International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Development of ETL pipeline for Electronic Health Record to support Machine Learning based Approaches for Security and Prediction

[1] Birendra Kumar Saraswat, [2]Neeraj Varshney, [3]P. C. Vashist

Submitted: 07/01/2024 **Revised**: 13/02/2024 **Accepted**: 21/02/2024

Abstract: Electronic Health Records (EHRs) contain a wealth of information about a patient's medical history, treatments, and health outcomes. However, the data in EHRs is often unstructured and scattered across multiple systems, making it challenging to extract meaningful insights. Developing an extract, transform, and load (ETL) pipeline for EHRs is crucial to address this challenge. This pipeline will enable the efficient integration and transformation of EHR data into a standardized format that can be used for machine learning-based approaches for security and prediction. This paper uses an ETL pipeline to identify the data sources and types of data to be extracted. These can range from structured data, such as diagnosis codes and lab results, to unstructured data, such as doctors' notes and imaging reports. Once the data sources are identified, the pipeline needs to be designed to extract the data from these sources securely and efficiently. The proposed model transforms the extracted data into a standardized format that can be used for machine learning algorithms. It involves cleaning the data, dealing with missing values, and converting it into a structured form. The proposed model obtained 96.59% accuracy, 96.36% precision, 95.64% recall, 97.56% f1-score, 96.66% false positive rate, 93.39% false negative rate.

Keywords: *Electronic Health Record, Extract, Transform, Load, Pipeline, Data Source, Machine Learning*

1. Introduction

Electronic health record security refers to the measures and protocols to protect sensitive medical information stored in electronic health records (EHRs). EHRs contain highly personal and confidential information such as patient demographics, medical history, diagnoses, medications, and lab results [1]. This information is valuable to patients and hackers who can use it for financial gain or to commit fraud. The primary concern of EHR security is to protect patient privacy. Patients trust healthcare providers with their most personal and sensitive information, and any breach of this trust can have serious consequences [2]. With proper security measures, patients can have peace of mind knowing their information is safe and only accessible by authorized personnel. Data breaches can happen for various reasons, including human error, malicious attacks, or system failures [3]. A secure EHR system can minimize the risk of data breaches by implementing robust encryption methods, access controls, and audit logs to track user activity. Health information is protected by several federal and state laws, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States. These laws mandate that healthcare providers implement appropriate security measures to protect patient data [4]. Failure to comply with these regulations can result in hefty fines and damage to the healthcare organization's reputation. In the event of a disaster or system failure, a secure EHR system can ensure the continuity of care for patients [5]. With proper backup and disaster recovery procedures, healthcare providers can quickly restore patient data and continue providing quality care without interruption [6]. EHR data can be used to identify patterns and trends in healthcare, allowing for better prediction of future health concerns and needs. However, the EHR data must be high quality and protected from tampering or fraudulent activities to ensure accurate predictions [7-8]. A secure EHR system can help maintain the integrity of the data and ensure that the predictions are reliable. Electronic health record security is critical for protecting patient privacy, preventing data breaches, complying with regulations, ensuring continuity of care, and enabling predictive analytics [9-10]. Healthcare organizations must continuously assess and improve their EHR security measures to avoid potential threats and protect their patients and businesses [11].

Machine learning is a branch of artificial intelligence that focuses on algorithms and statistical models used to analyze and interpret data to make predictions and decisions without explicit instructions [12]. In the context of Electronic Health Records (EHRs), machine learning can be extremely helpful in enhancing both security and prediction. Machine learning algorithms can detect patterns and anomalies in large sets of data, which can be especially useful in detecting potential security breaches

[1]*Computer Sciences & Engineering, GLA University, Mathura -281406, U.P., India.*
saraswatbirendra@gmail.com
[2]*Department of Computer Engineering and Applications GLA University, Mathura ,UP, India*
neeraj.varshney@gla.ac.in
[3]*Department of Information Technology, GL Bajaj Institute of Technology and Management, Greater Noida- 201306, U.P., India.*
pcvashist@gmail.com

in EHRs [13-14]. These algorithms can learn normal behavior and identify abnormal or suspicious activity, such as unauthorized access to patient records or attempted hacks [15]. It can help healthcare organizations quickly identify and respond to potential threats, significantly enhancing the security of EHRs. It can also assist in predicting and preventing potential security breaches [16]. By continuously analyzing and learning from EHR data, these algorithms can identify vulnerabilities and potential risks in the system [17]. It can allow healthcare organizations to proactively address and mitigate these risks before they can be exploited [18].

Machine learning can also improve the authentication process for EHR access by using data such as typing patterns and user behavior to verify the user's identity, making the system more secure against unauthorized access [19]. It can also greatly enhance the predictive capabilities of EHRs. These algorithms can identify patterns and trends to help predict and prevent potential health issues by analyzing vast amounts of patient data. For instance, machine learning can detect when a patient may be at risk for developing a specific condition, allowing healthcare providers to intervene early and provide preventive care. It can ultimately improve patient outcomes and reduce healthcare costs. Machine learning can be a powerful tool in enhancing the security and prediction capabilities of Electronic Health Records [20]. By continuously learning and adapting, these algorithms can improve the overall efficiency and effectiveness of EHRs, making them a valuable asset in the healthcare industry. However, ensuring that these systems are properly trained and regularly monitored is essential to maintain their accuracy and effectiveness. Machine learning (ML) is a type of artificial intelligence that enables computers to learn and improve from experience without being explicitly programmed [21]. In recent years, ML has significantly contributed to enhancing the security and prediction of Electronic Health Records (EHRs). Here are some specific ways ML has impacted EHRs:

- Fraud Detection: ML algorithms can analyze vast amounts of data from different sources to identify patterns and anomalies that may indicate fraudulent activities in EHRs. It helps healthcare organizations to protect against insurance fraud, identity theft, and other forms of EHR fraud.
- Predictive Analytics: By processing large amounts of data, including patient history, lab results, and medication data, ML algorithms can predict the likelihood of a patient developing a specific disease. It can help healthcare providers to take preventive measures and provide personalized treatment plans.
- Cyber security: EHRs contain sensitive and confidential patient information, making them a prime target for cybercriminals. ML algorithms can detect threats and protect against cyber-attacks, safeguarding patient privacy and preventing data breaches.
- Real-time Monitoring: ML algorithms can continuously monitor EHRs in real-time and detect any unusual activities or trends, such as an unauthorized person accessing patient records. It helps in the early detection and prevention of potential security breaches.
- Natural Language Processing: ML algorithms can analyze and interpret unstructured data, such as doctor's notes and patient health narratives, to extract relevant information and convert it into structured data for more accessible analysis and prediction.
- Disease Detection and Diagnosis: ML algorithms can analyze medical images, such as X-rays and MRIs, to identify early signs of diseases that may not be visible to the human eye. It can aid in early detection and accurate diagnosis of diseases.
- Personalized Medicine: ML algorithms can help healthcare providers determine the most effective treatment plan for each patient by analyzing individual patient data, including genetic information. It leads to better health outcomes and reduced healthcare costs.

Machine learning has significantly enhanced the security and prediction capabilities of EHRs, making them more accurate, efficient, and secure. As technology advances, we can expect further advancements in the use of ML in healthcare, ultimately leading to improved patient outcomes and healthcare delivery.

2. Related Works

The transformation process also includes mapping the data to a standard data model, enabling data integration from multiple sources. The ETL pipeline will load the transformed data into a suitable storage system. It will allow for easy access and analysis of the data by machine learning algorithms.

Sengan, S., et al.[22] have discussed efficiently retrieving relevant patient data, diagnoses, treatments, and outcomes. The models utilize fuzzy logic and mathematical reasoning to handle uncertainty, imprecision, and ambiguity often present in medical data, providing accurate and personalized health care information. Goodrum, H., et al.[23] have discussed a process that uses computer algorithms to categorize and organize electronic health records based on their content. It can help healthcare providers quickly find and access relevant information, improve data accuracy and

efficiency, and aid decision-making.. Wang, Z. Q., e al. [24] have discussed a system that uses a combination of digital twin and deep learning models to identify potential threats from within an organization. It analyses behavior patterns and uses self-attention techniques to accurately detect anomalies and predict insider attacks. This framework can help organizations prevent security breaches and protect their sensitive data. Cremonesi F. et al. [25] have discussed the essentials for creating a comprehensive and integrated healthcare platform. It allows for integrating diverse data types from various sources, enabling a more accurate and holistic understanding of patients' health. It is crucial to effectively use federated learning techniques in healthcare, ensuring better patient outcomes.

Brito, C. V. et al. [26] have discussed a technique that helps maintain the privacy of sensitive data while performing machine learning tasks on the distributed computing framework Apache Spark. It involves encrypting the data and only allowing certain authorized users to access it, ensuring the confidentiality of the data. Misra, D., et al.[27] have discussed a study aiming to develop an accurate and understandable predictive model for identifying patients at risk of septic shock. It utilizes the power of artificial intelligence to analyze clinical data and provide timely warning for healthcare providers. Ozonze O. et al. [28] have discussed that Automating electronic health record data quality assessment involves using software tools and algorithms to check for errors, inconsistencies, and missing data in EHRs. It can help improve data accuracy, completeness, and reliability, leading to better healthcare decisions and patient outcomes. Ramahlosi, M. N. et al. [29] have discussed a decentralized approach that utilizes cryptographic techniques to ensure the integrity and security of data as it moves through various systems and networks. Recording and verifying transactions on a distributed ledger provides a transparent and immutable system for securing sensitive data. Javaid, M., et al.[30] have discussed that machine learning has become a crucial tool in healthcare, allowing for faster and more accurate diagnosis and treatment of diseases. It can analyze large amounts of data to identify patterns and make predictions, ultimately improving patient outcomes and saving lives. Additionally, it can assist in drug development and clinical trials and automate administrative tasks, freeing up time for healthcare professionals to focus on patient care.

Gupta U. et al. [31] discussed an open-source technology for storing, processing, and analyzing large datasets. It incorporates artificial intelligence capabilities, such as machine learning and natural language processing, to efficiently handle complex and unstructured data. It enables organizations to gain valuable insights and make data-driven decisions. Keloth, V. K. et al. [32] have discussed that Representing and utilizing clinical textual data for real-world studies involves extracting relevant information from medical records and other textual sources to analyze patterns, trends, and outcomes in a real-world setting. This method allows for comprehensive and accurate analysis of clinical data to improve patient care and inform healthcare policies. Manickam V. et al. [33] have discussed a software system designed for healthcare management. It uses advanced data extraction, transformation, and loading techniques to collect and integrate data from different sources, creating a unified database. The framework is adaptable to changing healthcare needs and can make intelligent decisions for efficient data management. Ehwerhemuepha, L., et al.[34] have discussed how Cloud computing offers a scalable and cost-effective solution for hosting, managing, and analyzing large amounts of data in the healthcare industry. It allows for advanced analytics and predictive modeling techniques, improving medical diagnosis, treatment planning, and patient outcomes. One application could be using machine learning algorithms to identify patterns in medical data for early disease detection. Sarkar, S., et al.[35] have discussed the Machine learning-based approach using class-imbalanced proactive and reactive data, which involves using advanced algorithms and techniques to train models on imbalanced data sets, where one class of data is significantly more prevalent than the other. This approach aims to address the challenge of accurately predicting rare events and improving the model's overall performance.

Pirmani A. et al. [36] have discussed a framework designed to analyze large amounts of data for Multiple Sclerosis research. It involves three layers of analysis: local, federated, and central, allowing for scalability and collaboration across institutions while protecting the privacy of patient data. This approach can significantly enhance research efforts and improve our understanding of Multiple Sclerosis. AlZubi, A. A., et al.[37] have discussed that Cyber-attack detection in healthcare is the process of identifying and mitigating malicious activities in a healthcare system using a combination of cyber-physical systems and machine learning techniques. This approach utilizes real-time data from sensors, medical devices, and network traffic to detect anomalies and potential cyber threats, ensuring the safety and security of sensitive healthcare data. Barron-Lugo et al. [38] have discussed a system that can be easily deployed on any cloud service without being tied to one specific provider. It allows flexibility and ease of use in building high-availability data science services, which utilize data-driven design principles for optimal performance and reliability. The comprehensive analysis of the related works has shown in the following table.1 and identified

performance issues have shown in the following table.2. Fig.1 expresses the graphical representation of identified issues.

Table.1: Comprehensive analysis

Authors	Year	Advantage	Limitation
Sengan, S., et al.[22]	2020	Improved accuracy in diagnosis and treatment recommendations due to the use of advanced machine learning techniques	Limited accuracy due to potential errors or gaps in input data for the fuzzy based machine learning model.
Goodrum, H., et al.[23]	2020	Efficient analysis of large volumes of data, leading to faster and more accurate patient diagnosis and treatment	Possibility of misclassification due to diverse formatting and structure of electronic health record documents
Wang, Z. Q., et al.[24]	2023	The digital twin allows for real-time monitoring and analysis of user behavior, reducing the risk of insider threats.	Lack of real-world testing may lead to inaccurate results and limitations in generalizability.
Cremonesi, F., et al.[25]	2023	The practical approach allows for the integration of multiple data types to enhance the accuracy and effectiveness of healthcare data analysis.	Possible lack of scalability due to complex data gathering and management processes.
Brito, C. V., et al.[26]	2023	Protects sensitive data while allowing for large-scale data processing and machine learning on Apache Spark platform.	Restrictions on datasets and computational performance due to privacy-preserving techniques and distributed computing
Misra, D., et al.[27]	2021	It can provide timely and accurate diagnosis, leading to prompt treatment and improved patient outcomes.	Dependence on accurate and timely input data for accurate prediction.
Ozonze, O., et al.[28]	2023	Improved accuracy and reliability of patient information for better decision-making and healthcare outcomes	Subjectivity of data interpretation can lead to inconsistent results and difficulty in standardizing quality assessments.
Ramahlosi, M. N., et al.[29]	2023	Enhanced security and immutability of data through distributed and encrypted ledger technology	High energy consumption due to complex encryption processes on a large scale.
Javaid, M., et al.[30]	2022	It can analyze large amounts of medical data quickly and accurately, aiding in early diagnosis and treatment planning.	The need for large and diverse datasets to properly train and validate machine learning algorithms.
Gupta, U., et al.[31]	2023	Efficiently processes large and complex datasets to provide valuable insights for decision making through advanced machine learning algorithms.	Limited support for real-time processing and lack of integration with other AI frameworks.
Keloth, V. K., et al.[32]	2023	Accurate and comprehensive representation of patient information allows for better understanding and	Missing data or lack of standardization can limit

		analysis of real-world treatment outcomes.	generalizability and reliability of findings in clinical studies.
Manickam, V., et al.[33]	2023	Efficient data integration and analysis for improved decision-making and patient care in the healthcare industry.	It may be difficult to understand and implement in real-world healthcare management scenarios.
Ehwerhemuepha, L., et al.[34]	2020	A cloud computing solution allows for easy scaling of resources to handle large and complex datasets in healthcare analytics	The cost of implementing and maintaining the solution can be prohibitively high for smaller healthcare organizations.
Sarkar, S., et al.[35]	2020	Improved prediction accuracy due to the use of machine learning algorithms and balanced data sets.	Limited by the availability of high-quality, diverse and balanced data for both proactive and reactive cases.
Pirmani, A., et al.[36]	2023	It enables efficient data processing and analysis across multiple locations, increasing scalability and reducing data transfer costs.	Limited to specific diseases, may not be applicable to other research fields.
AlZubi, A. A., et al.[37]	2021	Improved security measures against cyber threats in the healthcare sector through advanced technology and data analysis.	Limited access to real-time data may affect the accuracy and effectiveness of cyber-attack detection in healthcare.
Barron-Lugo, et al.[38]	2023	It have the ability to easily switch between cloud providers for cost optimization or in case of service outage.	Limited flexibility to adapt to specific cloud providers' capabilities and features.

Table.2: Identified Performance Issues

Authors	**Performance Issues**					
	Privacy	**Integrity**	**Authentication**	**Data Transfer**	**Security**	**Data Breaches**
Sengan, S., et al.[22]	High	Moderate	Critical	Low	Critical	Very Low
Goodrum, H., et al.[23]	Moderate	High	Low	Very Low	Low	Very Low
Wang, Z. Q., et al.[24]	Very Low	Critical	High	Very Low	Low	Low
Cremonesi, F., et al.[25]	Very Low	Low	Critical	High	Very Low	Very Low
Brito, C. V., et al.[26]	High	Critical	Low	Very Low	High	Low
Misra, D., et al.[27]	Low	Moderate	Very Low	Critical	Very Low	High
Ozonze, O., et al.[28]	Very Low	Low	High	Low	Low	Critical
Ramahlosi, M. N., et al.[29]	Low	Very Low	Moderate	Very Low	High	Moderate

Javaid, M., et al.[30]	Very Low	Low	Low	Very Low	Critical	Very Low
Gupta, U., et al.[31]	Low	Very Low	Very Low	Very Low	Low	Moderate
Keloth, V. K., et al.[32]	Very Low	Low	Very Low	Critical	Moderate	High
Manickam, V., et al.[33]	Low	Very Low	Critical	Low	High	Very Low
Ehwerhemuepha, L., et al.[34]	Very Low	Critical	Very Low	High	Very Low	Low
Sarkar, S., et al.[35]	Moderate	Critical	High	Moderate	Low	Very Low
Pirmani, A., et al.[36]	Very Low	High	Moderate	Critical	Very Low	Low
AlZubi, A. A., et al.[37]	High	Moderate	Critical	Very Low	Low	Very Low
Barron-Lugo, et al.[38]	Low	Critical	High	Moderate	Very Low	Moderate

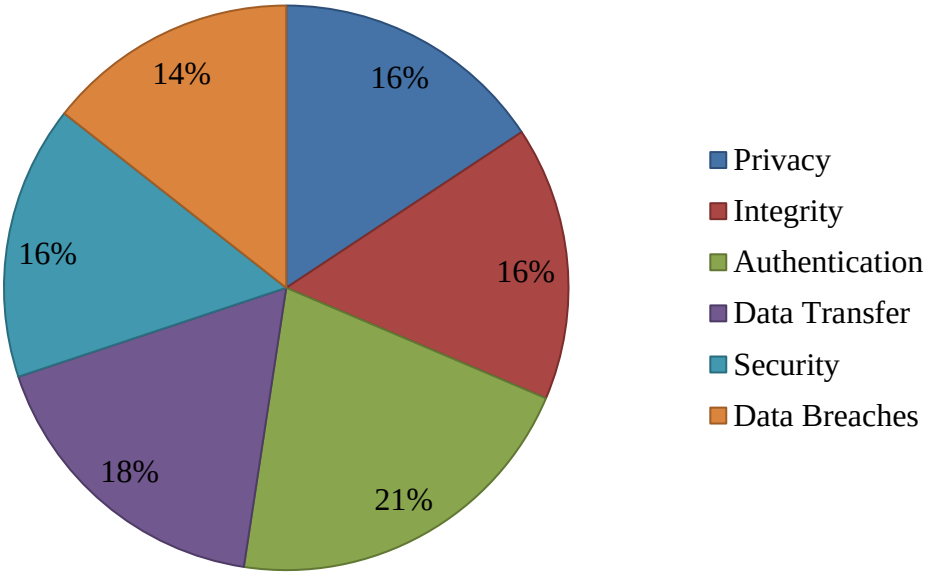

Fig 1: Identified Performance Issues

2.1. Research Gap Analysis

- Data Privacy: The sensitive nature of health information requires strict privacy measures in EHR systems to prevent unauthorized access and use of patient data. Without proper security protocols, patient data is vulnerable to cyber attacks and can lead to identity theft, fraud, and other data breaches.
- Data Integrity: EHR systems must ensure that the data entered into the system is accurate, complete, and reliable. It requires constant monitoring and maintenance to prevent data tampering, deletion, or modification by unauthorized users.
- User Authentication and Authorization: EHR systems must have robust user authentication procedures to ensure that only authorized personnel can access patient data. User access levels should also be appropriately defined to restrict access to sensitive information.
- Data Transfer and Exchange: EHR systems must have secure data transfer and exchange protocols, especially when sharing patient information with other healthcare providers. Secure these channels to avoid unauthorized access or interception of patient data.

- System and Network Security: EHR systems are vulnerable to hacking and malware attacks, which can compromise the entire system and patient data. Ensuring proper system and network security measures, such as firewalls, anti-virus software, and regular updates, is critical in preventing these attacks.
- Data Breaches and Cyber security Threats: EHR systems are increasingly targeted by cybercriminals due to the sensitive nature of patient data and the potential for financial gain. A lack of proper security measures and protocols can result in data breaches and expose patient information to theft or misuse.

An ETL (Extract, Transform, Load) pipeline for EHRs refers to a set of processes that extract data from EHR systems, transform it into a format suitable for machine learning methods, and load it into a centralized data repository. This type of ETL pipeline is novel because it provides a comprehensive and structured approach to data management to support machine learning-based approaches for security and prediction in healthcare.

- An ETL pipeline uses machine learning methods in the data transformation stage. Traditional ETL pipelines often rely on manual or rule-based transformations, which can be time-consuming and error-prone. By implementing machine learning algorithms, the ETL pipeline can automatically identify patterns and relationships in the data, reducing the need for manual interventions.
- An ETL is specifically designed to support machine learning-based approaches for security and prediction. It means that the pipeline is tailored to extract and transform data elements that are relevant for these purposes, such as patient demographics, clinical data, and medical codes. It ensures that the data used for machine learning models is accurate, relevant, and meaningful for the specific use case.
- An ETL pipeline is its ability to handle large and complex datasets. EHR systems typically contain a vast amount of data, including structured and unstructured data from multiple sources. An ETL pipeline can efficiently extract and transform this data, making it easier to manage and analyze. It is essential for machine learning-based approaches, which require extensive training datasets to perform accurate predictions.
- An ETL pipeline incorporates data security measures in the data transformation process. It includes techniques such as data masking, anonymization, and encryption, which help protect sensitive patient information and ensure compliance with privacy regulations such as HIPAA.

An ETL pipeline for EHRs represents a novel approach to managing healthcare data to support machine learning-based approaches for security and prediction. By leveraging machine learning techniques, tailoring the pipeline for the use case, and incorporating data security measures, this type of ETL pipeline offers a robust, efficient, and innovative healthcare data management solution. The remaining part of the paper is organized as follows. Section 3 shows the proposed model and technical information about it. Section 4 shows the comparative analysis of the existing and proposed model. Section 5 expresses the conclusion and future scope of the proposed research.

3. Proposed Model

An ETL (extract, transform, and load) pipeline is a data integration process used to extract data from various sources, transform it into a format suitable for analysis, and load it into a target data warehouse or database. In the EHR context, an ETL pipeline transforms large volumes of complex and diverse healthcare data into a structured, usable format for analysis and reporting.

3.1. Methodology

Machine learning-based approaches can play a crucial role in enhancing the security and prediction of the ETL pipeline for electronic health records. ETL pipelines are the backbone of any data integration process, and they play a vital role in ensuring the accuracy and reliability of data for decision-making in the healthcare industry. However, with the increasing use of electronic health records, the risk of data breaches has also increased. Machine learning algorithms can be utilized to identify any unusual or suspicious patterns in the data and prevent potential security threats. These algorithms can also aid in identifying any discrepancies or errors in the data during the transformation and loading process, thereby ensuring data accuracy. Fig.2 shows the proposed block diagram

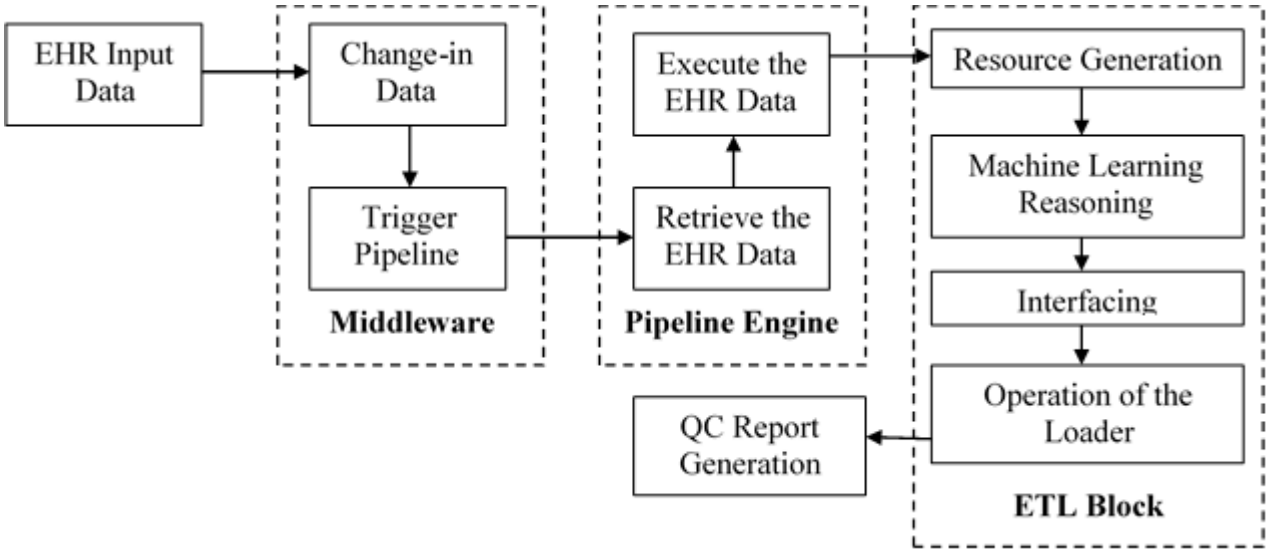

Fig.2: Proposed Block diagram

Machine learning can also improve the prediction capabilities of the ETL pipeline for electronic health record data. These algorithms can learn patterns and predict future events or trends by analyzing historical data. It can be beneficial for predicting potential health risks for individuals or identifying trends in diseases and healthcare utilization. Predictive models can also help healthcare organizations allocate resources more efficiently and plan for potential healthcare needs in advance. The ETL pipeline for EHR typically involves the following operations:

3.2. Extraction

The extraction pipeline process for EHR is the systematic method used to collect, standardize, and organize data from various sources to create a digital record of a patient's medical history. This process is essential in ensuring access to accurate and relevant patient information for healthcare providers. Let's initiate the extraction of documents. The number of initial document length has captured in eq.1

$$p''(o) = \lim_{o \to 0} \left(\frac{p(p+o) - p(o)}{o} \right) \quad (1)$$

The first step in the extraction pipeline process is data collection. It involves gathering data from different sources such as hospitals, clinics, laboratories, or pharmacies. The collection of information extraction has shown in eq.2

$$p'(o) = \lim_{p \to 0} \left(\frac{p^{p+o} - p^p}{o} \right) \quad (2)$$

The data can include patient demographics, medical history, lab results, medications, and more. This data can be in various formats, such as handwritten notes, images, or digital records. Next, the collected data is processed to standardize the information. The extracted data can characterized as per the eq.3

$$p(o) = \lim_{p \to 0} \left(\frac{(p^p * p^o) - p^p}{o} \right) \quad (3)$$

This step involves converting the data into a standardized format that is easily stored, searched and shared. For example, handwritten notes can be scanned and converted into digital text, or lab results can be transformed into a standardized code. The digital texts code can be demonstrated in eq.4

$$f_e^2 = 2 * f * F_e \quad (4)$$

The third step is data organization. In this stage, the standardized data is categorized based on the type of information, such as medications, lab results, or procedures. It helps create a structured record that is easily accessible and understandable for healthcare providers. Once the data is collected, standardized, and organized, it is loaded into a central database to create the patient's EHR. The organization of data can shown in the following eq.5

$$p(o) = e^p * \lim_{p \to 0} \left(\frac{1 - e^p}{o} \right) \quad (5)$$

The database can be a secure system accessible to authorized healthcare providers. The final step is data maintenance and updates. EHRs require regular maintenance to ensure the data is accurate and up-to-date. It can involve periodically reviewing and correcting any errors, as well as documenting any new information. Overall, the extraction pipeline process for EHRs is a crucial component of healthcare operations. It enables healthcare providers to access comprehensive patient information quickly and efficiently, leading to better

patient care, improved health outcomes, and streamlined healthcare processes.

3.3. Transformation

The transformation pipeline process is crucial to storing, accessing, and sharing EHRs. This process involves converting raw data from various sources into a standardized and structured format that healthcare professionals can quickly process and utilize. The first step in the transformation pipeline process is data ingestion. Let's consider the level-1 transformation has the following eq.6

$$\partial o = \partial e^{p} - 1 \quad (6)$$

$$p^{''}(o) = e^{p} * \lim_{o \to 0} \frac{o}{ln(o+1)} \quad (7)$$

It involves collecting and aggregating data from different sources, such as hospital systems, laboratories, pharmacies, and patient portals. Then the level-2 transformation has the following eq.8

$$\partial p^{o} = \partial o + 1 \quad (8)$$

$$p^{''}(o) = e^{p} * \lim_{o \to 0} \frac{1}{\frac{1}{o} ln(o+1)} \quad (9)$$

This data can include patient demographics, medical history, lab results, and medication records. Next, the data goes through a normalization process, which is standardized and cleansed to remove any errors or inconsistencies. It is essential because EHRs may contain data in different formats, and normalization helps ensure that all the information is organized consistently and unified. Now the level-3 transformation has the following eq.10

$$\partial p = \ln(o+1) \quad (10)$$

$$p^{''}(o) = e^{o} * \lim_{o \to 0} \frac{1}{ln(o+1)^{\frac{1}{o}}} \quad (11)$$

After normalization, the data moves to the mapping stage and is mapped to the appropriate fields in the EHR system. It involves identifying and matching data with the correct codes and terminologies, such as ICD (International Classification of Diseases) and SNOMED (Systematized Nomenclature of Medicine), to ensure compatibility and interoperability with other healthcare systems has shows in eq.12

$$f_e = \sqrt{2 * f * E_e} \quad (12)$$

The final step in the transformation pipeline process is data transformation, where the EHR data is converted into a format that can be easily stored, retrieved, and analyzed. It can include converting data into a structured format and organizing it into a logical and hierarchical structure for future use has shows in eq.13

$$p^{''}(o) = e^{o} * \frac{1}{ln * \lim_{o \to 0} (o+1)^{\frac{1}{o}}} \quad (13)$$

$$p^{''}(o) = e^{o} * \frac{1}{ln\, o} \quad (14)$$

The transformation pipeline process is essential in ensuring that EHRs are accurate, consistent, and interoperable, ultimately improving healthcare quality. It also allows for efficient data sharing and analysis, enabling healthcare professionals to make informed decisions and provide better patient care.

3.4. Data Mapping

Data Mapping is transforming and organizing data from one format to another. In the context of EHRs, data mapping is a crucial step in storing and retrieving patient information. The Data Mapping pipeline process in EHRs involves multiple operations that ensure patient data's accurate and efficient management. The first step in the Data Mapping pipeline process is to identify and collect the relevant data from various sources, such as medical forms, laboratory results, and patient interviews. The level-1 mapping cycle has the following eq.15

$$\partial P_1 = -\partial O + \sum_{o=1} \partial P_o = 0 \Rightarrow \frac{\partial P_o}{\partial O_p} = 1 \quad (15)$$

This data is then standardized and normalized to ensure consistency in the format and structure has shows in eq.16

$$f^{''} = \lim_{e \to 0} \left(\frac{g^{e} * (g^{f} - 1)}{f} \right) \quad (16)$$

$$f^{''} = \lim_{e \to 0} \left(\frac{g(e+f) - g(e)}{f} \right) \quad (17)$$

This step ensures that all systems and applications can access and interpret the data correctly. After establishing the relationships, the data is transformed into a common standard format. Then the level-2 mapping has the following eq.18

$$\partial P_2 = O + \sum_{o=1} \omega_o * P_o = 0 \Rightarrow \frac{\partial P_2}{\partial O_o} = 1 \quad (18)$$

The next operation involves defining the relationships between different data elements, for example, mapping a patient's name to their unique identification number or a

specific diagnosis to its corresponding medical code has shows in eq.19

$$\frac{\partial ln(\omega_p)}{\partial O_p} + \psi_1 \frac{\partial P_1}{\partial O_p} + \psi_2 \frac{\partial P_2}{\partial O_p} = 0 \quad (19)$$

This standardization enables seamless data exchange between EHR systems, improving interoperability and data sharing. The data mapping process also involves data cleansing and validation has shows in eq.20

$$ln(P_o) - ln(O_p) + \psi_1 - \psi_2 O_p = 0 \quad (20)$$

To ensure accuracy, it includes identifying and resolving any missing, duplicate, or erroneous data. This step is crucial in maintaining the integrity of the patient's health information and preventing potential errors in diagnosis and treatment. The mapped data is stored in a database or warehouse for easy access and retrieval. It enables healthcare providers to retrieve relevant patient information quickly and accurately, improving the quality and efficiency of patient care. The data mapping pipeline process is critical in managing patient data in EHRs. Standardizing, organizing, and validating data ensures the accuracy and accessibility of patient information, facilitating effective healthcare delivery.

3.5. Load

The load pipeline process for EHR is a method of organizing and loading large amounts of data into an EHR system. This process involves a series of steps designed to ensure the accurate and efficient loading of data while maintaining the integrity and security of patient information. The first step in the process is data preparation. It involves converting the data from its original format, such as paper charts or digital files, into a format compatible with the EHR system. It can include data cleansing to remove any duplicate or irrelevant information. Once the data is prepared, it is loaded into the EHR system. It can be done manually or through automated processes like bulk uploads. The load management has the following eq.21

$$N(p|o) = \left(\frac{N(p,o)}{N(o)}\right) \quad (21)$$

During this step, the system will check the data for errors or inconsistencies and flag them for review. The next step is data mapping, where the system matches the data fields from the source to the corresponding fields in the EHR system has shows in eq.22

$$N(p|o) = \frac{1}{N(o)} * \frac{1}{N''} exp\{p^o o + o^p p + O\} \quad (22)$$

Ensuring that the data is accurately organized and searchable within the EHR is crucial. After data mapping, the system will validate the loaded data against predefined criteria to ensure its accuracy and completeness has shows in eq.23

$$N(p|o) = \frac{1}{N'} exp\{p^o o + oP\} \quad (23)$$

It is an important quality control step to ensure the data is reliable for patient care. Once the data has been successfully loaded and validated, it is indexed for easy retrieval has shows in eq.24

$$N(p|o) = \frac{1}{N''} exp\left\{\sum_{p=1}^{o_e} p_o * o_p + \sum_{p=1}^{o_f} o_p O_p\right\} \quad (24)$$

It allows for quick access to specific patient records and information. In addition to the technical aspects, the load pipeline process includes security measures to protect patient data has shows in eq.25

$$N(p|o) = \frac{1}{N''} \prod_{p=1}^{o_f} exp\{p_o * o_p + o * p_o N_o\} \quad (25)$$

It can include encryption, data backup, and access controls to ensure only authorized personnel can access the information. The load pipeline process is crucial in successfully implementing and using EHR systems. It ensures that patient data is organized, accurate, and secure, allowing healthcare providers to provide quality patient care.

3.6. Data Quality Checks

The Data Quality Checks pipeline process for EHR is a crucial and continuous workflow that ensures the accuracy, completeness, and consistency of data gathered and stored in the EHR system. This pipeline process is designed to identify and address any potential errors or discrepancies in the data to maintain data integrity and trustworthiness. The first step in this process is data collection, where patient information such as demographics, medical history, and clinical data is collected and entered into the EHR system has shows in eq.26

$$N(e_p = 1|o) = \frac{N(e_p=1|o)}{N(e_p=0|o)+N(e_p=1|o)} \quad (26)$$

This data is then subjected to various quality checks, including completeness, accuracy, and consistency. These checks are essential because missing or incorrect data can lead to inaccurate diagnosis and treatment decisions.

$$N(p_o = 1|o) = \frac{exp\{p_o + p^o O_{p,o}\}}{exp\{o,p\} + exp\{p_o + p^o O_{p,o}\}} \quad (27)$$

The second step involves data profiling, where the system analyzes the data to identify any outliers or patterns that may indicate data errors or anomalies. For instance, the

system can flag duplicate entries or inconsistent values in a patient's vital signs. The third step is data cleansing, where data is cleaned and standardized to ensure consistency across the system. It involves removing or correcting any identified errors, such as misspellings or incorrect data formats. This step is crucial in maintaining data accuracy and consistency, as it eliminates potential errors during data entry has shows in eq.28

$$N(p_o = 1|o) = \psi(p_o + o^p O) \quad (28)$$

Once the data has been cleansed and standardized, it goes through a data validation process. The system compares the data against predefined rules and standards to ensure it meets quality criteria. Any data that does not pass validation is flagged, and the necessary corrections are made. The data goes through a data monitoring process, where data quality is regularly monitored and tracked to identify any potential issues or changes in the data quality. This step ensures that the data remains accurate and reliable as healthcare professionals constantly use it for patient care.

3.7. Data Governance

The Data Governance pipeline is a process that ensures the smooth and effective management and utilization of data in EHR. The process involves several operations that work together to ensure the accuracy, security, and accessibility of data. The first operation in the Data Governance pipeline is data collection. It involves capturing patient information such as name, age, medical history, and test results has shows in eq.29

$$N = \lim_{p \to 0} \left(\frac{o(p+o) - o(p)}{o} \right) \quad (29)$$

The data is collected from various sources, including hospitals, clinics, laboratories, and pharmacies. This process ensures that all relevant data is collected accurately from different sources has shows in eq.30

$$N = \lim_{p \to 0} \left(\frac{\left(\frac{1}{o+p-1}\right) - \left(\frac{1}{o-1}\right)}{p} \right) \quad (30)$$

The next operation in the pipeline is data integration. Here, the collected data is combined and stored in a centralized database has shows in eq.31

$$N = \lim_{p \to 0} \left(\frac{\left(\frac{1}{o+p-1}\right)\left(\frac{o-1}{o-1}\right) - \left(\frac{1}{o-1}\right)\left(\frac{o+p-1}{o+p-1}\right)}{o} \right) \quad (31)$$

It allows for better access and management of the data. In this step, data from different sources is standardized to ensure consistency and compatibility across the system. The third operation is data validation has shows in eq.32

$$N = \lim_{p \to 0} \left(\frac{(o-1) - (o+p-1)}{p(o-1)(p+o-1)} \right) \quad (32)$$

This step involves checking the accuracy and completeness of the data. Any inconsistencies or errors are identified and corrected at this stage. Data validation is crucial to ensure the reliability and integrity of the data has shows in eq.33

$$N = \lim_{p \to 0} \left(\frac{(-p)}{p(o-1)^*(o+p-1)} \right) \quad (33)$$

The fourth operation is data governance. It involves establishing policies and procedures for managing and protecting the data. It includes setting access controls, data sharing agreements, and data retention policies has shows in eq.34

$$N = \lim_{p \to 0} \left(\frac{(-1)}{(o-1)^*(o+p-1)} \right) \quad (34)$$

Data governance ensures that the data is used appropriately and by legal and ethical guidelines. The final operation in the pipeline is data analytics. Here, the data is analyzed to gain insights and make informed decisions has shows in eq.35

$$N = \left(\frac{(-1)}{(o-p)^2} \right) \quad (35)$$

It involves using various tools and techniques to identify data patterns, trends, and correlations. Data analytics helps improve patient care, predict health outcomes, and identify areas for process improvement. EHR's Data Governance pipeline process involves efficiently managing data from collection to analysis. It ensures that the data is accurate, secure, accessible, and used in a responsible manner. It ultimately leads to better patient care and improved healthcare outcomes.

3.8. Proposed Framework

The ETL pipeline transforms raw, complex, disparate EHR data into a unified, standardized, usable form for healthcare organizations' analysis, reporting, and decision-making. Providing accurate and timely information is vital to support clinical and administrative processes, improve patient care, and drive operational efficiency. The scoring prediction system involves a straightforward approach where the scores of items proposed by neighboring extractions are summed immediately has shows in eq.36

$$\hat{r}_{ui} = \frac{1}{|U_i(H)|} \sum_{d \in U_i(H)} r_{di} \quad (36)$$

The prediction score of initial extraction U to item I, produced by the mean value approach, has denoted as r_{ui}. The second step involves presenting the mean score values of extractions H and D using method 1, followed by the computation of the weighted average has shows in eq.37

$$\hat{r}_{ui} = \frac{\sum_{d \in Ui(h)} sim_{du} \times r_{vi}}{\sum_{v \in Ni(u)} sim_{uv}} \quad (37)$$

$$\hat{r}_{ui} = \hat{r}_u + \frac{\sum_{v \in N_{i(u)}} sim_{uv} \times (r_{vi} - r_v)}{\sum_{v \in N_{i(u)}} sim_{uv}} \quad (38)$$

Assuming that diff_time represents the time difference between when the extraction borrows loads and the current time and is a time attenuation factor that is less than 1, the extraction preference score for the TP3 class can be determined as follows:

$$\frac{3*\varepsilon^{difftimel}+4*\varepsilon^{difftime2}+3*\varepsilon^{difftime3}}{3} \quad (39)$$

Simultaneously, the level of interest in borrowing will vary significantly over time. Therefore, it is essential to consider factors such as time when determining the preference for borrowing. The functional flow diagram has shown the following fig.3

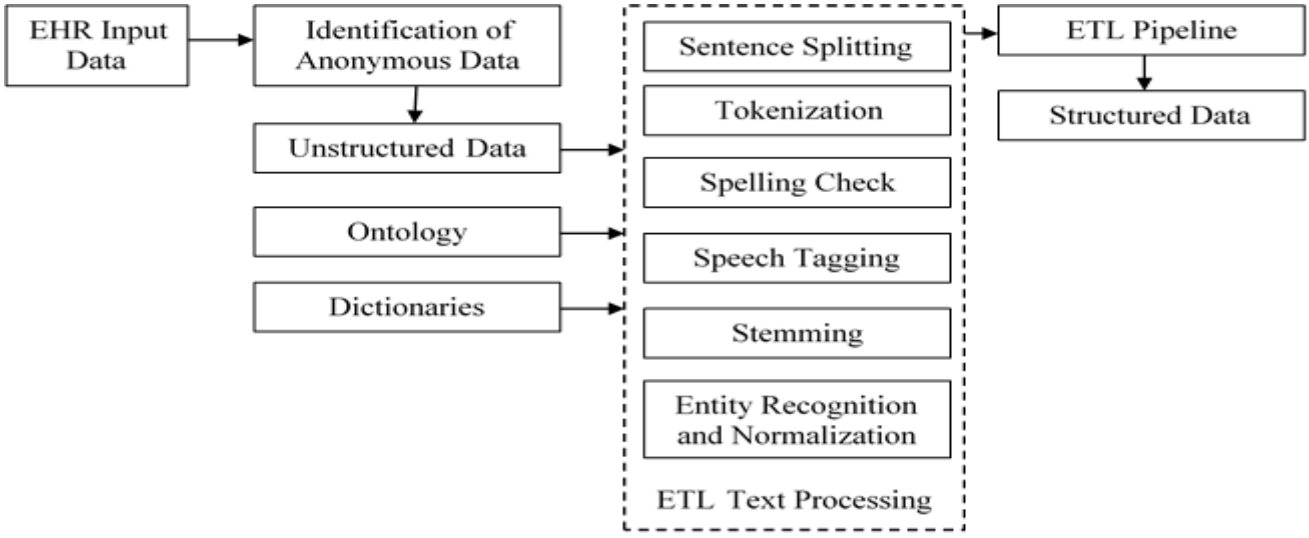

Fig.3: Functional flow diagram

The Data Quality Checks pipeline process for EHR is a comprehensive and continuous process that identifies and corrects any data errors or anomalies to maintain the accuracy, completeness, and consistency of data in the EHR system. It is crucial to ensure that healthcare professionals can rely on the data to make informed decisions for patient care. Within the weight calculation layer, the Wide & Deep model initially converts the problem of calculating recommendation weights into a highly complex multi-classification task using the softmax function. The operational flow diagram has shown in the following fig.4

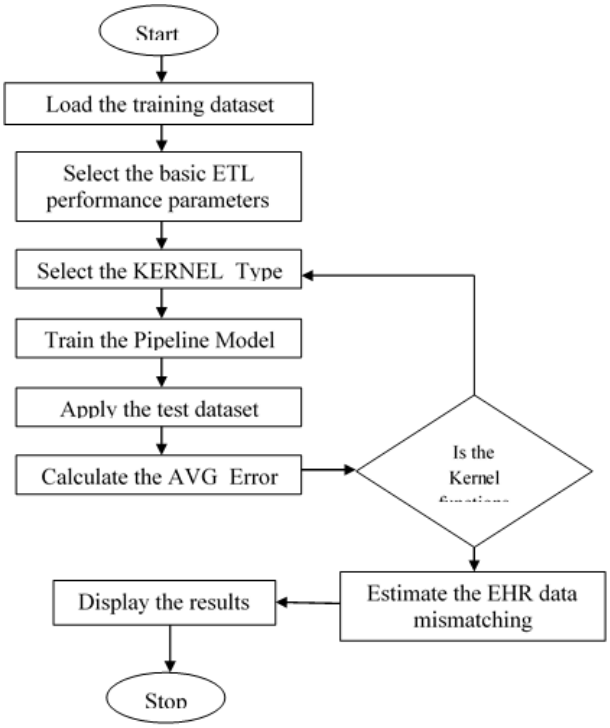

Fig.4: Operational flow diagram

Subsequently, during the online serving stage, it generates the recall results by employing nearest-neighbor retrieval. The expression for Softmax in this scenario is given by equation 40.

$$P\left(s_t = r/H, N\right) = \frac{r^{b,iu}}{\sum_{i \in H} e^{bi,u}} \quad (40)$$

Where, $P\left(s_t = r/H, N\right)$ represents the probability that the item I belongs to a class in the video library D for the user H when t is predicted. The energy function in a given state $\left(b, h\right)$ is defined as

$$E_\phi = -\sum_{j=1}^{r} s_j h_j - \sum_{i=1}^{nm} k_i c_i + \sum_{j=1}^{n} \sum_{i=1}^{k} h_j e_{ji} c_i \quad (41)$$

Applying regularization to the formula yields adjusted results has shows in eq.42

$$P(h, c; \phi) = \frac{1}{g(\phi)} exp\left(-R_\phi(h, c)\right) \quad (42)$$

Where, $g\left(\phi\right)$ By incorporating regularization into the formula, the results are modified.

$$g(\phi) = \sum_{j=1} exp\left(-R_\phi(h, c)\right)_i \quad (43)$$

The activation probability of the implicit element C can be determined when the explicit layer vector H is provided.

$$P\left(b_i = \frac{1}{h}\right) = R\left(\sum_{u=1}^{n} w_{ji} c_j + h_j\right) \quad (44)$$

The second tier of the Wide and deep model is the ranking model, which shares an overall architecture similar to the sorting link. However, the most significant distinction resides in the process of feature engineering. The ultimate goal of this project is to create an ETL pipeline that healthcare organizations can utilize to improve data quality, facilitate cross-platform data integration, and enable the use of machine learning-based approaches for security and prediction. It will help healthcare providers make more informed decisions, improve patient outcomes, and address security challenges in the healthcare sector.

4. Comparative Analysis

The proposed ETL model has compared with the existing fuzzy based machine learning model (FMLM), interactive recurrent neural networks (IRNN), ML-based cyber incident detection (MCID) and deep learning-based identification (DLI). The python simulator has taken here the simulation tool and Synthea dataset [39] has used to implement the results.

4.1. Estimation of Accuracy

EHRs contain sensitive and complex data, making it essential to ensure the accuracy and integrity of this data throughout the entire ETL pipeline. Accuracy in the ETL pipeline for EHRs is crucial for data security. Inaccurate data can lead to incorrect predictions and can compromise patient privacy and security. This is particularly important in the healthcare industry, where data privacy and security regulations must be followed. Any errors in the ETL process could lead to unauthorized access to sensitive patient data, jeopardizing patient confidentiality and trust in the healthcare system.

$$A_{ETL} = \left\{\frac{P_{True} + N_{True}}{P_{True} + N_{True} + P_{False} + N_{False}}\right\} \quad (45)$$

Where, A is accuracy, P is the positive and N is the negative prediction values. Moreover, in order to make accurate predictions using machine learning algorithms, the data must be of high quality and free from errors. Table.3 shows the estimation of accuracy.

Table.3: Estimation of Accuracy (in %)

No. of Documents	FMLM	IRNN	MCID	DLI	ETL
100	65.11	70.14	91.79	86.13	98.38
200	65.22	70.12	91.96	86.40	98.88
300	65.24	69.24	91.23	86.10	98.76
400	62.14	66.41	87.89	82.59	95.53
500	60.94	65.09	87.16	81.27	95.15
600	60.33	64.26	86.27	80.73	94.58
700	59.92	63.86	86.19	80.43	94.88

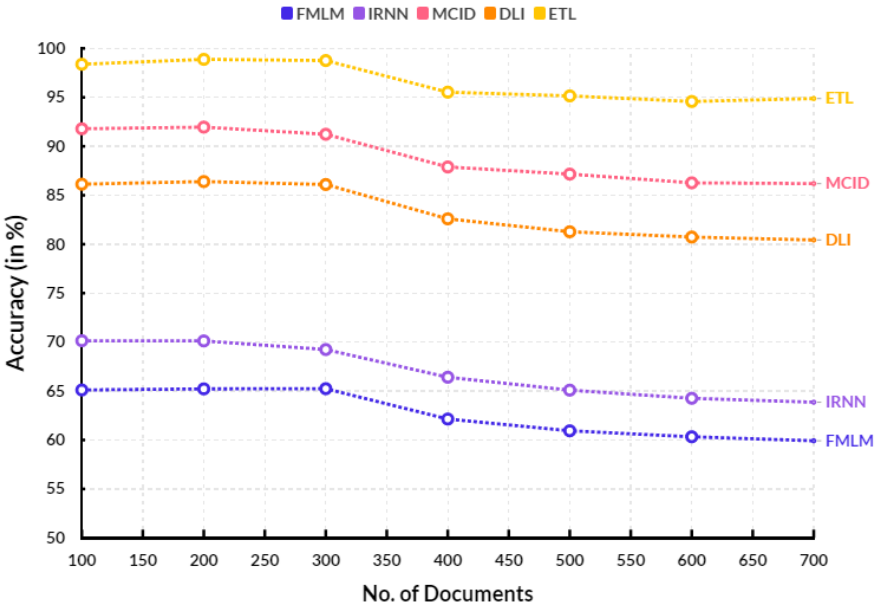

Fig.5: Estimation of Accuracy

Fig.5 shows the estimation of accuracy. In a computation point, the existing FMLM reached 59.92%, IRNN reached 63.86%, MCID reached 86.19% and DLI reached 80.43% accuracy. The proposed ETL reached 94.88% accuracy. The ETL process plays a crucial role in ensuring the accuracy and quality of data. Inaccuracies or missing data during the extraction, transformation, or loading stages can significantly impact the performance of machine learning models. This can potentially result in incorrect or biased predictions, leading to potential harm to patients and loss of trust in the healthcare system.

4.2. Estimation of Precision

EHR data is sensitive and contains valuable information about a patient's medical history, making it essential to ensure the accuracy and integrity of the data throughout the ETL process. The extraction stage must be precise in its selection of data. This involves identifying and retrieving the appropriate data from various sources such as hospital systems, laboratory results, and patient records.

$$P_{ETL} = \left\{\frac{P_{True}}{P_{True}+P_{False}}\right\} \qquad (46)$$

Where, P_{ETL} is precision, P is the positive prediction value. The precision of this step is critical as any missing or incorrect data can lead to biased or flawed predictions. Table.4 shows the estimation of precision.

Table.4: Estimation of precision (in %)

No. of Documents	FMLM	IRNN	MCID	DLI	ETL
100	73.62	66.51	89.28	81.93	98.54
200	72.13	64.54	86.86	79.73	98.55
300	71.33	63.41	86.45	78.93	97.35
400	69.00	62.22	84.85	78.26	96.87
500	67.99	61.83	82.53	76.83	95.44
600	67.35	60.31	81.28	75.74	94.28
700	66.69	60.07	78.55	75.26	93.51

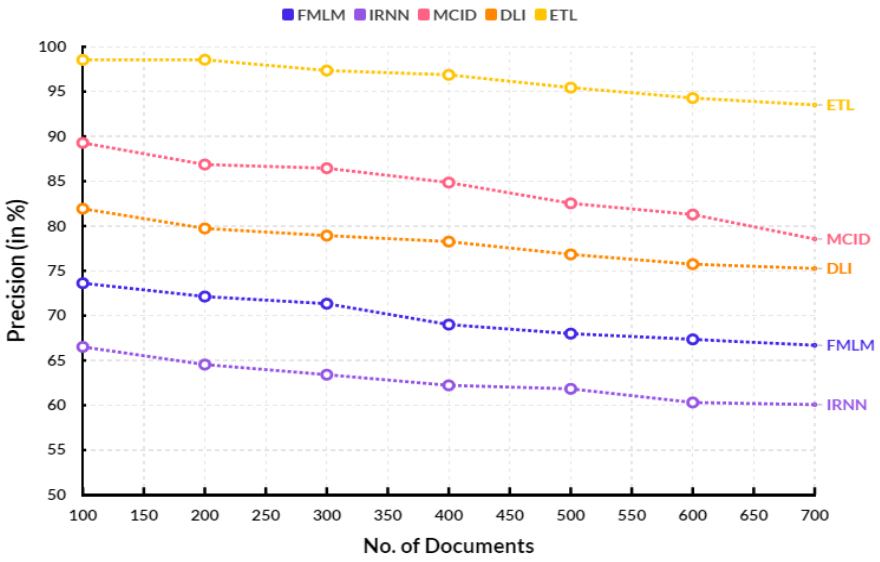

Fig.6: Estimation of precision

Fig.6 shows the estimation of precision. In a computation point, the existing FMLM reached 66.69%, IRNN reached 60.07%, MCID reached 78.55% and DLI reached 75.26% precision. The proposed ETL reached 93.51% precision. The transformation stage involves cleaning, integrating, and standardizing the data. This step is crucial in ensuring data consistency and reliability for ML models. Any errors or discrepancies at this stage can affect the accuracy of predictions and jeopardize the security of patient information. In the load stage, the data is loaded into the target repository, such as a data warehouse or database. The precision of this stage is critical in maintaining data integrity and consistency. Any errors in the loading process can result in data loss or duplication, which can significantly impact the performance of ML algorithms.

4.3. Estimation of Recall

Machine learning-based approaches have been increasingly used in the healthcare industry for security and prediction purposes. However, these approaches require a clean and comprehensive dataset to produce accurate results. Hence, the recall of ETL pipeline becomes crucial in maintaining the integrity and quality of the data. The first step in the ETL pipeline is to extract the relevant data from the EHRs. This includes identifying and retrieving the necessary information from various sources, such as different medical facilities and offices. This extraction process needs to be accurate and comprehensive to ensure all the relevant data is included. Missing or incorrect data can significantly affect the results of any machine learning-based approach. Next, the data needs to be transformed into a format that is suitable for analysis.

$$R_{ETL} = \left\{ \frac{P_{True}}{P_{True} + N_{False}} \right\} \quad (47)$$

Where, R is recall, P is the positive and N is the negative prediction values. This may include standardizing the data, filling in missing values, and removing any redundant or irrelevant data. Table.5 shows the estimation of recall.

Table.5: Estimation of recall (in %)

No. of Documents	FMLM	IRNN	MCID	DLI	ETL
100	63.73	70.61	89.44	82.94	98.54
200	62.10	68.87	87.86	81.52	97.25
300	61.62	66.53	85.66	80.26	96.24
400	60.33	65.72	84.03	78.27	95.35
500	58.22	63.43	82.89	75.80	94.98
600	56.73	61.50	80.69	74.36	93.94
700	54.92	59.77	79.54	72.64	93.17

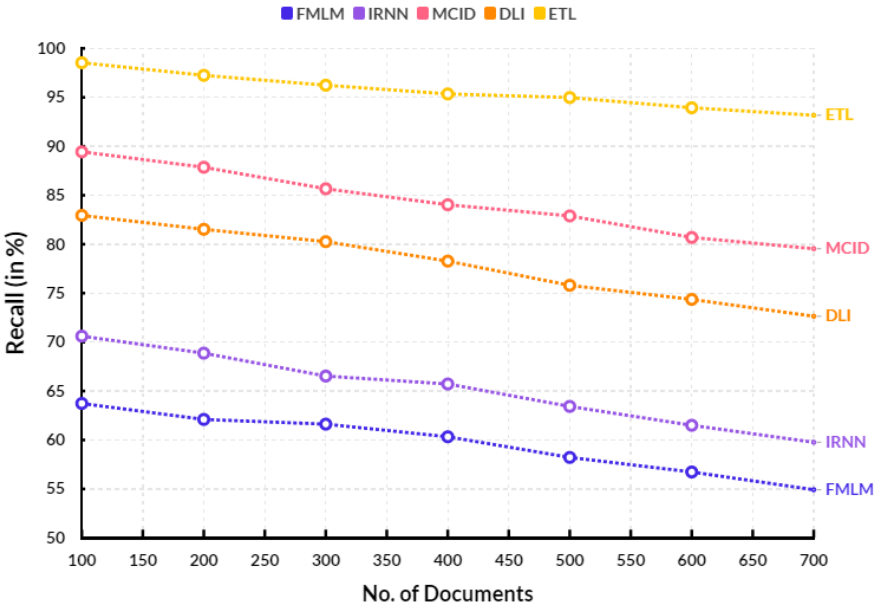

Fig.7: Estimation of recall

Fig.7 shows the estimation of recall. In a computation point, the existing FMLM reached 54.92%, IRNN reached 59.77%, MCID reached 79.54% and DLI reached 72.64% recall. The proposed ETL reached 93.17% recall. The ETL pipeline also involves data cleansing, where any errors or inconsistencies in the data are identified and corrected. This step is crucial in ensuring the accuracy and reliability of the data used for machine learning. The final step in the ETL pipeline is to load the transformed data into the machine learning models for analysis. This step is critical in ensuring the models have access to clean and accurate data for producing reliable results. Any errors or omissions in the ETL pipeline can lead to biased or incorrect predictions, impacting the overall effectiveness of the machine learning-based approach.

4.4. Estimation of F1-Score

The F1-score is a measure of its effectiveness in supporting machine learning based approaches for security and prediction. EHRs contain a wealth of sensitive patient information, making them a prime target for data breaches and cyber attacks. As a result, strong security measures must be in place to protect patient privacy and prevent unauthorized access to the data. An ETL pipeline is a critical component of any machine learning based approach for EHR security and prediction. It is responsible for extracting data from various sources, transforming it into a usable format, and loading it into a target system.

$$F1_{ETL} = \left\{\frac{P_{ETL} * R_{ETL}}{P_{ETL} + R_{ETL}}\right\} \quad (47)$$

Where, F1 is F1-score, P_{ETL} is the precision and R_{ETL} is the recall values. This process is essential for making accurate predictions and ensuring the security of EHR data. Table.6 shows the estimation of F1-Score.

Table.6: Estimation of F1-Score (in %)

No. of Documents	FMLM	IRNN	MCID	DLI	ETL
100	64.99	62.87	81.88	74.50	99.28
200	64.66	61.37	81.29	72.63	98.24
300	63.32	60.26	80.31	71.80	98.11
400	62.18	59.88	79.10	70.89	97.15
500	61.13	58.87	77.96	69.97	97.58
600	60.42	57.94	76.85	68.64	96.34
700	59.12	56.94	76.15	67.77	96.23

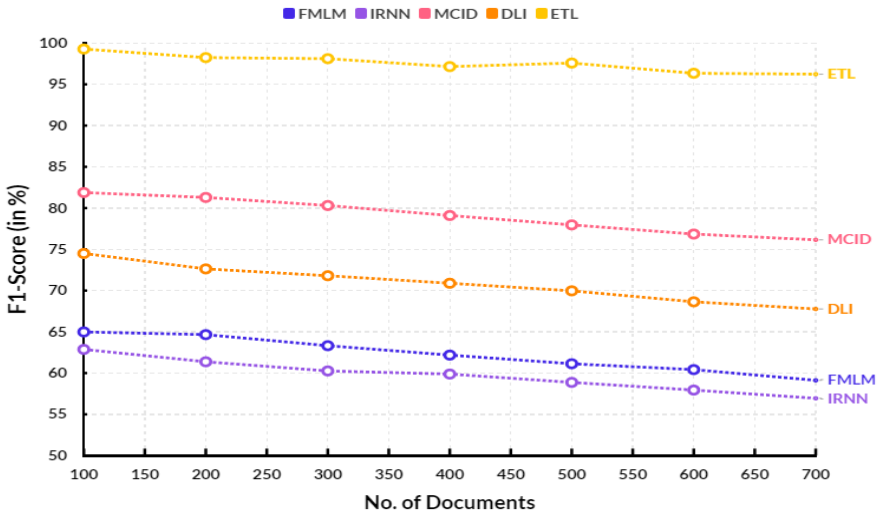

Fig.8: Estimation of F1-Score

Fig.8 shows the estimation of F1-Score. In a computation point, the existing FMLM reached 59.12%, IRNN reached 56.94%, MCID reached 76.15% and DLI reached 67.77% F1-Score. The proposed ETL reached 96.23% F1-Score. The F1-score of an ETL pipeline reflects its ability to effectively handle the diverse and often complex data present in EHRs. This includes structured data such as patient demographics and medical records, as well as unstructured data such as doctor's notes and diagnostic images. A high F1-score indicates that the ETL pipeline is correctly extracting all of the necessary data and transforming it accurately for analysis.

4.5. Estimation of False positive rate

The ETL involves extracting data from various sources, transforming it into a usable format, and loading it into the ML model for analysis and predictions. However, like any other system, the ETL pipeline for EHR is not perfect and can produce false positive results. It refers to the percentage of incorrect or inaccurate information included in the final dataset. This means that the pipeline can mistakenly include data that is not relevant or does not accurately represent the patient's health information. These false positives can occur due to human error during the data extraction and transformation process or due to inconsistencies in the source data. It can lead to incorrect predictions and decisions regarding patient care, and it can also pose security risks. For example, if false positive information is included in the EHR dataset, it can lead to wrong diagnoses or incorrect treatment plans, putting patients' health at risk.

$$FPR_{ETL} = \left\{\frac{P_{False}}{P_{False}+N_{True}}\right\} \tag{49}$$

Where, FPR is false positive rate, P is the positive and N is the negative prediction values. The incorrect or irrelevant data can also compromise the confidentiality of patients' sensitive information, leading to potential security breaches. Table.7 shows the estimation of false positive rate.

Table.7: Estimation of False positive rate (in %)

No. of Documents	FMLM	IRNN	MCID	DLI	ETL
100	62.69	60.57	85.28	77.24	98.37
200	62.36	59.07	84.69	75.37	97.36
300	61.02	57.96	83.71	74.54	97.20
400	59.88	57.58	82.50	73.63	96.24
500	58.83	56.57	81.36	72.71	96.67
600	58.12	55.64	80.25	71.38	95.47
700	56.82	54.64	79.55	70.30	95.31

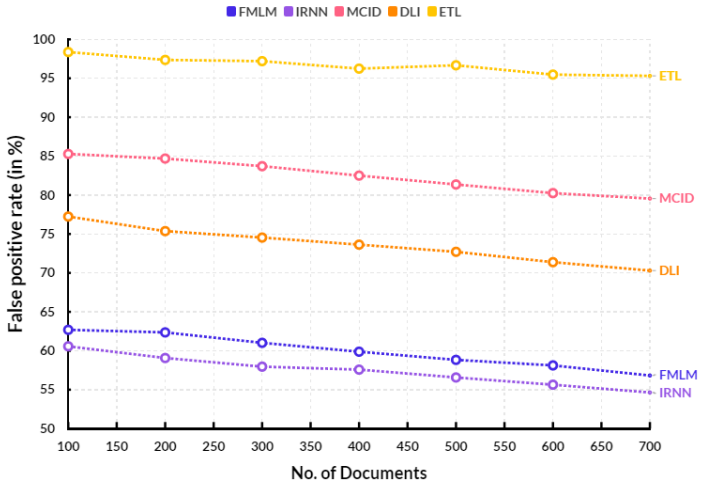

Fig.9: Estimation of False positive rate

Fig.9 shows the estimation of False positive rate. In a computation point, the existing FMLM reached 56.82%, IRNN reached 54.64%, MCID reached 79.55% and DLI reached 70.30% False positive rate. The proposed ETL reached 95.31% False positive rate. It is crucial to have stringent quality control measures in place. These can include regular data audits and checks for data accuracy, as well as having experienced data analysts and medical professionals involved in the process. Furthermore, using advanced technologies, such as natural language processing and data validation techniques, can also help reduce the chances of false positives. It must be carefully monitored and controlled to ensure the accuracy and security of patient information. By maintaining a low false positive rate, healthcare organizations can effectively use ML-based approaches to enhance security and make accurate predictions for patient care.

4.6. Estimation of False Negative rate

The false negative rate refers to the percentage of data that is incorrectly labeled as negative or normal when it should have been labeled as positive or abnormal. This can happen during the extraction, transformation, or loading stages of the pipeline. In the context of using machine learning approaches for security and prediction in EHR, a high false negative rate can have serious consequences. A false negative in the security aspect of EHR could mean that the pipeline did not flag a potential security breach or threat, leading to vulnerable patient data.

$$FNR_{ETL} = \left\{\frac{N_{False}}{N_{False}+P_{True}}\right\} \qquad (50)$$

Where, FNR is false negative rate, P is the positive and N is the negative prediction values. This can put patients at risk and compromise their privacy and trust in the healthcare system. Table.8 shows the estimation of False Negative rate.

Table.8: Estimation of False Negative rate (in %)

No. of Documents	FMLM	IRNN	MCID	DLI	ETL
100	58.56	88.11	88.22	66.79	91.68
200	58.89	89.61	88.81	68.66	92.69
300	60.23	90.72	89.79	69.49	92.85
400	61.37	91.10	91.00	70.40	93.81
500	62.42	92.11	92.14	71.32	93.38
600	63.13	93.04	93.25	72.65	94.58
700	64.43	94.04	93.95	73.73	94.74

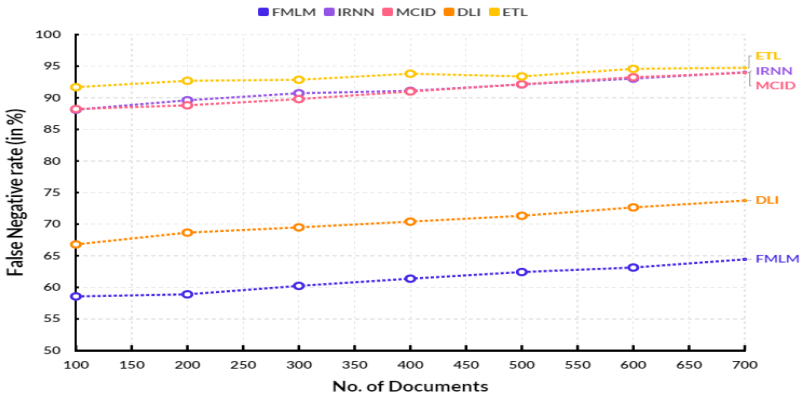

Fig.10: Estimation of False Negative rate

Fig.10 shows the estimation of False Negative rate. In a computation point, the existing FMLM reached 64.43%, IRNN reached 94.04%, MCID reached 93.95% and DLI reached 73.73% False Negative rate. The proposed ETL reached 94.74% False Negative rate. A high false negative rate means that the pipeline failed to identify important patterns or trends in the data, leading to inaccurate projections or predictions. This can result in incorrect diagnoses or treatments, which can have negative impacts on patient outcomes and overall healthcare quality. The false negative rate can be influenced by various factors such as data quality, the complexity of the data, and the algorithms used in the pipeline. It is important to continuously monitor and improve the ETL pipeline to minimize the false negative rate, as it can greatly affect the reliability and effectiveness of machine learning-based approaches for security and prediction in EHR. The convergence of performance has show n in the following table. 9

Parameters	FMLM	IRNN	MCID	DLI	ETL
Accuracy (A)	62.70	67.02	88.93	83.38	96.59
Precision (P)	69.73	62.70	84.26	78.10	96.36
Recall (R)	59.66	65.20	84.30	77.97	95.64
F1-Score (F)	62.26	59.73	79.08	70.89	97.56
False Positive Rate (FPR)	59.96	57.43	82.48	73.60	96.66
False Negative Rate (FNR)	61.29	91.25	91.02	70.43	93.39

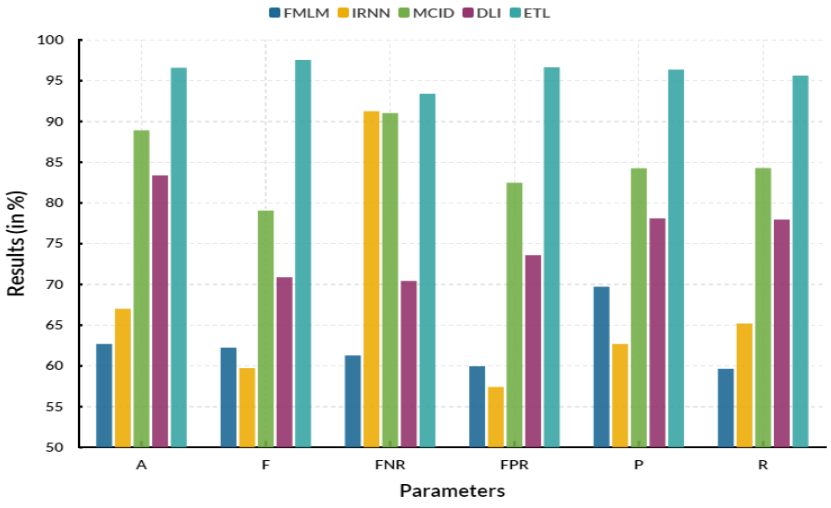

Fig.11: Convergence of performance

Fig.11 shows the convergence of performance. In a comparison point the existing FMLM obtained 62.70% accuracy, 69.73% precision, 59.66% recall, 62.26% f1-score, 59.96% false positive rate, 61.29% false negative rate. IRNN obtained 67.02% accuracy, 62.70% precision, 65.20% recall, 59.73% f1-score, 57.43% false positive rate, 91.25% false negative rate. MCID obtained 88.93% accuracy, 84.26% precision, 84.30% recall, 79.08% f1-score, 82.48% false positive rate, 91.02% false negative rate. DLI obtained 83.38% accuracy, 78.10% precision, 77.97% recall, 70.89% f1-score, 73.60% false positive rate, 70.43% false negative rate. The proposed ETL obtained 96.59% accuracy, 96.36% precision, 95.64% recall, 97.56% f1-score, 96.66% false positive rate, 93.39% false negative rate. EHR data is a critical component of healthcare industries, as it contains patients' health history, diagnoses, treatments, and other relevant information. However, EHR data can be challenging to manage and analyze due to its sheer volume, complexity, and diversity. The proposed ETL pipeline model addresses these challenges by streamlining the process of collecting, transforming, and loading EHR data into a centralized database. The data extraction process in the ETL pipeline involves pulling data from various sources, such as electronic medical records, lab results, and radiology reports. This step eliminates the need for manual data entry, reducing the chances of human error and saving time for healthcare professionals. Moreover, the ETL pipeline can handle large amounts of data from multiple sources simultaneously, ensuring that all relevant information is included in the final dataset. Tthe data transformation phase involves cleaning, formatting, and standardizing the extracted data. This step is crucial for EHR data, as it often comes in different formats and structures, making it challenging to analyze. The ETL pipeline utilizes data cleansing techniques to remove duplicate or incomplete records, improving the overall quality of the data. Additionally, the pipeline can also standardize data using industry-specific codes and terminologies, making it easier to compare and analyze data across different healthcare systems. The load phase of the ETL pipeline involves transferring the cleaned and transformed data into a centralized database. This centralized database serves as a single source of truth for all EHR data, eliminating the need to access multiple systems or software. Having a central repository for EHR data allows for seamless data integration and analysis, leading to better insights and decision-making for healthcare professionals. The proposed ETL pipeline model addresses the challenges of managing EHR data by automating data extraction, cleansing, transformation, and loading processes. This model not only improves data accuracy and consistency, but it also allows for easier data integration and analysis. Overall, these improvements can lead to better outcomes for EHR data, such as more accurate diagnoses, improved patient care, and enhanced research capabilities for healthcare organizations.

5. Conclusion

The ETL pipeline model for Electronic Health Record (EHR) data is an efficient system for managing the extraction, transformation, and loading of data from different sources into a central repository. In this model, the raw data is first extracted from various sources such as hospital systems, medical devices, laboratory equipment, and electronic health record systems. Then, the data is transformed into a standardized format to ensure consistency and accuracy. Finally, the transformed data is loaded into a central repository, which can then be accessed for analysis and reporting. Using this model, healthcare organizations can collect and store large volumes of diverse data in a structured and organized manner. This allows for easier data management and analysis, providing valuable insights into patient health and healthcare practices. By integrating data from various sources, the ETL pipeline model allows for a holistic view of patient health, which can help healthcare providers make more informed decisions about patient care. Moreover, the ETL pipeline model also ensures data quality and integrity by applying data cleansing and validation processes during the transformation stage. The proposed model obtain 96.59% accuracy, 96.36% precision, 95.64% recall, 97.56% f1-score, 96.66% false positive rate, 93.39% false negative rate. This helps in identifying and correcting data errors and inconsistencies, ensuring the accuracy and reliability of the data. The ETL pipeline model for EHR data is a crucial component in the healthcare industry as it enables efficient data management, analysis, and informed decision-making. It helps in improving patient outcomes, reducing costs, and streamlining healthcare operations. The model has become an essential tool for healthcare organizations to effectively leverage the vast amount of data available and make data-driven decisions for the benefit of patients and the healthcare system as a whole.

References

[1] Harerimana, G., Kim, J. W., Yoo, H., & Jang, B. (2019). Deep learning for electronic health records analytics. IEEE Access, 7, 101245-101259.

[2] Latif, J., Xiao, C., Tu, S., Rehman, S. U., Imran, A., & Bilal, A. (2020). Implementation and use of disease diagnosis systems for electronic medical records based on machine learning: A complete review. IEEE Access, 8, 150489-150513.

[3] R. Saklani, K. Purohit, S. Vats, V. Sharma, V. Kukreja and S. P. Yadav, "Multicore Implementation of K-Means Clustering Algorithm," 2023 2nd International Conference on

Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 171-175

[4] Corey, K. M., Kashyap, S., Lorenzi, E., Lagoo-Deenadayalan, S. A., Heller, K., Whalen, K., ... & Sendak, M. (2018). Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. PLoS medicine, 15(11), e1002701.

[5] Annapragada, A. V., Donaruma-Kwoh, M. M., Annapragada, A. V., & Starosolski, Z. A. (2021). A natural language processing and deep learning approach to identify child abuse from pediatric electronic medical records. PLoS One, 16(2), e0247404.

[6] Balch, J. A., Ruppert, M. M., Loftus, T. J., Guan, Z., Ren, Y., Upchurch, G. R., ... & Bihorac, A. (2023). Machine Learning–Enabled Clinical Information Systems Using Fast Healthcare Interoperability Resources Data Standards: Scoping Review. JMIR Medical Informatics, 11, e48297.

[7] Ramesh, G., Logeshwaran, J., & Aravindarajan, V (2022). A Secured Database Monitoring Method to Improve Data Backup and Recovery Operations in Cloud Computing. BOHR International Journal of Computer Science, 2(1), 1-7

[8] López-Martínez, F., Núñez-Valdez, E. R., García-Díaz, V., & Bursac, Z. (2020). A case study for a big data and machine learning platform to improve medical decision support in population health management. Algorithms, 13(4), 102.

[9] Yadav, S. P., & Yadav, S. (2019). Mathematical implementation of fusion of medical images in continuous wavelet domain. Journal of Advanced Research in dynamical and control system, 10(10), 45-54

[10] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics, 19(6), 1236-1246.

[11] Palanisamy, V., & Thirunavukarasu, R. (2019). Implications of big data analytics in developing healthcare frameworks–A review. Journal of King Saud University-Computer and Information Sciences, 31(4), 415-425.

[12] Fleuren, L. M., Dam, T. A., Tonutti, M., de Bruin, D. P., Lalisang, R. C., Gommers, D., ... & Elbers, P. W. (2021). The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients. Critical Care, 25, 1-12.

[13] V. A. Mohammed, M. A. Mohammed, M. A. Mohammed, J. Logeshwaran and N. Jiwani, Machine Learning-based Evaluation of Heart Rate Variability Response in Children with Autism Spectrum Disorder, 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 1022-1028

[14] Miller, D. D. (2020). Machine intelligence in cardiovascular medicine. Cardiology in Review, 28(2), 53-64.

[15] Yadav, S. P., & Yadav, S. (2019). Fusion of Medical Images using a Wavelet Methodology: A Survey. In IEIE Transactions on Smart Processing & Computing (Vol. 8, Issue 4, pp. 265–271). The Institute of Electronics Engineers of Korea

[16] Bates, D. W., Auerbach, A., Schulam, P., Wright, A., & Saria, S. (2020). Reporting and implementing interventions involving machine learning and artificial intelligence. Annals of internal medicine, 172(11_Supplement), S137-S144.

[17] Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial intelligence in medicine, 104, 101822.

[18] Rehman, A., Naz, S., & Razzak, I. (2022). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimedia Systems, 28(4), 1339-1371.

[19] Haendel, M. A., Chute, C. G., Bennett, T. D., Eichmann, D. A., Guinney, J., Kibbe, W. A., ... & Gersing, K. R. (2021). The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. Journal of the American Medical Informatics Association, 28(3), 427-443.

[20] Ben Ali, W., Pesaranghader, A., Avram, R., Overtchouk, P., Perrin, N., Laffite, S., ... & Hussin, J. G. (2021). Implementing machine learning in interventional cardiology: the benefits are worth the trouble. Frontiers in Cardiovascular Medicine, 8, 711401.

[21] Mohammed, M. A., Mohammed, M. A., Mohammed, V. A., Logeshwaran, J., & Jiwani, N. (2023, February). An earlier serial lactate determination analysis of cardiac arrest patients using a medical machine learning model. In 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS) (pp. 263-268). IEEE

[22] Sengan, S., Kamalam, G. K., Vellingiri, J., Gopal, J., Velayutham, P., & Subramaniyaswamy, V. (2020). Medical information retrieval systems for e-Health care records using fuzzy based machine

learning model. Microprocessors and Microsystems, 103344.

[23] Goodrum, H., Roberts, K., & Bernstam, E. V. (2020). Automatic classification of scanned electronic health record documents. International journal of medical informatics, 144, 104302.

[24] Wang, Z. Q., & El Saddik, A. (2023). DTITD: An Intelligent Insider Threat Detection Framework Based on Digital Twin and Self-attention Based Deep Learning Models. IEEE Access.

[25] Cremonesi, F., Planat, V., Kalokyri, V., Kondylakis, H., Sanavia, T., Resinas, V. M. M., ... & Uribe, S. (2023). The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform. Journal of Biomedical Informatics, 141, 104338.

[26] Brito, C. V., Ferreira, P. G., Portela, B. L., Oliveira, R. C., & Paulo, J. T. (2023). Privacy-Preserving Machine Learning on Apache Spark. IEEE Access, 11, 127907-127930.

[27] Misra, D., Avula, V., Wolk, D. M., Farag, H. A., Li, J., Mehta, Y. B., ... & Abedi, V. (2021). Early detection of septic shock onset using interpretable machine learners. Journal of Clinical Medicine, 10(2), 301.

[28] Ozonze, O., Scott, P. J., & Hopgood, A. A. (2023). Automating electronic health record data quality assessment. Journal of Medical Systems, 47(1), 23.

[29] Ramahlosi, M. N., & Akanbi, Y. M. A. (2023). A Blockchain-based Model for Securing Data Pipeline in a Heterogeneous Information System. Published Online by the SAICSIT 2023 Organising Committee Potchefstroom: South African Institute of Computer Scientists & Information Technologists, 167.

[30] Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars and applications. International Journal of Intelligent Networks, 3, 58-73.

[31] Gupta, U., & Sharma, R. (2024). Apache Hadoop framework for big data analytics using AI. In Artificial Intelligence and Blockchain in Industry 4.0 (pp. 130-140). CRC Press.

[32] Keloth, V. K., Banda, J. M., Gurley, M., Heider, P. M., Kennedy, G., Liu, H., ... & Xu, H. (2023). Representing and utilizing clinical textual data for real world studies: An OHDSI approach. Journal of Biomedical Informatics, 142, 104343.

[33] Manickam, V., & Rajasekaran Indra, M. (2023). Dynamic multi-variant relational scheme-based intelligent ETL framework for healthcare management. Soft Computing, 27(1), 605-614.

[34] Ehwerhemuepha, L., Gasperino, G., Bischoff, N., Taraman, S., Chang, A., & Feaster, W. (2020). HealtheDataLab–a cloud computing solution for data science and advanced analytics in healthcare with application to predicting multi-center pediatric readmissions. BMC medical informatics and decision making, 20, 1-12.

[35] Sarkar, S., Pramanik, A., Maiti, J., & Reniers, G. (2020). Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data. Safety science, 125, 104616.

[36] Pirmani, A., De Brouwer, E., Geys, L., Parciak, T., Moreau, Y., & Peeters, L. M. (2023). The Journey of Data Within a Global Data Sharing Initiative: A Federated 3-Layer Data Analysis Pipeline to Scale Up Multiple Sclerosis Research. JMIR Medical Informatics, 11(1), e48030.

[37] AlZubi, A. A., Al-Maitah, M., & Alarifi, A. (2021). Cyber-attack detection in healthcare using cyber-physical system and machine learning techniques. Soft Computing, 25(18), 12319-12332.

[38] Barron-Lugo, J. A., Gonzalez-Compean, J. L., Lopez-Arevalo, I., Carretero, J., & Martinez-Rodriguez, J. L. (2023). Xel: A cloud-agnostic data platform for the design-driven building of high-availability data science services. Future Generation Computer Systems, 145, 87-103.

[39] https://www.kaggle.com/datasets/krsna540/synthea-dataset-jsons-ehr