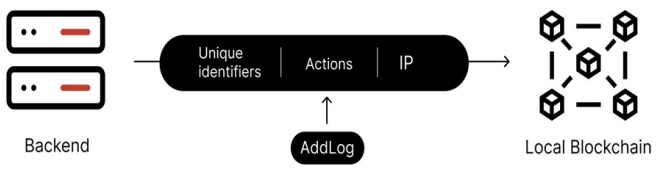

Fig. 5. RPC request from back-end to blockchain.

The second mechanism comes into play when an information security specialist seeks to audit these interactions. In this scenario, the back-end system again utilizes RPC requests to retrieve the logged actions from the blockchain. These logs are then processed and presented through the user interface (UI), allowing security specialists to review and analyze the interactions for any potential anomalies or breaches. To ensure the integrity and confidentiality of transactions on the blockchain, access control mechanisms are implemented within smart contract logic. This design paradigm establishes an ownership model, where the smart contract is assigned a specific owner, typically the back-end infrastructure managing the contract. Consequently, this architecture restricts data access: only transactions initiated from the back-end, authenticated through its unique private key, are permitted to write or read data. Access attempts by any other entity, despite possession of a private key, are systematically denied, thereby preserving the exclusivity of interaction to the designated back-end system.

The accompanying Fig. 6 vividly illustrates the mechanism of preventing unauthorized access to the blockchain, serving as a testament to the robust defense scheme embedded within the system.

3.4. Automatization Process

Data entered into the blockchain undergoes a meticulous processing and conversion procedure to ensure its suitability for machine learning (ML) analysis. This process is pivotal for transforming raw data into a structured format that can be effectively utilized by various ML algorithms. Sessions are initiated every three hours, during which comprehensive statistics are aggregated.

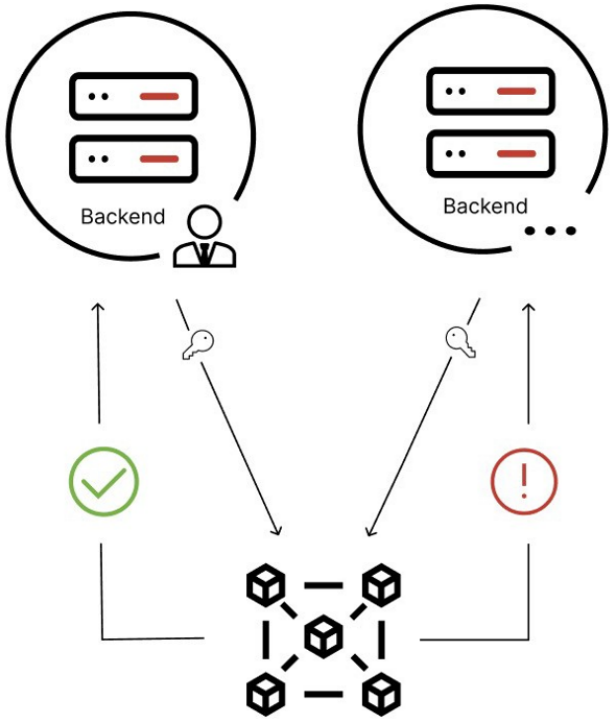

Fig. 6. Preventing unauthorized access to the blockchain.

These statistics include the number of requests, unique doctor identifiers, and IP addresses used. Each session is meticulously analyzed, incorporating an 'anomaly' feature designed to categorize activities into normal, suspicious, or anomalous. This feature significantly enhances the system's security posture by enabling real-time detection and categorization of potential security threats.

The dataset generated from these sessions comprises approximately 25,000 records, providing an extensive and accurate reflection of the system's operational and transactional dynamics. This dataset serves as a robust foundation for the development and evaluation of ML models, offering deep insights into the behavioral patterns within the system and facilitating the identification of potential anomalies. The structured dataset, as summarized in Table 2, is essential for understanding the intricate details of the system's operations.

To process this dataset, was leveraged the capabilities of Scikit-learn's comprehensive ML pipeline, which facilitated an extensive exploration of various models, including Logistic Regression, Support Vector Classification (SVC), and KNeighborsClassifier. The selection process for these models was guided by a rigorous evaluation based on multiple performance metrics such as accuracy, F1 score, cross-validation results, and an overfit indicator. These metrics are critical for assessing the predictive capabilities of the models and ensuring their robustness. The accuracy metric evaluates the model's overall ability to correctly classify data points, calculated as shown in Equation 1. The weighted F1 score, detailed in Equation 2, balances precision and recall, providing a more nuanced evaluation of model performance across different classes.

Table 2. Medical dataset features summary.

Feature	Description	Type
date	Session timestamp	datetime
doctor_id	Doctor identifier	int64
requests_per_session	Data requests count	int64
used_ip_addresses	Unique IP addresses count	int64
anomaly	Suspicious activity indicator (0: Normal, 1: Suspicious, 2: Anomaly)	int

To enhance the generalizability of these models, cross-validation techniques were employed. This involves partitioning the dataset into training and validation sets multiple times to evaluate model performance consistently. The cross- validation F1 score, denoted as μWeighted in Equation 3, averages the F1 scores across all validation folds, ensuring a reliable assessment.

$$Ovrft\ ind. = /CV\ F1\ \mu_{Training} - CV\ F1\ \mu_{Validation}/ \quad (1)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$F_1 weighted = 2\frac{Precision_{weighted} * Recall_{weighted}}{Precision_{weighted} + Recall_{weighted}} \quad (3)$$

This multi-faceted approach, integrating advanced ML algorithms, blockchain technology for data integrity, and a high-performance back-end infrastructure, sets a new benchmark for security in healthcare data management. By delivering a system that not only detects anomalies with high accuracy but also ensures the indelible recording of data interactions, was pave the way for a future where healthcare data security is uncompromised and fully accountable. The robust architecture and comprehensive analysis techniques employed ensure that the system is both reliable and scalable, capable of meeting the stringent demands of modern healthcare environments.

The practical evaluation of these models involved the use of a Flask-based back-end framework to execute machine learning predictions on blockchain- derived data. This back-end was specifically designed to interface with blockchain systems to extract transaction logs, which were subsequently formatted into a structured dataset for anomaly detection using a pre-trained machine learning model. The Flask application acted as the back-end server with a dedicated endpoint /predict that handled POST requests. It accepted JSON-formatted data representing blockchain logs, primarily detailing doctor activities. Each log entry typically included identifiers for doctors and patients, along with metadata such as IP addresses used during the transactions.

Upon receiving the data, the application performed several aggregation operations. It tallied the total number of requests or transactions performed by each doctor, facilitating a measure of activity volume. It tracked the number of unique patients each doctor had interacted with, which helped in assessing the breadth of doctor-patient interactions. It recorded and counted unique IP addresses to monitor the diversity of network locations from which the doctor accessed the blockchain, adding a layer of scrutiny for potential security concerns. The Flask application leveraged numpy and pandas for numerical operations and data handling. The aggregated data points (number of requests, unique patients, and IP addresses) were structured into a feature array. This array was then scaled using a pre-loaded scaler object to match the input format expected by the machine learning model. The scaled features were fed into a pre-trained Support Vector Machine (SVM) model, loaded via joblib, to predict the likelihood of anomalous behavior. Each prediction was mapped to a descriptive label (Normal, Suspicious, Anomalous) based on the model's output.

For each doctor, a result object was constructed containing the doctor's ID, the anomaly prediction label, and aggregated metrics such as the number of requests, unique IP addresses, and unique patients. This result was then formatted into a JSON response, providing a comprehensive view of each doctor's activities and the associated risk assessment.

The application included robust error handling to manage and report exceptions during the data processing and prediction phases, ensuring that any operational issues were communicated back to the client effectively. It was configured to run on a local server with debugging enabled, facilitating development and testing. It listened on port 5000, allowing for straightforward integration with local development environments or deployment setups. This comprehensive analysis and practical implementation underscore the importance of selecting the right ML model based on the specific requirements and characteristics of the data at hand, highlighting the potential of advanced machine learning techniques in sophisticated anomaly detection scenarios where high accuracy and low overfitting are paramount

3.5. Medical Dataset Collection and Database Configuration

At first, was decided to use patient data from Kaggle, which included information such as first name, last name, phone number, and email. However, was soon realized that this dataset was not comprehensive enough for research needs. To address this, was opted for an alternative approach to obtaining patient data by developing a Python script responsible for generating synthetic medical data.

This Python script is designed to simulate a realistic dataset for research purposes. The data generation process is implemented through several key functions that create individual attributes such as dates of birth, individual identification numbers (IINs), telephone numbers, and email addresses. Each function is crafted to ensure the authenticity and variability of the generated data, closely mimicking real-world medical datasets. This approach allows a robust and detailed dataset to be generated, tailored specifically to the requirements of the research domain.

The script initiates by creating an empty list named *medical_data*, which serves as the repository for each generated medical record. A looping construct iterates over a predetermined number of iterations (set by *num_entries*), with each iteration responsible for the generation and assembly of a single patient's record. At each iteration, the script selects a gender randomly from the predefined list of genders, influencing subsequent selections of first names and surnames from gender-specific lists. The middle name is dynamically generated based on the selected gender, adhering to cultural naming conventions. The *generate_dob*() function is invoked to produce a random but plausible date of birth within the specified range, utilizing the Python's *datetime* module for accurate date manipulation. The *generate_iin_corrected*(*dob*) function constructs an IIN based on the generated date of birth. It intelligently adjusts the IIN by incorporating a century digit indicative of the birth year and follows with a random numeric sequence to complete the IIN format, ensuring uniqueness and realism akin to actual identification numbers. Unique contact details are fabricated using the *generate_phone_number*() and *generate_email*(*name*) functions. These functions ensure the generation of believable Kazakhstani phone numbers and email addresses, the latter of which combines elements of the individual's name with random digits and a choice of domain, reflecting common email format practices. Each record is further enriched with randomly selected city and workplace data from predefined lists, mirroring the probable diversity of an individual's living and working environment. A medical diagnosis is also randomly assigned from a list of common conditions, adding a critical layer of health-related data to each record. All individual attributes generated and gathered comprising full name, gender, date of birth, city, workplace, IIN, medical diagnosis, phone number, and email are consolidated into a comprehensive list format and appended to the *medical_data* list. Upon completion of the data generation loop, the accumulated data in medical_data is transferred into a pandas *DataFrame*. This *DataFrame* is structured with designated column labels corresponding to each attribute, ensuring organized data storage. The complete dataset is then exported to a CSV file, named *Kazakhstan_Medical_Dataset.csv*, facilitating easy access and usability for subsequent data analysis tasks.

Figure 7 shows an artificial medical dataset created by the script and intended for use in medical research simulators. The dataset, formatted as a CSV file, includes multiple fields such as name, gender, date of birth, city, workplace, individual identification number (IIN), diagnosis, phone number, and email. It contains 200 patients and each record is carefully crafted to provide variety and realism, reflecting the variability expected in real medical records.

Next, the generated CSV file containing patient data needed to be transferred to a database for further use in the platform. After careful consideration, PostgreSQL was selected as the database system due to its effectiveness, reliability, and widespread adoption. PostgreSQL is well-regarded for its robust performance and extensive feature set, making it a popular choice among developers and data professionals. Moreover, familiarity with PostgreSQL allowed its full potential to be leveraged efficiently. The data transfer process from a CSV file to a PostgreSQL database involves several meticulous steps to ensure accuracy and efficiency. Initially, was parsed the CSV file, which stores data in a comma-separated format, by writing a script capable of reading the file, splitting the data based on commas, and mapping each value to the corresponding column in the PostgreSQL table. This scripting ensured that data was transferred accurately and systematically from the CSV file to the database.

Kazakhstan_Medical_Dataset

Name	Gender	Date of Birth	City	Place of Work	ИИН	Diagnosis	Phone Number	Email
Saltanat Aidarkyzy Aigerim	Female	1991-08-11	Almaty	Y-Holding	91081136234	Anemia	77851925943	saltanat26@test.org
Alina Zhanatkyzy Dana	Female	1983-03-22	Astana	X-Corp	83032238345	Anemia	77808098795	alina64@test.org
Dana Nurkyzy Madina	Female	1985-10-31	Almaty	E-Bank	85103132526	Influenza	77958947815	dana12@example.com
Nurlan Aidaruly Kuanysh	Male	1979-05-02	Almaty	Y-Holding	79050235161	Type 2 Diabetes	77582493811	nurlan70@test.org
Nurzhan Aslanuly Arman	Male	1987-05-10	Almaty	W-Media	87051038284	Gastritis	77517821194	nurzhan27@example.com
Timur Rustemuly Ruslan	Male	1978-09-22	Astana	E-Bank	78092236686	Migraine	77547865749	timur57@example.com
Talgat Maratuly Talgat	Male	1979-10-20	Astana	W-Media	79102038676	Migraine	77624209758	talgat38@test.org
Arman Madiuly Kuanysh	Male	1986-01-18	Almaty	Z-Health	86011833883	Dermatitis	77392560004	arman43@example.com
Kuanysh Nuruly Nurzhan	Male	2000-09-04	Almaty	Z-Health	90454133	Hypertension	77832558026	kuanysh86@demo.net
Nurlan Zhanatuly Nurzhan	Male	1981-06-22	Astana	E-Bank	81062239237	Dermatitis	77594289152	nurlan64@test.org
Karina Nurkyzy Saltanat	Female	1970-05-08	Almaty	Y-Holding	70050838347	Dermatitis	77151455377	karina96@test.org
Talgat Rustemuly Daulet	Male	1999-07-25	Astana	Z-Health	99072539244	Influenza	77925255753	talgat55@test.org
Aigerim Rustemkyzy Zhansaya	Female	1989-07-10	Almaty	Y-Holding	89071032281	Common Cold	77356905126	aigerim79@demo.net
Talgat Nuruly Daulet	Male	1999-11-26	Almaty	X-Corp	99112639794	Chronic Kidney Disease	77305296210	talgat58@example.com
Daulet Olzhasuly Nurlan	Male	1982-08-26	Astana	X-Corp	82082634496	Chronic Kidney Disease	77158545253	daulet40@demo.net
Saltanat Daurenkyzy Aigerim	Female	1989-08-31	Astana	W-Media	89083132857	Type 2 Diabetes	77901552993	saltanat66@demo.net
Bibigul Olzhaskyzy Svetlana	Female	2004-02-22	Astana	W-Media	4022252295	Hypertension	77532376911	bibigul63@demo.net
Timur Zhanatuly Ruslan	Male	1993-08-31	Astana	X-Corp	93083137635	Hypertension	77337841598	timur19@example.com
Daulet Zhanatuly Zhanbolat	Male	1992-04-22	Astana	Z-Health	92042233510	Type 2 Diabetes	77930873795	daulet85@demo.net
Svetlana Zhanatkyzy Madina	Female	1970-04-03	Almaty	E-Bank	70040338631	Gastritis	77237857138	svetlana78@demo.net
Alina Olzhaskyzy Gulnara	Female	1971-07-31	Almaty	Y-Holding	71073131979	Dermatitis	77918075864	alina52@test.org
Kuanysh Madiuly Arman	Male	2008-12-09	Almaty	W-Media	8120956255	Migraine	77484588819	kuanysh68@test.org
Saltanat Aidarkyzy Dana	Female	1992-09-16	Astana	Y-Holding	92091638630	Gastritis	77325475439	saltanat27@demo.net

Fig. 7. Artificially created emulation of medical dataset in Kazakhstan.

Although the dataset is synthetic, its schema was strictly designed to mirror the structure of real Electronic Health Records (EHR) compliant with national standards. Specifically, the 'IIN' field follows the cryptographic generation algorithm used by the Ministry of Justice of the Republic of Kazakhstan (containing birth date and checksum logic), and the diagnostic codes are formatted to resemble ICD-10 entries used in hospital information systems. This structural fidelity ensures that while the behavioral patterns are simulated, the data processing pipeline remains transferable to legitimate hospital databases without significant architectural changes.

To achieve this transfer, was leveraged SQL scripts to manage the data insertion process directly within PostgreSQL. The process commenced with reading the CSV file using PostgreSQL's built-in capabilities, which facilitated the efficient handling of large datasets. Then proceeded to create the database table by defining a schema that matched the structure of the CSV data, ensuring seamless data integration. Finally, SQL was executed commands to load the data from the CSV file into the PostgreSQL table, meticulously verifying each step to maintain data integrity and consistency throughout the transfer process.

By following these steps, was ensured that the data was accurately parsed, the database schema was appropriately defined, and the data was correctly inserted into the PostgreSQL table, thereby achieving a reliable and systematic transfer from the CSV file to the database.

The sql code shows an example of the script used for this purpose, illustrating the creation of a table and copying data into this table from a CSV file.

The 'anomaly' feature in our dataset was generated based on simulated threat scenarios common in healthcare IT. Normal activity was defined as a doctor accessing a moderate number of their own patients' records from a stable IP address. Suspicious activity included accessing an unusually high number of records in a single session (potential data exfiltration). Anomalous activity involved high-volume access from multiple IP addresses simultaneously (potential credential compromise). Although the patient profiles are synthetic, the data fields (Name, IIN, Diagnosis, etc.) are structured to mirror typical Electronic Health Record (EHR) formats. The interactions logged (GET, PUT) are analogous to standard Create, Read, Update, Delete (CRUD) operations found in any modern Hospital Information System (HIS). We acknowledge that a synthetic dataset cannot capture the full complexity and noise of real-world user behavior. The patterns are inherently cleaner and more distinct than would be found in a live environment. Therefore, the high performance of our models should be interpreted as a successful proof-of-concept, and the results' generalizability to real hospital systems requires further validation with anonymized, real-world data.

3.6. Back-end and Blockchain Interaction

The back-end of the platform is divided into two main components: a platform for managing medical data and a platform for viewing and analyzing logs, including identifying anomalies within them. For both back-end applications, the Golang programming language was selected. This decision was influenced by familiarity with Golang and its efficiency in integrating with blockchain technology using libraries such as go-ethereum. Golang's concurrency model and performance make it an ideal choice for building robust and scalable back-end systems.

The development of a comprehensive platform that simulates the behavior of a real-world medical system was undertaken as part of their article. This platform allows users, in the role of doctors, to update and view patient data. The core functionalities of this application include creating, reading, updating, and deleting patient records. Each interaction with patient data is meticulously logged to ensure traceability and accountability. Whenever a doctor updates or views patient data, the application generates a log entry detailing the action. This log entry is then sent to the blockchain using the go-ethereum library. By leveraging blockchain technology, it is ensured that all actions are immutably recorded, providing a transparent and tamper-proof audit trail. This not only enhances the security of patient data but also ensures compliance with regulatory requirements for data integrity and access control.

The second component of the back-end infrastructure is designed to receive and analyze logs related to doctor activities. This application retrieves log data from the blockchain and processes it to detect any anomalies. The logic of this application involves several critical steps. The blockchain is periodically queried to fetch log data related to doctor activities, ensuring continuous monitoring of all interactions with patient data. Once the log data is retrieved, it is sent to the machine learning (ML) application. The ML application, developed using state-of- the-art anomaly detection algorithms, analyzes the log data to determine whether any activities are anomalous. Based on the analysis, the ML application provides a verdict on whether the log data indicates normal or anomalous behavior. This verdict, along with the detailed log data, is then sent back to the log analysis platform. The complete response, containing the logs and the ML application's verdict, is sent to the client application. This enables administrators and security professionals to view and act upon the analysis results in real-time.

For both back-end applications, several key technologies and methodologies were utilized to ensure their effectiveness and reliability. By using Golang and the go-ethereum library, the integration with the blockchain was streamlined, allowing seamless logging of doctor activities and efficient data retrieval. The ML application was developed using Python and leverages powerful libraries such as scikit-learn for anomaly detection. This integration allows the platform to utilize advanced analytical capabilities to identify potential security threats.

Additionally, the back-end infrastructure includes robust security measures to protect the integrity and confidentiality of the data. Access to the blockchain is controlled through smart contracts that ensure only authorized actions are recorded. Each log entry includes comprehensive metadata such as timestamps, doctor and patient identifiers, and the nature of the action performed. This detailed logging ensures

a high level of transparency and accountability within the medical platform. Furthermore, the back-end is designed to handle large volumes of data efficiently, ensuring that the platform can scale to accommodate the needs of healthcare providers of varying sizes.

Additionally, one of the most critical aspects of the article was the blockchain component of the application. The blockchain served as a secure log storage base, ensuring that the data remains immutable and cannot be maliciously modified or deleted. The primary focus was on the reliability of data storage rather than the consensus mechanism typically emphasized in blockchain technologies. This approach provided assurance that the log data would be preserved in a tamper- proof manner.

For the prototype application, Ethereum was found to be a highly convenient platform due to its robust ecosystem and well-supported development tools. Instead of deploying the application on public testnets, Ganache, a personal blockchain for Ethereum development, was utilized, allowing a local Ethereum blockchain to be run. This setup provided a controlled environment where the application could be tested and refined without the complexities and potential delays associated with public testnets.

To deploy the blockchain, Hardhat, a powerful development framework for Ethereum, was used. Hardhat facilitated the development, testing, and deployment of the smart contracts. By writing a deployment script in JavaScript, the contracts were efficiently deployed to the Ganache blockchain. This deployment process involved several key steps, including compiling the smart contracts, preparing the deployment scripts, and executing these scripts to deploy the contracts to the local blockchain.

Smart contracts were designed to handle the logging of all doctor-patient interactions within the medical platform. Each log entry recorded on the blockchain included comprehensive metadata, such as the action performed, the timestamp, and the identifiers of the involved parties. This detailed logging ensured that every interaction was transparently recorded and could be independently verified at any time.

The use of Ethereum and tools like Ganache and Hardhat significantly streamlined the development process. Ganache provided a reliable local blockchain environment where quick iterations on smart contracts and testing of their functionality were possible. Hardhat's development tools enabled the automation of many aspects of the deployment process, ensuring that the contracts were correctly deployed and interacted with the blockchain as expected.

3.7. Regulatory Compliance and Privacy

A critical design constraint for blockchain in healthcare is compliance with data protection regulations, specifically the Law of the Republic of Kazakhstan 'On Personal Data and Its Protection' and the General Data Protection Regulation (GDPR). To comply with the 'Right to be Forgotten' (Art. 17 GDPR) and local data sovereignty laws, our architecture strictly separates personal data from the immutable ledger.

1. Off-Chain Storage: All Sensitive Personal Data (SPD) and medical records are stored in a standard PostgreSQL database (off-chain). This allows for the modification or deletion of patient records upon request, fulfilling legal deletion obligations.

2. On-Chain Pseudonymity: The blockchain stores only technical metadata (access logs, timestamps, hashes of actions) and pseudonymized identifiers (e.g., Doctor ID, generic session tokens). No names, diagnoses, or IINs are written to the blockchain. This hybrid approach ensures that the system provides immutable auditability of access without violating the mutability requirements of personal data storage.

4. Results and Discussion

To assess the efficacy of various machine learning (ML) models in anomaly detection, a rigorous comparative analysis was performed, evaluating models such as Logistic Regression, Support Vector Classifier (SVC), and KNeighborsClassifier. This analysis focused on several performance metrics including accuracy, F1 score, and overfitting indices, crucial for understanding each model's capability to generalize unseen data.

The evaluation of these models was systematically structured. Logistic Regression, often used as a baseline for classification tasks, demonstrated robust performance with a test accuracy of 0.9872 and an F1 score of 0.9853. These metrics indicate a high level of precision and recall, making it a reliable choice for initial anomaly detection frameworks.

However, the Support Vector Classifier (SVC) emerged as particularly effective, achieving near-perfect accuracy and F1 scores of 0.9998. Such high values suggest that SVC is exceptionally adept at distinguishing between normal and anomalous instances, with minimal error. The overfitting index for SVC was recorded at less than 0.0001, underscoring its ability to perform well across various datasets without tailoring the model excessively to the training data.

The KNeighborsClassifier also displayed exemplary performance, attaining perfect scores in both accuracy and F1 metrics (1.0000), affirming its strength in handling anomaly detection tasks. Despite its perfect score, the KNeighborsClassifier had a slightly higher overfitting index of 0.0003, comparable to that of Logistic Regression, which may indicate a need for cautious tuning of its parameters to avoid overfitting.

These results are visualized in Fig.8, which presents a bar plot illustrating the comparative performance of the models across the discussed metrics. This visual representation aids in quickly identifying the strengths and weaknesses of each model at a glance.

Additionally, Table 3. provides a detailed breakdown of these performances, offering insights into not only the primary accuracy and F1 scores but also cross- validation metrics such as mean F1 and standard deviation of F1 from cross-validation, which further describe each model's stability across different subsets of data.

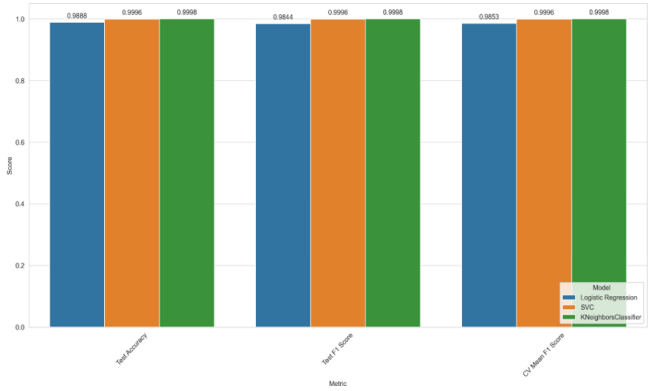

Fig. 8. Comprehensive model performance comparison.

The cross-validation F1 mean and its standard deviation are particularly telling, with SVC showing a mean F1 of 0.9994 and a remarkably low standard deviation of 0.0002, indicating high model stability and reliability. Logistic Regression and KNeighborsClassifier also showed good stability, though their standard deviations were higher, pointing to slightly less consistency across datasplits compared to SVC.

Overall, this comprehensive analysis underscores the importance of selecting the right ML model based on the specific requirements and characteristics of the data at hand. The high performance of SVC and KNeighborsClassifier, in particular, highlight their potential in sophisticated anomaly detection scenarios where high accuracy and low overfitting are paramount.

It is important to address the exceptionally high accuracy metrics (>99%) achieved by the classifiers. We acknowledge that these results are largely influenced by the nature of the synthetic dataset. Unlike real-world clinical data, which contains significant noise, missing values, and ambiguous user behaviours, the synthetic logs contain clearly defined patterns for 'normal' and 'anomalous' activities. Therefore, the reported accuracy serves as a proof-of-concept for the logical architecture of the detection module rather than a benchmark for performance on wild data. In a production environment, we anticipate a decrease in precision due to behavioral variability, which would necessitate the use of unsupervised learning methods such as Isolation Forests or Autoencoders to handle novel, undefined threat vectors.

The application for executing machine learning predictions on blockchain- derived data operates through a Flask-based back-end framework. This back-end is specifically designed to interface with blockchain systems to extract transaction logs, which are subsequently formatted into a structured dataset for anomaly detection using a pre-trained machine learning model.

The Flask application acts as the back-end server with a dedicated endpoint /predict that handles POST requests. It accepts JSON-formatted data representing blockchain logs, primarily detailing doctor activities. Each log entry typically includes identifiers for doctors and patients, along with metadata such as IP addresses used during the transactions.

Upon receiving the data, the application performs several aggregation operations. It tallies the total number of requests or transactions performed by each doctor, facilitating a measure of activity volume. It tracks the number of unique patients each doctor has interacted with, which helps in assessing the breadth of doctor-patient interactions. It records and counts unique IP addresses to monitor the diversity of network locations from which the doctor accesses the blockchain, adding a layer of scrutiny for potential security concerns.

The next step is to examine the dataset on which the training of the ML model depends. Due to the lack of logs in similar systems, a dataset was generated to emulate the data in the logs. Figure 9 illustrates the overall summary of logs in correlation with the date and amount of anomaly on those days.

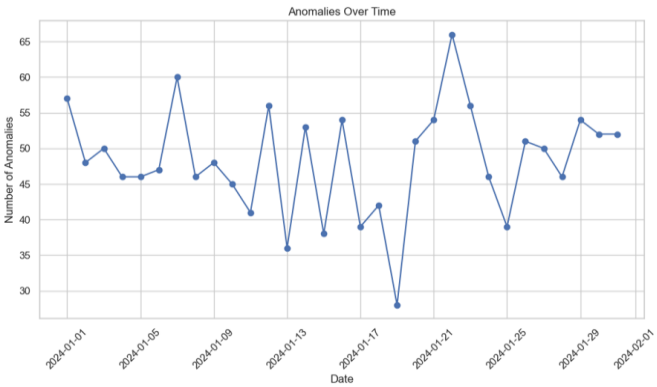

Fig. 9. Illustration of the number of anomalies by date in the dataset.

Table 3. Performance metrics of ML models.

Model	Acc. (Test)	F1 (Test)	CV F1 (Mean)	CV Std. (Dev.)	Ovft. Ind.
Logistic Reg.	0.9872	0.9853	0.9861	0.0018	0.0003
SVC	0.9998	0.9998	0.9994	0.0002	< 0.0001
KNeighbors	1.0000	1.0000	0.9995	0.0004	0.0003

The Flask application leverages *numpy* and pandas for numerical operations and data handling. The aggregated data points (number of requests, unique patients, and IP addresses) are structured into a feature array. This array is then scaled using a pre-loaded scaler object to match the input format expected by the machine learning model. The scaled features are fed into a pre-trained Support Vector Machine (SVM) model, loaded via *joblib*, to predict the likelihood of anomalous behavior. Each prediction is mapped to a descriptive label (Normal, Suspicious, Anomalous) based on the model's output.

For each doctor, a result object is constructed containing the doctor's ID, the anomaly prediction label, and aggregated metrics such as the number of requests, unique IP addresses, and unique patients. This result is then formatted into a JSON response, providing a comprehensive view of each doctor's activities and the associated risk assessment. The application includes robust error handling to manage and report exceptions during the data processing and prediction phases, ensuring that any operational issues are communicated back to the client effectively. It is configured to run on a local server with debugging enabled, facilitating development and testing. It listens on port 5000, allowing for straightforward integration with local development environments or deployment setups.

All systems are fully deployed via Docker. The deployment architecture comprises several Docker containers, each fulfilling a specific role within the system. Figure 10 illustrates all the active Docker containers on the deployment server, as listed by the docker ps command. The back-end container handles the processing of requests for both the medical and administrative platforms. It interfaces with the blockchain, database, and machine learning model to ensure seamless operations. The blockchain container hosts the deployed blockchain, which the back-end queries, securely storing all logs of actions performed on the medical platform to ensure immutability and transparency. The database container manages the database, storing critical information about doctors and patients, and supports the back-end by providing necessary data for processing and querying.

The admin front-end container, shown in Fig.11, hosts the front-end of the administrative system, providing a comprehensive view of all actions on the platform and highlighting any detected anomalies in doctors' activities. This interface is crucial for monitoring and maintaining the integrity of the medical platform. The admin front-end was built using React and is deployed using Nginx.

The medical platform front-end container, depicted in Fig.12, runs the front-end that emulates the platform's logic, allowing doctors to track their patients' information efficiently and serving as the primary user interface formedical professionals interacting with the system. Similarly, this front-end was built using React and is deployed using Nginx. For deployment, the front-end applications are built using React and then containerized using Docker. Nginx is then used as a reverse proxy to serve the built React applications. This setup ensures that the front-end applications are efficiently managed within Docker containers and served through Nginx, providing a reliable and scalable deployment environment.

Fig. 10. Illustration of docker containers on the deployment server.

Fig. 11. Illustration with patient information.

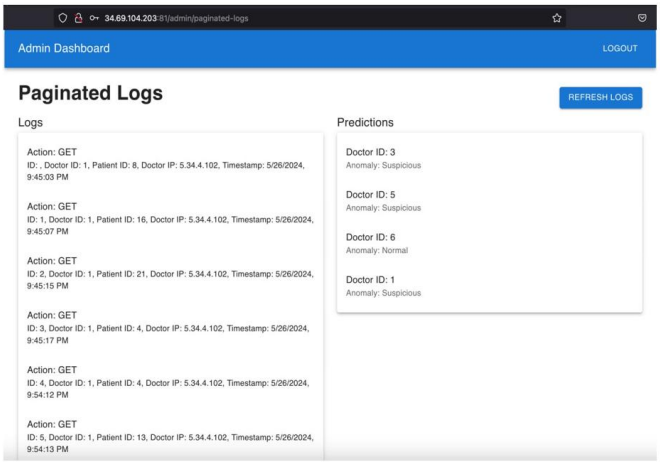

Fig. 12. Illustration from the panel with logs and predictions.

The machine learning model container receives logs from the back-end, analyzing these logs to detect the presence or absence of anomalies, and plays a crucial role in maintaining the system's security. This deployment setup ensures a robust and scalable environment for the medical platform, leveraging Docker's containerization to maintain isolation and manageability of each system component.

In our Ganache simulation, transaction latency was instantaneous. However, on a production permissioned blockchain like Hyperledger Fabric, log confirmation might take several seconds. This is acceptable for post-incident auditing but would not be suitable for real-time transaction blocking. A busy hospital could generate thousands of log entries per hour. On a public blockchain like Ethereum, the associated 'gas' fees would be prohibitive. Even on a permissioned chain, high throughput could be a challenge. A more scalable approach would involve batching—bundling hundreds of log entries and recording a single cryptographic hash of the batch on-chain. This provides the same immutability guarantee at a fraction of the cost and transactional load. Storing the full text of every log on-chain is not scalable. The aforementioned hashing approach (also known as 'anchoring') is the industry-standard solution, where logs are stored in a conventional database and only their cryptographic proofs are stored on the blockchain.

The security of our system hinges on the protection of the back-end's private key, which controls the smart contract. In a

production deployment, this key must be secured within a Hardware Security Module (HSM) to prevent theft. Doctor and administrator authentication would be managed by the hospital's existing Identity and Access Management (IAM) system, not on the blockchain itself.

Our system is designed to detect anomalous behavior from insiders, but it cannot prevent collusion. A malicious actor with legitimate access could still perform harmful actions. Furthermore, a compromised administrative account poses a significant risk. Mitigation requires strong organizational policies, such as the principle of least privilege and multi-factor authentication, in addition to this technical solution. While the logging smart contract is simple, any production-level smart contract must undergo a formal third-party security audit to check for vulnerabilities like reentrancy or integer overflows before deployment.

Handling electronic health records is subject to strict data protection laws, such as HIPAA in the US, GDPR in Europe, and Kazakhstan's Law 'On Personal Data and Its Protection.' Any implementation of this system must be fully compliant with these regulations. A key design choice to ensure compliance is that no Personal Health Information (PHI) is stored directly on the blockchain. The logs only contain metadata: pseudonymized identifiers, timestamps, IP addresses, and action types (e.g., 'GET'). This minimizes privacy risks and avoids violating regulations that grant patients the 'right to be forgotten,' which is incompatible with immutable ledgers. Our use of a local, permissioned blockchain (emulated by Ganache) is intentional. A public blockchain would be unsuitable for this use case, as it would expose metadata to the public. A private, permissioned blockchain ensures that all data remains within the control of the healthcare organization. Our experiments utilized Ganache, which offers instantaneous transaction sealing and zero cost. However, deploying this system on a public ledger like Ethereum Mainnet would be cost-prohibitive due to 'gas' fees for every log entry. For a hospital generating thousands of logs daily, a public chain is not a viable solution.

Therefore, the proposed production architecture relies on a permissioned blockchain (e.g., Hyperledger Fabric or a private Ethereum fork). In such a setup, transaction latency typically increases from instantaneous (in Ganache) to 2-5 seconds depending on the consensus mechanism (e.g., Raft or IBFT). While this introduces a slight delay in audit availability, it does not impact the real-time operation of the medical frontend. Future stress testing will focus on measuring throughput (TPS) limits on a permissioned network to ensure it handles peak hospital loads.

5. Conclusion

In conclusion, this work establishes a sophisticated framework that integrates blockchain technology and machine learning to significantly enhance the security of healthcare data systems. The use of blockchain provides an immutable and transparent log storage mechanism, ensuring that all data transactions are permanently recorded and verifiable. This method addresses fundamental vulnerabilities in traditional data management systems by preventing tampering and unauthorized access. Simultaneously, the incorporation of machine learning techniques automates the detection of data anomalies and security breaches. By systematically analyzing patterns within transaction logs, the system quickly identifies deviations that may signify potential threats, thereby facilitating rapid and accurate threat response.

The adoption of machine learning not only streamlines the detection of anomalies but also enhances the system’s ability to adapt to evolving cyber threats. As machine learning models are exposed to new data, they continuously learn and improve, thereby maintaining high detection accuracy over time. This adaptive capability is crucial in the context of healthcare, where data security needs constantly evolve due to changing technologies and attack vectors.

6. Future Work

Looking forward, the integration of deep learning technologies could significantly refine this system. Deep learning, with its ability to process and learn from vast amounts of data, could uncover more complex patterns and subtle anomalies that simpler machine learning models might overlook. This could lead to earlier detection of sophisticated breaches and more nuanced understanding of data flows, further strengthening the security measures in place. The implementation of deep learning could also enhance predictive capabilities, allowing the system to forecast potential breaches before they occur, based on emerging patterns. Thus, this article not only demonstrates the efficacy of combining blockchain and machine learning for data security but also sets the stage for future enhancements using advanced deep learning techniques to create more resilient healthcare data environments.

Our immediate priority is to partner with a healthcare institution to test the system on a real, anonymized dataset of audit logs. This will provide a true benchmark of the model's performance and reveal challenges not present in the synthetic data. To train models without centralizing sensitive data from multiple institutions, we will investigate the use of Federated Learning. This privacy-preserving technique would allow a global anomaly detection model to be trained without raw log data ever leaving the hospital's premises.

We will deploy the system on a realistic permissioned blockchain testbed (e.g., Hyperledger Fabric) to measure key performance indicators like transaction throughput, latency, and CPU/memory usage under simulated high-load conditions. We plan to conduct simulated attacks to evaluate the system's resilience. This includes stress testing against Denial-of-Service (DoS) attacks and attempting to evade the ML model with carefully crafted adversarial examples of user behaviour.

Author Contributions

A.B. drafted the manuscript and conducted the investigation; L.R. supervised the project, conceptualized the study, and reviewed the manuscript; B.T. and P.K. developed the software foundation and methodology; Z.K. and A.I.

contributed to data curation and validation; all authors reviewed and approved the final version.

Acknowledgements

The authors acknowledge the support of the Research and Innovation Center "CyberTech" at Astana IT University.

Conflict of Interest:

The authors declare no conflict of interest.

References

[1] IBM Security, "Cost of a data breach 2023," IBM Report, 2023. [Online]. Available: https://www.ibm.com/reports/data-breach

[2] J. Jiang, "Evaluation of causes of protected health information breaches," JAMA Internal Medicine, vol. 179, no. 2, p. 265, 2019. doi: 10.1001/jamainternmed.2018.5295

[3] e-Estonia, "Estonian blockchain technology (FAQ)," e-estonia.com, 2023. [Online]. Available: https://e-estonia.com/wp-content/uploads/2023-nov-nochanges-faq-a4-v03-blockchain-1-1.pdf

[4] R. Naik, A. Diwadkar, H. Amonkar, "SecureHealth: A blockchain-based healthcare application," IARJSET, vol. 10, no. 6, 2023.

[5] S. Bankuoru, A. Peter, D. Dorcas, "Determinants of blockchain technology application in primary healthcare delivery: An integrated best-worst approach," Cogent Engineering, vol. 10, no. 1, p. 2202032, 2023.

[6] V. Merlo, G. Pio, F. Giusto, and M. Bilancia, "On the exploitation of the blockchain technology in the healthcare sector: A systematic review," Expert Systems with Applications, vol. 213, Art. no. 118897, 2023.

[7] R. Jurdak, J. M. Corchado, J. Hyuk, "Editorial: Blockchain-based sustainable, secure healthcare systems," Computer Networks, vol. 214, Art. no. 109175, 2022.

[8] A. Buldas, D. Draheim, M. Gault, "An ultra-scalable blockchain platform for universal asset tokenization: Design and implementation," IEEE Access, vol. 10, pp. 77,284–77,322, 2022.

[9] G. Nagasubramanian, R. Sakthivel, R. Patan, "Securing e-health records using keyless signature infrastructure blockchain technology in the cloud," Neural Computing and Applications, vol. 32, no. 11, 2020.

[10] G. R., "A study on KSI-based authentication management and communication for secure smart home environments," KSII Trans. Internet Inf. Syst., vol. 12, no. 2, pp. 892–905, 2018.

[11] P. Benedikt, "A secure and auditable logging infrastructure based on a permissioned blockchain," Computers & Security, vol. 87, Art. no. 101602, 2019.

[12] J. Huang, "Blockchain-based log system," in Proc. 2018 IEEE Int. Conf. Big Data (Big Data), 2018. doi: 10.1109/BigData.2018.8622204

[13] K. Salah, M. H. U. Rehman, N. Nizamuddin, and A. Al-Fuqaha, "Blockchain for AI: Review and open research challenges," IEEE Access, vol. 7, pp. 10,127–10,149, 2019. doi: 10.1109/ACCESS.2018.2890507

[14] W. Pourmajidi, "Logchain: Blockchain-assisted log storage," in Proc. 2018 IEEE 11th Int. Conf. Cloud Comput., 2018. doi: 10.1109/CLOUD.2018.00150

[15] K. Wang, J. Dong, Y. Wang, and H. Yin, "Securing data with blockchain and AI," IEEE Access, vol. 7, pp. 77,981–77,989, 2019. doi: 10.1109/ACCESS.2019.2921555

[16] S. Barbaria, M. C. Mont, E. Ghadafi, "Leveraging patient information sharing using blockchain-based distributed networks," IEEE Access, vol. 10, pp. 106,334–106,351, 2022. doi: 10.1109/ACCESS.2022.3206046

[17] H. S. V. S. Jennath, S. Anoop, S. Asharaf, and H. S. Jennath, "Blockchain for healthcare: Securing patient data and enabling trusted artificial intelligence," Int. J. Interact. Multimedia Artif. Intell., vol. 6, no. 3, p. 15, 2020. doi: 10.9781/ijimai.2020.07.002

[18] G. Bhavya, M. Swetha, M. S. Muneshwara, and R. Anand, "Soft computing technique for blockchain-enabled secure healthcare system," in Proc. 2021 IEEE ICICCS, 2021. doi: 10.1109/ICICCS51141.2021.9432133

[19] U. Palani, S. S. Mahesh, D. Vasanthi, and D. Kumar, "Ethereum blockchain-based healthcare industry ecosystem," in Proc. 2020 7th Int. Conf. Smart Struct. Syst. (ICSSS), 2020. doi: 10.1109/ICSSS49621.2020.9202232

[20] C. Klinkmüller, I. Weber, A. I. Ponamorev, "Efficient logging for blockchain applications," arXiv preprint arXiv:2001.10281, 2020. [Online]. Available: https://arxiv.org/pdf/2001.10281

[21] [21] P. Tagde, S. Tagde, T. Bhattacharya, "Blockchain and artificial intelligence technology in e-health," Environmental Science and Pollution Research, vol. 28, no. 38, pp. 52,810–52,831, 2021. doi: 10.1007/s11356-021-16223-0

[22] T. F. Heston, "A case study in blockchain healthcare innovation," SSRN Working Paper, 2017. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3077455

[23] A. F. Abbas, N. A. Qureshi, N. Khan, "The blockchain technologies in healthcare: Prospects, obstacles, and future recommendations; lessons learned from digitalization," Int. J. Online Biomed. Eng., vol. 18, no. 9, pp. 144–159, 2022. doi: 10.3991/ijoe.v18i09.32253

[24] T. M. Kim, S. Lee, D. Chang, "DynamiChain: Development of medical blockchain ecosystem based on dynamic consent system," Applied Sciences, vol. 11, no. 4, p. 1612, 2021. doi: 10.3390/app11041612