International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Exploring Privacy-Preserving Disease Diagnosis: A Comparative Analysis

Jaskaranbeer Kaur[1*] and Manu Bansal[2,] Bharat Garg[3]

[1*]Thapar Institute of Engineering and Technology, Patiala Email:-jkaur_phd19@thapar.edu
[2]Thapar Institute of Engineering and Technology, Patiala Email:-mbansal@thapar.edu
[3]Thapar Institute of Engineering and Technology, Patiala Email:-bharat.garg@thapar.edu

***Corresponding Author:** Jaskaranbeer Kaur
[*]Thapar Institute of Engineering and Technology, Patiala Email:-jkaur_phd19@thapar.edu

Abstract:
In the healthcare sector, data is considered as a valuable asset, with enormous amounts generated in the form of patient records and disease-related information. Leveraging machine learning techniques enables the analysis of extensive datasets, unveiling hidden patterns in diseases, facilitating personalized treatments, and forecasting potential health issues. However, the flourish of online diagnosis and prediction still faces some challenges related to information security and privacy as disease diagnosis technologies utilizes a lot of clinical records and sensitive patient data. Hence, it becomes imperative to prioritize the development of innovative methodologies that not only advance the accuracy and efficiency of disease prediction but also ensure the highest standards of privacy protection. This requires collaborative efforts between researchers, healthcare practitioners, and policymakers to establish a comprehensive framework that addresses the evolving landscape of healthcare data while safeguarding individual privacy. Addressing this constraint, numerous researchers integrate privacy preservation measures with disease prediction techniques to develop a system capable of diagnosing diseases without compromising the confidentiality of sensitive information. The survey paper conducts a comparative analysis of privacy-preserving techniques employed in disease diagnosis and prediction. It explores existing methodologies across various domains, assessing their efficacy and trade-offs in maintaining data confidentiality while optimizing diagnostic accuracy. The review highlights the need for robust privacy measures in disease prediction, shortcomings related to existing techniques of privacy preserving disease diagnosis, and provides insights into promising directions for future research in this critical intersection of healthcare and privacy preservation.

Keywords: Privacy Preservation, Disease Diagnosis, Machine Learning, Cryptography

1. Introduction

Recent developments in Information and Communication Technology (ICT), Internet of Things (IoTs), Cloud Services, Wireless Sensors and Smart Devices have tremendously enhanced the human life expectancy. This evolution has also influenced the traditional approach of healthcare practices. Healthcare industry contributes in maintaining healthiest lifestyle and works with an agenda "*to get people healthy*". Modern healthcare incorporates ICT in order to improve healthcare services by addressing the shortcomings of traditional healthcare approaches [1]. This gradual shift towards digital healthcare provides various services such as remote patient monitoring, replace paper based records to Electronic Health Records (EHR), patient' remote access and control to EHR, health knowledge management, disease surveillance and diagnosis, clinical decision support etc. As a result of these diverse medical utilities, modern healthcare industry is emerging at a massive rate. The global medical device market is forecasted to grow at a compound annual growth rate of 4.5% from 2018 to 2023 and is expected to reach $409.5 billion by 2025 [2].

Modern healthcare systems generate voluminous amount of data on daily basis that includes clinical decision support systems data, healthcare based queries and their responses, hospital's administrative data, patients' prescriptions and medical suggestions, clinical data from Computerized Physician Order Entries (CPOE), Electronic Health Records (EHR), Patient's Health Insurance Data and Policies etc. This data is collected by the hospitals and stored on secure network servers to facilitate all-time data accessibility for patient care. Due to software vulnerabilities, human error and weak security policies, the high value healthcare information is sometimes exposed to unauthorized users [3]. Healthcare enterprises have been a frequent target of cyber-attacks from past few years as they contain extremely sensitive and private information about patients. This leads to illegal disclosure of Protected Health Information (PHI) in the form of data breaches. Several attacks, affecting more than a million users each, have occurred within a short span of time. As per the reports from many practitioners, the total number of 249.09 million individuals were affected from healthcare breaches from 2005 to 2019 [3]. Healthcare industry has faced the highest number of privacy breaches among all the industries. In year 2018, 536 healthcare breaches had reported from 65 countries [4]. 505 healthcare breaches have resulted in theft, exposure, or illegal disclosure of 41.2 million healthcare

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

records in year 2019 [5]. One of the biggest healthcare data breach had occurred in January 2015 that resulted in theft of around 78.8 million patient records from Anthem Health Insurance Database [6]. Healthcare data breaches have also affected financial economy of any nation. Medical data and records now are selling for an average of 40-50 USD per record in black market, having more worth than credit card numbers [1]. According to IBM report, the average cost of healthcare industry breach in 2019 was $6.45 million. The average cost of healthcare data breaches increased by 3.5% from 2015 to 2019. Also, the cost of healthcare breached record registered an increase of 5.0% in the same time period [3].

The abovementioned incidents highlighted an imminent need of data privacy in healthcare systems to protect and secure the integrity, confidentiality and availability of extremely private patient related information. A rigorous privacy preservation mechanism is precondition need to be met in order to ensure privacy of patient's identity, medical records, ongoing diagnosis, and treatments etc.

1.1. Privacy Preservation in Healthcare

Privacy is a multidimensional concept defined in philosophical, legal and technical context. Information Privacy can be defined as "the right of people, groups or organizations to decide for themselves when, how, and to what level information regarding to them is transferred to others". Health Information Privacy can be stated as "An individual's right to control the use, acquisition, or disclosure of his or her identifiable health data" [1]. Keeping in mind the need of privacy in modern healthcare, privacy is made fundamental right in legal terms. Privacy policies and various standards for have been legalized in various countries. The "Health Insurance Portability and Accountability Act (HIPAA)" states US health-informatics' privacy rules for Electronic Health Records (EHR). "Health Information Technology for Economic and Clinical Health Act (HITECH Act)" was created to derive the adoption and meaningful use of EHR technology by health care providers in US [1]. In Europe, to secure consistent data protection rules and to improve the information flows, the "General Data Protection Rules" have been formulated [7]. In India, the guidelines provided by Medical Council of India (MCI), Code of Ethics Regulations 2002 states that "Confidences concerning individual or domestic life entrusted by patients to a physician and defects in the disposition or character of patients observed during medical attendance should not be revealed unless their revelation is required by the laws of State" (MCI, 2002). Information Technology Act, 2000 (India) u/s 3 dealing with the 'authentication of electronic records' would provide the legal sanction and improvise the security of the data [8].

There are many techniques involved in privacy preservation of healthcare that are discussed as follows:

1.1.1. Pseudonymization Based Techniques

Pseudonymization is an integrated approach of data anonymization and data encryption. In this technique, instead of using one's real identity for various tasks in healthcare, a pseudo-identity is derived to replace user's real identity and other unique identifiable attributes. The pseudo identity can be traced back to the real user only if all information along with the answer to a pre-programmed secret question and encryption information linking real patient to his/her pseudo identity is available [1]. In this way, pseudonymization maintains a satisfactory balance between privacy and transparency as depicted in Figure 1.

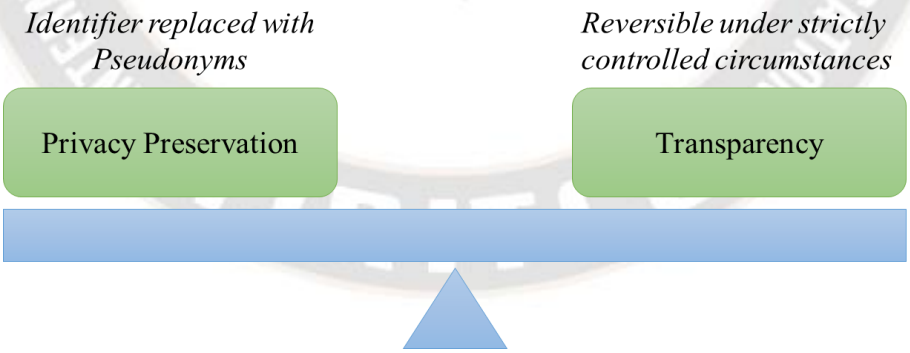

Figure 1 Trade-off between Privacy and Transparency in Pseudonymization [9]

Pseudonymization ensures the decoupling of medical data from patient identifying data, while the actual medical data is maintained and accessed by external applications. A pseudonymized database must contain at least two tables, one for the personal data and another for the pseudonyms [10]. The basic approach utilized for deriving pseudonyms (PSN) is encryption or hashing. Pseudonymization is basically applied for the secure sharing and processing of Electronic Health Records (EHR).

1.1.2. Access Control Based Techniques

Access Control allows a user to define who has access to their information and to what extent others can use it [1]. Access Control enforces the principle of least privilege i.e. a particular user has minimum set of permissions necessary to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

perform their jobs [11]. Figure 2 represents various forms of access privileges to patient records.

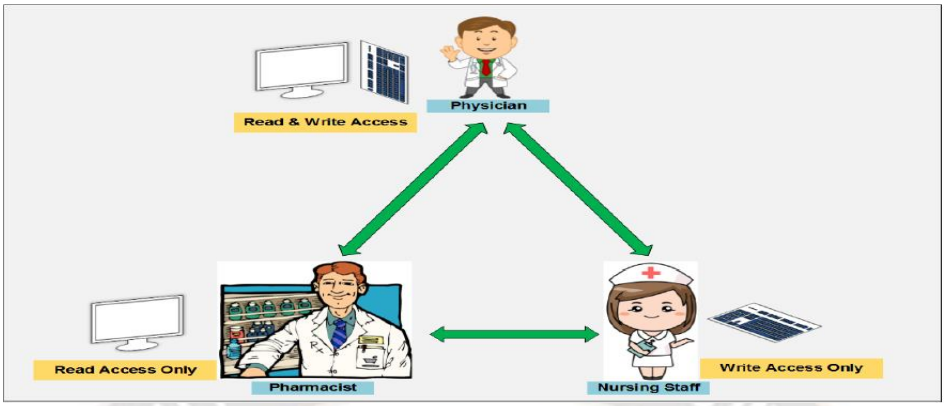

Figure 2 Access Privileges to Patient Records [1]

In a healthcare system, access control offers fundamental security barriers to data privacy by limiting the access and operations to healthcare data [12]. Access Control ensures identification, authentication, authorization and accountability (IAAA) of EHR [11]. Access control mechanisms are classified as Role based access control (RBAC), Mandatory access control (MAC), Discretionary access control (DAC), Attribute based access control (ABAC) and Identity based access control (IBAC). In RBAC, a particular role is allocated to each user and according to role-specific access privileges, it is defined that which actions can be operated on which data [1]. MAC is based on access policy decisions provided by a central authority i.e. hospitals [13] and even registered user can't change the access rights. DAC is a form of access control in which authorized user has full control over the data and can further grant access permissions to other users [11]. ABAC can provide fine-grained access control by making access privileges decisions based on attributes of database and users [11] and only the user having the attributes matched with access structure can have access to data [14]. IBAC is an approach having access control policies based on the authenticated identity of the users [1].

1.1.3. Cryptography Based Techniques

Cryptography refers to hidden writing that analyses and constructs protocols to prevent unauthorized users from accessing private data [12]. To ensure privacy in healthcare, encryption is applied at many stages. First of all, patient related sensitive data, hospital policies, diagnosis results etc. are encrypted before transmitting to outsourced environment. Also, medical data related queries need to be encrypted by authorized users before sending them for evaluation. In some cases, the query processing is also done on the encrypted query without revealing the original data [15]. The obtained diagnosis and prediction results are also reverted back in encrypted form which can be decrypted only by the authorized user. Cryptography approaches are categorized as Symmetric Key Encryption (SKE) and Public Key Encryption (PKE). SKE utilizes a single shared secret key for encryption and decryption. On the other hand, PKE entails two separate keys for encryption and decryption. One public key for encryption and one private key for decryption. Both the approaches are considered efficient for security of EHR based systems [7]. Along with SKE and PKE, some alternative cryptography primitives are also applied for medical data privacy. These include Attribute Based Encryption (ABE), Homomorphic Encryption, Proxy Re-Encryption, Searchable Encryption (SE) etc. ABE is a cryptographic primitive in which data is encrypted or decrypted on the basis of user attributes [12]. ABE enables a user to selectively share his/her EHR by encrypting the data under a set of attributes [16]. Homomorphic Encryption is the scheme that allows the computations to be performed over encrypted data without observing the original data [17]. The results obtained after processing on encrypted data is same as obtained on original data [18]. Proxy Re-Encryption permits a semi trusted proxy server to re-encrypt the cipher text into an another cipher text. Both the cipher texts are encrypted with the public keys of different users [12]. SE is a cryptographic approach that allows search operations on the encrypted data without disclosing the private information [12]. In many healthcare applications to provide privacy and efficient keyword search, SE is integrated with ABE and Proxy Re-Encryption called Attribute Based Encryption with Keyword Search (ABKS) and Proxy Re-Encryption with Keyword Search (PERS) respectively [19].

1.2. Disease Prediction in Healthcare

In this 21st century, humans are surrounded with technological advancements as these constitutes an important part in our day to day life. Also, due to environmental conditions and living habits, people come across various health related issues. People avoid to go to hospitals for small problems that may become a major disease in future. So, the prediction of disease at an earlier stage is an important task. But at the same time, the accurate prediction of disease is a tedious task. Owing to its automatic analyzing capability,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

Machine Learning (ML) is widely used in disease prediction from the past few years.
ML is a sub-field of Artificial Intelligence (AI) that enables the systems to learn by itself without being explicitly programmed [20]. ML is a method for analysis of data that iteratively learns from available data with aid of learning algorithms. Comprising of models trained from past experiences, ML make accurate predictions for the future [21]. In this way, by learning from symptoms of diagnosed patients, ML is used to predict disease in new and undiagnosed patients. ML techniques can be further classified as shown in Figure 3.

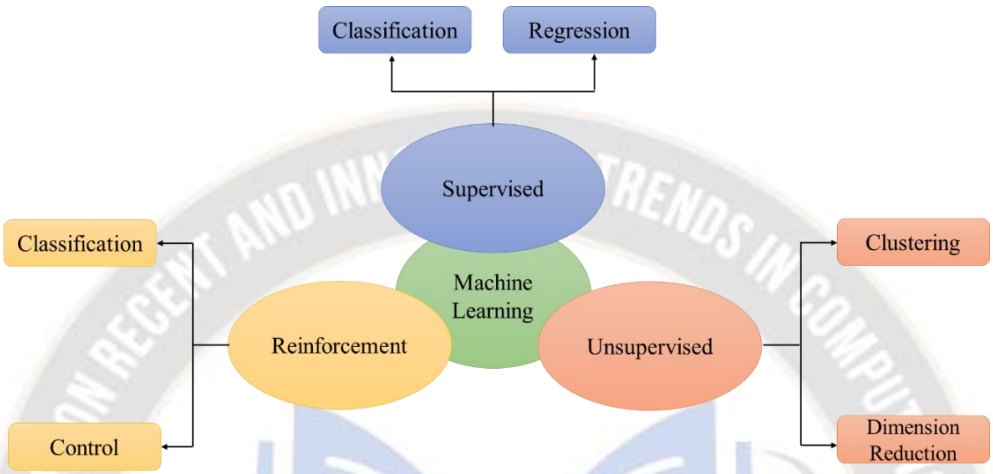

Figure 3 Classification of Machine Learning Techniques

In supervised learning, labeled data is used to train the model for future predictions [22]. Supervised learning involves Classification and regression algorithms [23]. Unsupervised Learning involves training of model from the unlabeled data i.e. target value is not known for the dataset [24]. Unsupervised learning includes Clustering and Dimension Reduction algorithms [23]. Reinforcement learning comprises of development of a system that improves performance by learning from the environment [22]. An another type of learning is Semi-supervised learning that involves the integration of supervised and unsupervised learning [24]. Semi-supervised learning enriches a small set of labeled data with additional unlabeled data [23]. Generally, disease prediction in healthcare with ML involves steps of Data Preprocessing, Feature Selection, and Prediction. The pipeline for machine learning analytics for disease prediction is shown in Figure 4. In some cases, the provided medical dataset has already preprocessed and disease prediction model can be directly applied in such cases.

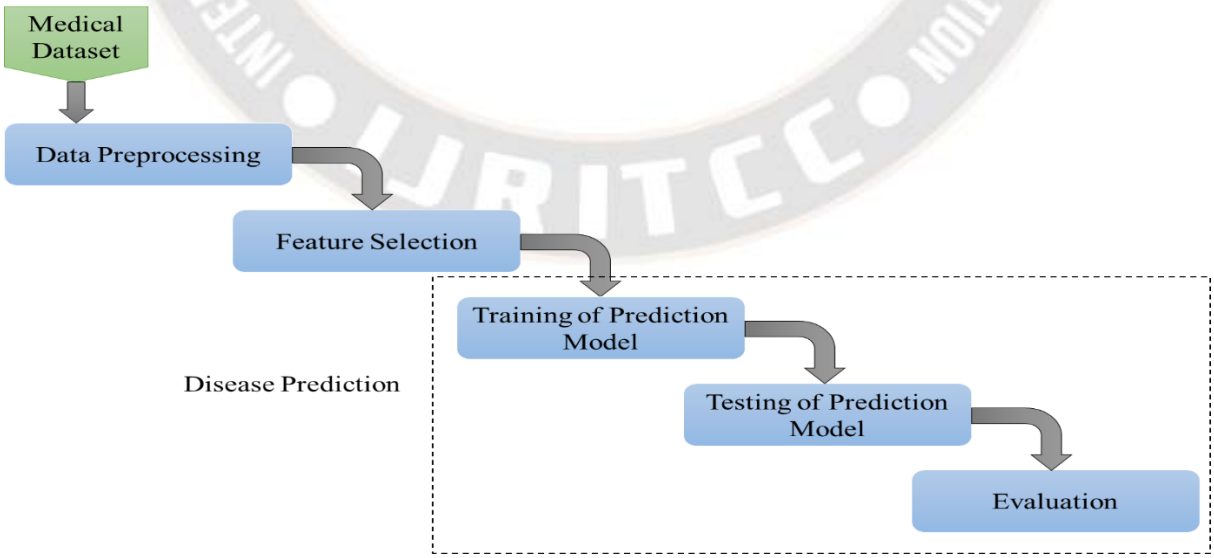

Figure 4 Framework for Disease Predictive Analytics

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

i. *Data Preprocessing:* It is the process of handling unwanted values contained in the data [25]. Any medical dataset may contain irrelevant, null, unexplained or repeated values. With the data preprocessing techniques, these values are either removed or filled with some values.
ii. *Feature Selection:* It is the process of selecting the most relevant features from the dataset [26]. In a medical dataset, there may be some attributes that do not affect the performance of prediction results. These attributes can be considered irrelevant for the further computations.
iii. *Disease Prediction:* For the disease prediction, the feature selected dataset is divided into two subsets: Training Set and Testing Set. Accordingly, disease prediction model consists of following two phases:
   a. *Training of Prediction Model:* It is the process of learning of predictive model with knowledge extracted from the training set [24]. In healthcare disease prediction, training set consists of some previously diagnosed patient's data containing symptoms, treatment suggestions etc.
   b. *Testing of Prediction Model:* It is basically the testing of the trained predictive model for testing dataset [24]. In disease prediction, the testing dataset consists of undiagnosed patients related data.

The performance of disease prediction model is evaluated by some parameters such as Precision, Accuracy, Sensitivity or Recall, and F-Measure etc.

1.3. Privacy-Preserving Disease Prediction

Disease prediction systems play a significant role in people's life as predicting the risk of disease is essential for people to lead a healthy life. Like any other healthcare data, the disease prediction systems also contain sensitive patient related information such as patient's symptoms, diagnosis results, treatment recommendations etc. Hence, the privacy of disease prediction system is also a major concern. Owing to abovementioned requirement, the disease prediction techniques are integrated with privacy preservation techniques.

The diagnosed patient's data is aggregated and outsourced in a privacy-preserved manner. For disease prediction, the model consists of two phases: Model Training and Disease Prediction [27]. In the model training phase, the historical data is utilized to train the model. The trained model further extracts the symptoms vector for the disease from the training set data. In disease prediction phase, the trained model obtains the disease prediction results based on the extracted training data [28]. The prediction results are also in a secure format and can be disclosed only by the authorized user. This ensures the privacy of sensitive medical data. Also, the use of ML techniques guarantees the accurate disease prediction.

Figure 5 represents a privacy-preserving disease prediction system with encryption and Single Layer Perceptron model.

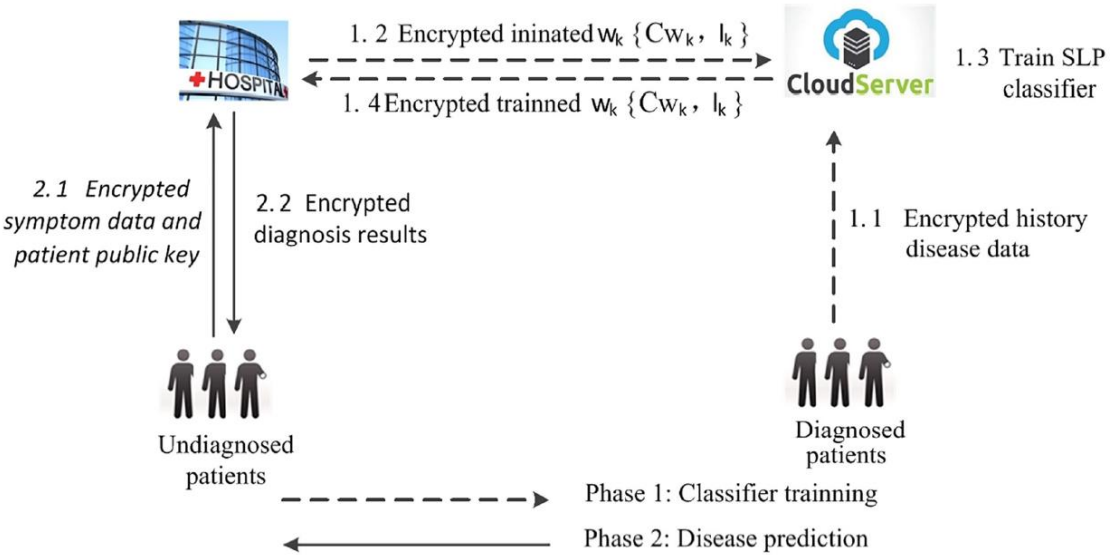

Figure 5 A Privacy-Preserving Disease Prediction Framework [27]

2. Literature Survey

Yogachandran Rahulamathavan *et al.* [29] designed a novel privacy-preserving clinical decision support system based on non-linear SVM Gaussian Kernel. In order to fulfil the privacy requirements, conventional Gaussian Kernel is re-designed by embedding Paillier Homomorphic cryptosystem. Diagnosed patients' data is encrypted and outsourced to cloud over internet. Service provider uses Paillier encryption to directly process encrypted data. In order to remove the limitation of Paillier cryptosystem to support only integer, a novel technique is introduced to scale the continuous variables involved in the process to integers. Further, the classification model is trained using this encrypted data. Trained model obtains the diagnosis results for new patients on the basis of their symptoms. The system model is implemented on Wisconsin Breast Cancer (WBC) and Puma India Diabetes (PID) dataset from UCI machine learning repository. Results demonstrate that designed system model

provides high accuracy results without compromising the privacy. **Ximeng Liu *et al.* [30]** proposed a new patient-centric clinical decision support system, called PPCD. The proposed system is based on Naïve Bayesian classification to allow the disease risk prediction in privacy-preserving way. Past patients' medical data are stored in cloud and utilized to train the classifier. The trained service provider can use the trained classifier to predict the disease risk for new patients according to the symptoms. In order to minimize the disclosure of past patients' sensitive data to service provider, a novel aggregation scheme called Additive Homomorphic Proxy Aggregation (AHPA) is introduced that allows training of classifier without having any information about patient's historical data. Also, to leverage the leakage of diagnosis results of classifier, a privacy preserving top-k disease names retrieval protocol is presented. It permits the patients to securely retrieve top-k diagnosed disease names according to their preferences. The proposed system technique is implemented with one real dataset called Acute Inflammations Dataset (AID) and another synthetic dataset. Simulations demonstrates that the system can predict disease risk with high accuracy in a privacy preserving way. **Guoming Wang *et al.* [31]** proposed a privacy-preserving disease prediction scheme for e-healthcare big data. The proposed scheme is characterized by employing Privacy-preserving Single Layer Perceptron (PSLP) learning scheme based on Paillier homomorphic cryptosystem. In this scheme, the hospital outsources the sensitive patient information to cloud in the encrypted form by using Paillier cryptosystem. Cloud server further executes the encrypted data for neural network training to obtain the disease prediction model. The scheme is executed on Wisconsin Breast Cancer (WBC) database (January 8, 1991). Detailed analysis reveals that proposed method achieves the prediction target with comparatively low computational cost and communication overhead. **Guoming Wang *et al.* [32]** developed an efficient and privacy-preserving pre-clinical guidance scheme, called PGuide. This smartphone based scheme allows on-the-go medical services and self-diagnosis to patients while preserving the privacy of medical users and service provider. A Privacy-Preserving Comparison Protocol (PPCP) is introduced in PGuide that enables users to provide more detailed health information securely and attain precise disease risk prediction without leaking the privacy out. The proposed scheme is demonstrated with an android based user-end application on smart-phone and a service application in Java. Security analysis shows that PPCP achieve enough security and hence efficient for practicality of PGuide. **P. M. Lavanya *et al.* [33]** proposed a privacy-preserving decision support system for disease prediction based on symptoms of patient. The personal health records of the patients stored in local database are outsourced to cloud server with the use of big data. For the secure storage of health records, each individual patient' information is stored in an encrypted manner by providing a unique ID. The information regarding a particular patient can be extracted with the help of this unique ID within an elapsed tolerable time. The Personal Health Records are encrypted using Homomorphic Based Encryption (HBE) to keep the information confidential. Hadoop is utilized for the disease prediction using clustering and classification techniques. **Qinghan Xue *et al.* [34]** introduced a Privacy-Preserving Disease Treatment, Complication Prediction System(PDTCPS). To enhance the search accuracy and storage efficiency, PDTCPS comprises of tree-based structure, fuzzy keyword search and Bloom Filter technique. An encrypted index tree is generated to enable fuzzy keyword queries. Top-level node of each tree index contains an encrypted category of a body part. Bloom Filter consists of disease keywords classified under this top level node and associated fuzzy keyword sets. 2nd level nodes of index tree contain three nodes, one for complication prediction, one for treatment and one for diagnosis. These nodes further stores information about k disease i.e. its training model, encrypted feature set and corresponding Bloom Filter fuzzy keyword set. Hence, following a tree like top-down approach, PDTCPS provide disease diagnosis, treatment and complication prediction. PDTCPS allows public cloud and healthcare providers to collectively generate disease training models. PDTCPS is implemented on Pima Indians Diabetes Dataset and Wisconsin Breast Cancer Dataset from UCI machine leaning repository. The results reveal that PDTPCS is highly secure and efficient in disease treatment and complication prediction. Also, it comparative analysis shows that it is better than two existing schemes CAM and HDBS. **Hui Zhu *et al.* [35]** proposed an efficient and privacy-preserving medical pre-diagnosis scheme, termed as eDiag. On the basis of improved expression for non-linear kernel Support Vector machine (SVM), a novel privacy-preserving classification scheme is introduced with lightweight polynomial aggregation and multiparty random masking. The medical user query data is pre-processed with random masking technique. Service provider further calculates the encrypted query information by decision function of classifier with polynomial aggregation. The medical user will obtain the pre diagnosis result in encrypted form and can be decrypted only by the registered user. Experiment analysis on real time environment reveals that proposed scheme achieves the prediction with enough security strength and privacy preserving ability. **Jiaping Lin *et al.* [36]** proposed a novel privacy-preserving Predictive Clinical Decision scheme (PCD) based on Recurrent Neural Network (RNN) for e-healthcare. Patients' historical data is used to train RNN models, which further predict disease according to real-time symptoms. In PCD, homomorphic encryption is utilized to encrypt the data in order to prevent privacy leakage. Sequential and averaged RNN model is designed for the improved real-time accuracy of disease prediction. Simulation results show that proposed scheme efficiently predict disease while preserving the privacy. **Ximeng Liu *et al.* [37]** designed a privacy-preserving clinical decision support system, referred as Peneus. The system is developed for outsourced cloud environment and is based on Naïve Bayesian Classifier. Patients' historical data is used to train the classifier. The trained model further predicts the disease

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

according to symptoms collected by real time patient monitoring. In order to allow multiple users to outsource PHI data to cloud platform, Peneus constructs secure Single Instruction Multiple Data (SIMD) integer circuits using Fully Homomorphic Encryption (FHE). A privacy-preserving PHI aggregation protocol is adopted so that the individual PHI from different owners can be securely accumulated for training of classifier. The trained Naïve Bayes Classifier utilizes encrypted PHI for privacy-preserving disease classification in the outsourced cloud. Simulation results illustrate that Peneus provides accurate health monitoring without any privacy loss to unauthorized parties. **Xiaoxia Liu *et al.* [38]** developed an efficient and privacy-preserving primary diagnosis scheme, called PDiag. Primary diagnosis service can be provided without compromising the privacy of sensitive health information of patient and diagnosis model of service provider. A lightweight polynomial aggregation technique based on improved expression of Naïve Bayes is introduced in PDiag. The encrypted user query is directly processed by service provider without decryption. Also, the diagnosis result can be decrypted only by the registered user. Detailed security analysis reveals that proposed scheme's privacy preserving ability and security strength with tolerable overhead in communication and computation, makes it efficient for real time environment. **Chuan Zhang *et al.* [39]** designed a Privacy-Preserving Disease Prediction system, termed as PPDP. PPDP comprises of medical data encryption, disease learning and disease prediction algorithms that novelly utilize random vectors and matrices. Diagnosed patients' medical data is encrypted and outsourced to cloud server. This data is further utilized to train the prediction models with Single Layer Perceptron algorithm. On the basis of prediction models, the risk of disease for undiagnosed patients can be computed. The designed system technique is applied on real datasets of Breast Cancer and Heart Disease as well as on randomly generated synthetic dataset. Detailed analysis shows that PPDP achieves high level of privacy preservation along with accurate disease prediction. **Alia Alabdulkarim *et al.* [40]** proposed a privacy-preserving healthcare system. It is integration of two subsystems, one is Privacy-Preserving Clinical Decision-Support System (PPCDSS) and another is Privacy-Preserving Mobile Health Social Network (MHSN). PPCDSS consists of two phases: Training and Testing. In the training phase, patients' historical data is collected and used to train the decision tree model. For testing, the trained classifier is utilized to diagnose new coming patients' without disclosing patients' information. MHSN is basically utilized for providing emergency call service. Patients' health condition is continuously monitored remotely through Wireless Body Sensors and in case of any abnormal condition, immediate care service is provided. Social Network connects together the people of having similar symptoms. Hence, it enables the service of seeking help of near-by passing people while the patient is waiting for an ambulance to arrive. The proposed model is expected to enhance healthcare service while providing privacy to patients' related data. **Jeongsu Park *et al.* [41]** designed a privacy-preserving medical diagnosis system based on *k*-Nearest Neighbor (*k*NN) classification. By using e-Health cloud service, the system provides diagnosis for medical databases owned by multiple data owners. The system model ensures the privacy of patient's symptoms, medical dataset and diagnosis results by hiding the data access pattern even from e-health cloud server. *k*NN classifier is redesigned as privacy-preserving protocol for finding the *k* data with highest similarity to a given symptom (PE-FTK). To avoid security issues related to multiple untrusted cloud servers, PP*k*NN is characterized by employing Multiparty Computation (MPC) based on secret sharing to compute *k*NN results in a distributed manner without any server. MPC further prevents collusion attacks among various servers and between data owner and cloud server. Experiment evaluations reveal that as compared with previous works, the designed model offers 35% reduction in average running time. Also, diagnosis results of the system are deterministic i.e. without any error probability. **Wei Guo *et al.* [42]** proposed a Privacy-Preserving Online Medical Pre-Diagnosis scheme, called POMP. Based on logistic regression, the proposed scheme is well-suited for cloud environment. Patients' historical data and hospital's pre-diagnosis model is encrypted and outsourced to cloud server. Homomorphic cryptosystem is employed to perform medical examination Med Exam (.) function directly on encrypted data. Further, the lightweight pre-diagnosis result is obtained by exploiting the pre-processing technique and Bloom Filter. Extensive security and performance analysis shows the privacy-preserving ability and security strength of proposed scheme with high efficiency in computation and communication overhead. **Ximeng Liu *et al.* [43]** developed a Hybrid Privacy-Preserving Clinical Decision Support system (HPCS) for fog-cloud computing. A lightweight single layer neural network data mining technique is adopted by fog server for the real time monitoring of patient's health condition. Specifically, a novel privacy-preserving outsourced inner-product protocol is designed for fog servers in order to achieve lightweight single layer neural network. Historical PHI and symptoms collected are securely outsourced to cloud server. For the prediction of disease, a privacy-preserving piecewise polynomial calculation protocol is used that enables cloud server to process non-linear activation functions in multiple layer neural network. A novel protocol termed as privacy-preserving fraction approximation is developed to avoid overflow problem occur as a result of unlimited iterative calculations. Real time experiments reveals that the proposed system attains the goal of privacy-preserved health monitoring with low communication overhead.**Xue Yang *et al.* [28]** proposed an Efficient and Privacy-Preserving Disease Risk Prediction scheme (EPDP) for e-healthcare environment. Keeping in mind the privacy preservation requirement, EPDP comprises of two phases: Disease Model Training and Disease Prediction. In disease model training phase, diagnosed patients' historical data is collected and encrypted with Okamoto-Uchiyama (OU) cryptosystem. To reduce the

encryption time for medical data, a super-increasing sequence is introduced in EPDP. This sequence compress multidimensional data to 1-D and further encryption is done on the compressed data. Encrypted data is outsourced to cloud server and used to train the Naïve Bayes Classifier. On the basis of extracted training results by classifier, Bloom Filter technique is used to obtain the prediction result for undiagnosed patients in disease prediction phase. Proposed technique is applied to one real dataset and one synthetic dataset. Real data set is Acute Inflammations Dataset (AID) from UCI machine learning repository. Extensive simulations illustrate that EPDP is more efficient in communication and computation overhead reduction with accurate prediction results than existing schemes. Hence, making EPDP suitable for real-time e-healthcare environment especially in case of medical emergency. **Alia Alabdulkarim *et al.* [44]** developed a clinical decision support system based on Privacy-Preserving Random Forest (PPRF) algorithm. Multiple parties can collaborate in a privacy-preserved manner to create a classification model. On the basis randomly drawn subset from its dataset, each party generates a decision tree to create local random forest model. After the aggregation of local random forest models, a global random forest ensemble is formed. This global random forest ensemble model diagnoses the new patients' symptoms without disclosing medical information to unauthorized parties. Developed system is implemented on Acute Inflammations Dataset taken from UCI machine learning repository with one modification of removing the "Temperature" attribute. Simulation results illustrate that due to avoidance of cryptography implementations and removal of unnecessary attributes, proposed PPRF performs better in terms of average runtime than existing ones. **Malathi D. *et al.* [45]** designed a hybrid reasoning-based Privacy-Aware Disease Prediction Support System (PDPSS). To enhance the accuracy, the system is designed by integration of Fuzzy Set Theory, *k*-Nearest Neighbor classifier and Case-Based Reasoning classifier (F*k*NN-CBR). Paillier homomorphic cryptosystem is employed to encrypt the sensitive patient information. To predict the disease occurrence, hybrid reasoning approach is applied on encrypted data. Further, the predicted results and treatment advices are transmitted to patient in a privacy preserving manner. Indian Liver Patient Dataset from UCI Machine Learning Repository is used for the system implementation. The experimental results demonstrate that the proposed FKNN-CBR offers high sensitivity, prediction accuracy and specificity over existing baselines. **Hui Ma *et al.* [27]** proposed a Privacy-Preserving Clinical Decision (PPCD) with cloud support based on single layer perceptron. The framework consists of two phases: SLP Model Training and Disease Prediction. In model training phase, diagnosed patient's historical data is encrypted and outsourced to cloud server. This encrypted data is utilized to train the single layer perceptron model for disease prediction. A Lightweight Secure Multiplication (LSM) is introduced to enhance the model training efficiency. The trained model is sent to hospital which further performs disease prediction for undiagnosed patients. The framework is applied on Heart Disease Dataset (HDD), Acute Inflammations Dataset (AID) and Wisconsin Breast Cancer Dataset (WBCD) from UCI machine learning repository. Experimental analysis witnesses the achievement of high accuracy of disease prediction in PPCD without the privacy disclosure risk. **Zhuoran Ma *et al.* [46]** introduced a Privacy-Preserving and High-Accurate outsourced Disease Predictor on Random Forest, termed as PHPR. It ensures secure model training with medical data and results in accurate disease prediction. Multiple data owners encrypt their data and collectively train the RF classifier on-the-fly and in a privacy-preserved manner. To secure the sensitive information of individual dataset, privacy-preserving computational protocols over rational numbers are designed in data outsourcing. These protocols avoid the only integer support limitation of Paillier Cryptosystem. Further, the trained RF classifier obtains the disease prediction results for undiagnosed patients. For the real time implementation of the prediction scheme, Heart Disease Dataset from UCI machine learning repository is selected. Experimental results reveal that PHPR provides highly accurate prediction result while maintaining the privacy of the medical information. **Alia Alabdulkarim *et al.* [47]** designed a Privacy-Preserving Clinical Decision Support System (PPCDSS). Based on novel Privacy-Preserving Single Decision Tree algorithm (PPSDT), the system diagnosis the new symptoms without the medical data disclosure risk. Multiple hospitals collaboratively share historical diagnosed patients' data via cloud in a privacy-preserved manner to build the classification model of CDSS. This trained classification model further utilized for disease diagnosis. In order to minimize privacy risks, homomorphic encryption is employed to process the aggregated data. On another end, nonce is used to prevent one party from decrypting other parties' data and diagnosis results. Designed system is implemented on Acute Inflammations Dataset from UCI machine learning repository with one modification of removing the "Temperature" attribute. Simulation results illustrate that designed PPCDSS outperforms Naïve Bayes Classifier by 46.67% along with providing secure transmission and processing of patient related data. **Dan Zhu *et al.* [48]** proposed a privacy-preserving multi-level pre-diagnosis scheme, termed as CREDO. Based on Multi-Level *k*-Nearest Neighbor (ML-*k*NN), CREDO provides an efficient pre-diagnosis service without the disclosure of sensitive data. Data collected from the multiple parties is encrypted and outsourced to cloud server. Service provider directly operates on encrypted data without obtaining the original information. Initially, using the k-mean clustering, service provider narrows down the scope of medical instances needed to be computed. Then, ML-*k*NN classification is utilized on medical instances to provide pre-diagnosis service. Also, in order to ensure privacy preservation, pre-diagnosis results can only be decrypted by the registered owner. The proposed scheme is applied to one real dataset called Medical from Mulan (a Java library for multi-label learning). Detailed analysis reveals that along with privacy-preserving ability of the CREDO, it provides an efficient pre-

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

diagnosis service with low computation overhead and high accuracy. **Fengwei Wang *et al.* [49]** developed an efficient and privacy-preserving disease risk assessment scheme, called CARER. Based on Naïve Bayes Classification, the CARER can securely train a disease risk prediction model over multi-sourced vertically distributed data. Modified Paillier Homomorphic cryptosystem is used to encrypt the medical centres' data which is further utilized for training of disease risk prediction model. A model updating strategy is designed in the system that enables medical centres to outsource their fresh collected data. For ensuring the privacy of trained model of service provider, a random masking technique is applied in disease risk prediction. The prediction results are obtained in encrypted form and can be decrypted only by registered user. The developed scheme is implemented on one real Breast Cancer Wisconsin(BCW) dataset and another randomly generated synthetic dataset. Results show that the improved disease risk predicting efficiency of CARER along with high privacy preserving ability and security. **T. Munirathinam *et al.* [50]** proposed a new e-healthcare system for monitoring the disease level. The system analyses the level of dead diseases i.e. Cancer, Heart Disease and Diabetes, by applying fuzzy rules and deep learning approach and using technologies such as Cloud and IoT. Medical information is collected from remotely located patients who are utilizing e-healthcare services. The data is encrypted and outsourced to the cloud. For the storage of data in cloud, a new ECC based Secure Storage Algorithm (ECC-SSA) is introduced in the system. Convolutional Neural Network (CNN) is utilized for learning of data and clear prediction of disease. The severity level of the disease is predicted by applying fuzzy rules. Proposed system is implemented on Heart Dataset (HDD), Diabetic Dataset and Wisconsin Diabetic Breast Cancer (WDBC) Dataset from UCI machine learning repository. Implementation results demonstrate that the proposed system is highly secure and obtains prediction results with 99% accuracy. **Mingwu Zhang *et al.* [51]** designed an efficient and privacy-preserving clinical diagnosis scheme based on Support Vector Machine (SVM). By using the outsourced cloud platform, the scheme provides medical diagnosis assistance for doctors in a without disclosing the medical data and diagnosis model. A privacy-preserving multiclass SVM scheme is introduced for secure multi-diseases diagnosis. The historical medical data and doctor's query is encrypted and sent to diagnosis server. The SVM model is pre-trained and encrypted parameters of the model are outsourced to medical cloud server. An encoding approach is applied for the encryption of negative parameters of SVM. To perform operations on the encrypted data directly and provide clinical diagnosis with privacy-preserving multiclass SVM, various building blocks are designed such as privacy-preserving classification, privacy-preserving computation of decision function and search of max decision functions on encrypted data. For any Based on the results of encrypted trained classifier, the encrypted disease diagnosis results are obtained and can be decrypted as an aided diagnosis only by authorized users. The scheme is applied to Dermatology Dataset from UCI machine learning repository. Experimental analysis reveals the practical feasibility of the designed scheme along with privacy-preserving ability and high efficiency. **Ma *et al.* [52]** introduced a Lightweight Privacy-Preserving diagnosis mechanism on Edge (LPME) computing. It reconstructed the Extreme Gradient Boosting (XGBoost) model within the edge-cloud paradigm. The LPME system develops an XGBoost-based diagnosis model using model parameters trained across multiple edge nodes, eliminating the need for storing extensive training data. Additionally, the LPME system employs Homomorphic Encryption (HE)-based secure computation in a single-cloud model, optimizing parameters over encrypted model parameters during training. With a randomly split secret key, only one part is stored in the single cloud, ensuring robust privacy preservation for lightweight XGBoost training and reliable privacy-preserving training on resource-limited edges. Users can submit encrypted requests to an edge and corresponding diagnosis results are returned to the user in a homomorphic encrypted format. The proposed approach is implemented on Thyroid disease dataset and Heart disease dataset. The experimental results validated the effectiveness and security of the LPME system in edge computing. **Liang *et al.* [53]** designed a privacy-preserving decision tree (PPDT) classification system for health monitoring systems. Initially, the decision tree classifier, representing the clinical decision model, is converted into Boolean vectors, and symmetric key encryption is applied to encrypt these Boolean vectors in order to protect the privacy of classification model and medical data. Breast Cancer Wisconsin dataset is utilized to evaluate the proposed PPDT scheme. Performance analysis demonstrate that the PPDT scheme is highly efficient in terms of computational time, storage consumption and complexity as it requires execution times only at the microsecond level, communication costs in the kilobyte range, and storage costs in the kilobyte range for the test dataset. **Xie et al. [54]** introduced a privacy-preserving scheme for cloud-aided diagnosis in the Internet of Medical Things (IoMT). The scheme employed a hybrid data encryption approach, combining homomorphic encryption and Advanced Encryption Standard (AES), for the efficient generation of user requests. Additionally, a set of secure two-party protocols leveraging homomorphic encryption is proposed. These protocols encompass secure computations for kernel functions, multiplication, and comparison. By utilizing these building blocks, we establish a privacy-preserving diagnosis scheme based on multi-class Support Vector Machine (SVM). The proposed technique is evaluated on synthetic as well as on real dataset of cardiotocography dataset from UCI machine learning repository. Obtained results revealed that the devised scheme not only safeguards users' input data and diagnosis results but also prevents the cloud from acquiring knowledge of the diagnosis model. Furthermore, the designed approach minimizes computation and communication costs on clients, making it well-suited for the Internet of Medical Things (IoMT) environment. **Gopalan *et al.* [55]** designed an

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

effective privacy-preserving (PP) scheme designed for patient healthcare data obtained from IoT devices, specifically tailored for disease prediction within the contemporary Health Care System (HCS). The proposed system employed Log of Round value-based Elliptic Curve Cryptography (LR-ECC) to secure the data transfer following the initial authentication phase. Only authorized healthcare personnel can securely retrieve patient data on the hospital side. The integration of the Herding Genetic Algorithm-based Deep Learning Neural Network (EHGA-DLNN) allows for the testing of this data using the trained system to predict diseases. The proposed scheme is implemented on Hungarian dataset which demonstrate superiority of the proposed work to the existing systems in disease prediction, offering enhanced privacy and security features. **Rehman *et al.* [56]** proposed a privacy preserving deep learning based diagnosis model for pneumonia disease. The process is divided into two phases: the initial phase is dedicated to privacy preservation in which, a chaos and convolutional neural network (CNN) based scheme is designed for the encryption of pneumonia image dataset. Further, multiple chaotic maps are integrated to form a random number generator, and the resulting random sequence is employed for both pixel permutation and substitution. The second phase consists of dep learning based diagnosis of pneumonia in which CNN is used as classification model and encrypted x-ray images is used as pneumonia dataset. The physiological parameters of pneumonia are extracted from provided x-ray images. In order to achieve the higher accuracy of the designed model, fine tuning and transfer learning are also incorporated with CNN. The proposed scheme is validated by implementing on 5300 verified chest x-ray images of healthy and pneumonia patients. Along with providing enough privacy, the CNN model achieves classification accuracy of 97% on the considered dataset which is higher as compared to existing machine learning and deep learning models. **Zhao et al. [57]** designed a novel privacy preserving outsourced multiclass SVM disease diagnosis model. By applying Paillier cryptosystem fundamental operation algorithms crucial for the storage and computation of outsourced data from multiple data owners, are generated. These include secure aggregation and multiplication algorithms, forming the foundation of proposed SVM training protocol. In the diagnosis phase, BFV cryptosystem based secure comparison algorithm and maximum finding algorithm are incorporated with the SVM model in order to achieve privacy of the user's symptoms and diagnosis results. The proposed scheme is implemented on Dermatology dataset from the UCI machine learning repository. The results show that the proposed scheme not only achieves an adequate diagnosis accuracy, but also reduces the computational overhead. **Chen *et al.* [58]** introduced a privacy-preserving medical diagnosis system utilizing the Distributed Two Trapdoors Public Key Cryptosystem (DT-PKC) and the Boneh-Goh-Nissim (BGN) cryptosystem. This ensures the confidentiality of user data and support vectors in SVMs classification for medical diagnosis services. A secure computing protocol tailored for multi-class SVMs, comprising multiple binary classification SVMs is designed. This protocol accommodates SVMs with diverse kernel functions and exhibits greater scalability compared to existing schemes limited to a single kernel function. Furthermore, the authors present a user authentication scheme to verify the legitimacy of user identity, preventing potential malicious attacks from unauthorized users on the medical diagnosis system. The designed privacy preserving SVM is evaluated on Dermatology dataset from the UCI machine learning repository. Results demonstrate the reliability, security and scalability with high classification accuracy. **Zhou *et al.* [59]** proposed a logistic regression based privacy preserving disease diagnosis system (LR-DDH). Initially, a linear homomorphic authenticated encryption algorithm is designed which is further utilized in online diagnosis scheme. In this scheme, patients can securely access medical diagnosis services remotely. Notably, the key for homomorphic authenticated encryption remains undisclosed to the cloud server, enabling outsourced computing even in a fully untrusted cloud environment. The proposed approach guarantees the confidentiality and integrity of learning model parameters, patient queries, and diagnostic results. **Shen *et al.* [60]** designed a privacy preserving and efficient federated learning mechanism (FLM) based online diagnosis scheme for e-healthcare environment. Initially, this scheme converts the data owner's data sharing challenge into a machine learning problem using FLM. This involves sharing computed local model parameters instead of the actual data, effectively safeguarding the privacy of training datasets. Subsequently, a homomorphic cryptosystem alongside the Support Vector Machine (SVM) algorithm is employed for the efficient classification of patients' physiological data without compromising their privacy. Additionally, an innovative method to reconstruct the decision function of the SVM model is introduced, effectively preventing any leakage of model parameters. The proposed scheme is implemented on Dermatology dataset, diabetes dataset and hepatitis C virus (HCV) dataset from UCI machine learning knowledge database. Security analysis and performance evaluation of the proposed scheme affirm its capability to fulfill security requirements while exhibiting high clinical diagnosis accuracy in real-world e-healthcare systems.

Table 1 represents the summary of state-of-the-art privacy preserving disease diagnosis/prediction techniques.

Table 1. Comparative analysis of existing privacy preserving disease diagnosis systems

Ref	Contribution	Techniques Used	Dataset / Disease	Limitations	Future Scope
[29]	Designed a novel privacy-preserving clinical decision support system	Non-linear Support Vector Machine Gaussian Kernel,	Wisconsin Breast Cancer (WBC),	The complexity of the system is high due to non-parametric nature of Gaussian Kernel.	To optimize the system to better predict diseases at low

		Paillier Homomorphic Encryption	Pima India Diabetes (PID)	Multiple communication rounds are required between patient and storage server which further leads to communication overhead.	communication costs without compromising security attributes
[30]	Proposed a new Patient-Centric Clinical Decision Support system (PPCD)	Naïve Bayesian Classification, Additive Homomorphic Proxy Aggregation (AHPA)	Acute Inflammations Dataset (AID)	Can only process symptoms with bit value during decision. For non-bit values i.e. temperature, it needs to expand these values into long bit string before prediction	To employ other advanced data mining techniques such as SVM classification in PPCD
[31]	Proposed a privacy-preserving disease prediction scheme for e-healthcare big data	Single Layer Perceptron Learning Scheme (PSLP), Paillier Homomorphic Encryption	Wisconsin Breast Cancer (WBC)	Complexity of the scheme increases with increase in number of input nodes in PSLP	To deal with more efficient and privacy-preserving big-data medical model training algorithms
[32]	Developed an efficient and privacy-preserving smartphone based pre-clinical guidance scheme (PGuide)	Privacy Preserving Comparison Protocol (PPCP)	----	----	To enhance the performance of scheme by dealing with other practical security and privacy issues in disease risk predication model
[33]	Proposed a privacy preserving decision support system for disease prediction	Homomorphic-Based Encryption (HBE), Hadoop using clustering and classification	Personal Health Records	Homomorphic Encryption bring heavy computation overhead due to massive time-consuming operations such as bilinear pairing.	To employ efficient privacy preserving techniques in terms of computation cost
[34]	Introduced a Privacy-Preserving Disease Treatment, Complication Prediction System (PDTCPS)	Encrypted Index tree-based structure, Fuzzy Keyword Search, Bloom Filter	Pima Indians Diabetes (PID) and Wisconsin Breast Cancer (WBC)	Complexity of system increases with number of top level nodes in tree based structure.	To enhance the proposed scheme to support more complex query types To improve the performance of the system using other types of encrypted health data
[35]	Proposed an efficient and privacy-preserving medical pre-diagnosis scheme, termed as eDiag	Non-linear kernel Support Vector machine (SVM), Lightweight Polynomial Aggregation, Multiparty Random Masking	Pima Indian Diabetes (PID) Dataset	Computation overhead increases when the number of support vectors is large Cannot deal with multi diseases classification	To deal with multi disease classification in order to enhance the performance
[36]	Proposed a novel Privacy-Preserving Predictive Clinical Decision scheme (PCD)	Homomorphic Encryption, Sequential Recurrent Neural Network (RNN)	e-Health Records Dataset	Integration of homomorphic encryption with sequential RNN results in higher computational load.	To implement integrated models of advanced data mining techniques to obtain higher accuracy for early prediction
[37]	Designed a privacy-preserving clinical decision support system, referred as Peneus	Fully Homomorphic Encryption (FHE), Naïve Bayesian Classifier	Breast Cancer Wisconsin (Diagnostic) Dataset	FHE based systems comes up with high complexity	To construct efficient bootstrapping techniques to improve the performance of FHE for more efficient decision making of the system
[38]	Developed an efficient and privacy-preserving primary diagnosis scheme (PDiag)	Lightweight Polynomial Aggregation Technique	Wisconsin Breast Cancer (WBC) [and Heart Disease (HD) datasets	Computation overhead of proposed scheme increases when the number of disease classes is large	To optimize the scheme for multi class diagnosis
[39]	Designed a Privacy-Preserving Disease Prediction system (PPDP)	Single Layer Perceptron algorithm	Breast Cancer, Heart Disease	Inefficient to capture disease patterns in high dimensional healthcare data	To design more efficient privacy preserving disease prediction model
[40]	Proposed a privacy-preserving healthcare system. Integration of two subsystems, one is Privacy-Preserving Clinical Decision-Support System (PPCDSS) and another is Privacy-Preserving Mobile Health Social Network (MHSN)	Decision Tree Algorithm	Health records from three hospitals	Decision tree algorithm sometimes provides inaccurate results due to overfitting problem in case of high-dimensional healthcare data	To employ advanced classification techniques to deal with large number of instances in data
[41]	Designed a privacy-preserving medical diagnosis system	k-Nearest Neighbor Classification, Multiparty Computation (MPC)	Patient Health Records (PHR)	Multiparty computation leads to massive communication overhead as it require multiple interactions to complete a specific operation over cipher text	To construct the privacy preserving and efficient protocols for data mining techniques other than *k*NN
[42]	Proposed a Privacy-Preserving Online Medical Pre-Diagnosis scheme (POMP), based on logistic regression	Homomorphic Encryption, Preprocessing Technique, Bloom Filter Technique	----	----	To extend the proposed scheme to other regression models
[43]	Developed a Hybrid Privacy-Preserving Clinical Decision Support System (HPCS) for fog-cloud computing	Lightweight Single Layer Neural Network	Breast Cancer Wisconsin (Diagnostic) Dataset	Can support only linear operations for clinical decision making	To design privacy-preserving models to support for non-linear operations i.e. Sigmoid function for achieving better decision rate

[28]	Proposed an Efficient and Privacy-Preserving Disease Risk Prediction scheme (EPDP) for e-healthcare	Okamoto-Uchiyama (OU) Cryptosystem, Super-Increasing Sequence, Naïve Bayes Classifier	Acute Inflammations Dataset (AID)	Difficult to use for high dimensional data encryption as the coefficients of super increasing sequence increase exponentially	To improve the accuracy by achieving the message integrity in e-healthcare To integrate access control with scheme in order to give access to professional authorities for the proper understanding of prediction results
[44]	Developed a clinical decision-support system	Random Forest Algorithm	Acute Inflammations Dataset (AID)	Cannot deal with classifier training from multiple data sources	To ensure integrity of the model with Message Authentication Code (MAC)
[45]	Designed a hybrid reasoning-based Privacy-Aware Disease Prediction Support System (PDPSS)	Homomorphic Encryption, Fuzzy Set Theory, k-Nearest Neighbor Classifier, Case-Based Reasoning Classifier (FKNN-CBR)	Indian Liver Patient Dataset	----	To develop an extensive privacy-aware model to better predict diseases at low computational and communication costs without compromising security attributes
[27]	Proposed a Privacy-Preserving Clinical Decision (PPCD) with cloud support	Single Layer Perceptron (SLP), Lightweight Secure Multiplication (LSM)	Heart Disease Dataset (HDD), Acute Inflammations Dataset (AID),Wisconsin Breast Cancer Dataset (WBCD)	The time cost for underlying operations increases with the number of sample cases	To optimize the model training using mini-batch for efficiency improvement To find an effective way of introducing other advanced machine learning methods to build the privacy-preserving disease prediction system
[46]	Introduced a Privacy-Preserving and High-Accurate outsourced Disease Predictor on Random Forest (PHPR)	Random Forest Classifier, Paillier Homomorphic Encryption	Heart Disease Dataset	Paillier Homomorphic Encryption can encrypt only integer values	To improve the efficiency of PHPR and extend the work to other efficient classification techniques
[47]	Designed a Privacy-Preserving Clinical Decision-Support System (PPCDSS)	Homomorphic Encryption, Privacy-Preserving Single Decision Tree Algorithm (PPSDT)	Acute Inflammations Dataset (AID)	Privacy of diagnosis model is not taken into account.	To generalize the algorithm to accept more forms of datasets
[48]	Proposed a privacy-preserving multi-level pre-diagnosis scheme, termed as CREDO	Multi-Level k-Nearest Neighbor (ML-kNN)	Medical Dataset from Mulan (a Java library for multi-label learning)	----	To incorporate disease treatment methods along with developed scheme
[49]	Developed an efficient and privacy-preserving disease risk assessment scheme, called CARER	Paillier Homomorphic Encryption, Naïve Bayes Classification	Breast Cancer Wisconsin (BCW)	Communication overhead of CARER linearly increases with number of attributes and disease classes	To optimize the scheme for high dimensional data
[50]	Proposed a new e-healthcare system for monitoring the disease level	ECC based Secure Storage Algorithm (ECC-SSA), Convolutional Neural Network (CNN)	Heart Disease Dataset (HDD), Diabetic Dataset, Wisconsin Diabetic Breast Cancer (WDBC)	ECC cryptography algorithm is highly susceptible to network attacks in data outsourcing	To introduce novel cryptographic algorithm instead of ECC To consider the temporal features in CNN for making decision effectively
[51]	Designed an efficient and privacy-preserving clinical diagnosis scheme	Multiclass Support Vector Machine (SVM), Encoding	Dermatology Dataset	Multiclass SVM may perform moderately in some problematic classes	To optimize the existing models of multiclass SVM by combining other classification methods and algorithms for healthcare diagnosis
[52]	Introduced a Lightweight Privacy-Preserving diagnosis mechanism on Edge (LPME) computing	Homomorphic Encryption, Extreme Gradient Boosting (XGBoost)	Thyroid Disease Dataset and Heart Disease Dataset	Increasing the tree height (h) in final decision node, can lead to overfitting and increase in runtime.	To explore strategies to mitigate the challenges associated with with decision nodes.
[53]	Designed a privacy-preserving decision tree (PPDT) classification system for health monitoring systems	Symmetric Key Encryption, Decision Tree Classifier	Wisconsin Breast Cancer Dataset	Overfitting problem in decision tree due to large number of parameters such as leaf nodes and decision nodes.	Further enhancement of the proposed PPDT approach to deal with malicious adversaries.
[54]	Introduced a privacy-preserving scheme for cloud-aided diagnosis in the Internet of Medical Things (IoMT).	Advanced Encryption Standard (AES), Homomorphic encryption, Support Vector Machine (SVM)	Cardiotocography Dataset	Higher complexity at the cloud side in achieving privacy of diagnosis model.	
[55]	Designed an effective privacy-preserving (PP) scheme designed for patient healthcare data obtained from IoT devices	Elliptic Curve Cryptography (ECC), Deep Learning Neural Network (DLNN)	Hungarian Dataset	Potential vulnerability of ECC-based security systems to quantum computing attacks.	To improve the model with more generic strategies to accommodate additional type of datasets.

[56]	Proposed a privacy preserving deep learning based diagnosis model for pneumonia disease.	Choas Encryption, Convolutional Neural Network (CNN)	5300 Chest X-Ray Images Dataset	----	To enhance the deep learning classification approach by using hybrid AI algorithms.
[57]	Designed a novel privacy preserving outsourced multiclass SVM disease diagnosis model	Paillier Cryptosystem, BFV cryptosystem, Support Vector Machine (SVM)	Dermatology Dataset	Correct result cannot be obtained after decryption in BFV cryptosystem, if effect of noise is not considered by setting the required value of noise budget.	Exploration and implementation of more efficient privacy preserving disease diagnosis scheme.
[58]	Introduced a privacy-preserving medical diagnosis system	Distributed Two Trapdoors Public Key Cryptosystem (DT-PKC) and the Boneh-Goh-Nissim (BGN) cryptosystem, Homomorphic encryption, Support Vector Machine (SVM)	Dermatology Dataset	Reduced efficiency in calculating inner product as BGN cryptosystem consumes substantial amount of time for bilinear mapping during homomorphic multiplication.	To design a novel system by using more efficient scheme than BGN cryptography.
[59]	Proposed a logistic regression based privacy preserving disease diagnosis system (LR-DDH)	Homomorphic Encryption, Logistic Regression	Dermatology Dataset	In the classification stage, computing time at the cloud side increases with increase in dimension of data.	Exploration and implementation of privacy preserving methods for distributed federated learning format.
[60]	Designed a privacy preserving and efficient federated learning mechanism (FLM) based online diagnosis scheme	Homomorphic Encryption, Support Vector Machine (SVM)	Dermatology Dataset, Diabetes Dataset and Hepatitis C Virus (HCV) Dataset	Communication overhead and latency introduced by the need to coordinate and aggregate model updates across distributed devices or servers.	To enhance the robustness of the proposed scheme to make it resistant against adversaries such as Byzantine attack.

3. Insights and Challenges

While privacy-preserving disease diagnosis techniques surveyed demonstrate considerable efficiency in concurrently achieving privacy and accurate diagnosis, there exist some general limitations associated with these approaches as outlined below:

- **Privacy in Multi-data sources:** Combining data from multiple sources for prediction models is powerful, but protecting privacy while doing so is challenging Existing privacy preserving disease diagnosis systems still lack effective methods to keep data secure and train accurate models in these complex settings. This gap highlights a critical need for exploration and innovation in safeguarding privacy across diverse data sources.

- **Inefficiency in terms of computation and communication overhead:** Current privacy-preserving techniques suffer from inefficiencies, particularly in terms of computation and communication overhead. These methods often impose significant computational burdens and require extensive communication resources, hindering their practicality and scalability in real-world applications. Addressing these efficiency challenges is crucial for advancing the adoption of robust privacy-preserving solutions.

- **Disease Prediction in Multi-label instances:** Limited research attention has been directed towards disease prediction models that cater to multi-label instances, where individuals may simultaneously exhibit multiple diseases. This gap underscores the need for more comprehensive investigations and innovative approaches to enhance the accuracy and applicability of disease prediction in scenarios involving multiple coexisting health conditions.

- **Privacy of prediction model:** The majority of current research overlooks the privacy dimension in prediction models, neglecting to address the critical aspect of safeguarding the confidentiality of the model itself. This oversight underscores the imperative for a more comprehensive integration of privacy considerations into existing works on predictive modeling.

- **High dimensional healthcare data:** There is a scarcity of methods for predicting diseases in high-dimensional healthcare data, primarily due to the associated challenge of overfitting. This gap emphasizes the need for innovative approaches to mitigate overfitting issues and enhance the effectiveness of disease prediction models in complex healthcare datasets.

- **Imbalanced medical datasets:** The prevailing disease prediction techniques are predominantly tailored for balanced medical datasets, with minimal attention given to the challenges posed by imbalanced datasets. There is a notable dearth of research addressing predictive analysis in the context of imbalanced medical data, highlighting the necessity for tailored methodologies to effectively handle such disparities.

4. Conclusion

This survey paper extensively examines the existing privacy-preserving techniques in disease diagnosis and prediction, offering detailed insights into methodologies and datasets. The surveyed literature also uncovers notable shortcomings of multi-source models, predicting diseases in high-dimensional data, privacy in prediction models, and gap in

addressing imbalanced medical datasets. By overcoming these identified challenges, researchers can pave the way for a future where disease prediction systems not only excel in diagnostic capabilities but also stand as paragons of secure and privacy-respecting technology in the healthcare domain. Future directions of research in the field of healthcare data privacy and disease prediction should prioritize the development of robust methods for safeguarding privacy in multi-data source settings, addressing the challenges of combining diverse datasets. Additionally, there is a critical need to explore innovative solutions for enhancing efficiency in privacy-preserving techniques, overcoming limitations in computation and communication overhead. These avenues of inquiry will contribute to advancing the accuracy, applicability, and scalability of disease prediction models in high-dimensional healthcare data and imbalanced medical datasets.

References

1. M.A. Sahi, H. Abbas, K. Saleem, X. Yang, A. Derhab, I. Rashid and A. Yaseen, "Privacy Preservation in e-Healthcare Environments: State of the Art and Future Directions," IEEE Access, vol. 6, pp. 464-478, 2018.
2. 24x7 Magazine "Global medical device market to grow 4.5%". [Online]. Available: https://www. 24x7mag.com/medical-equipment/global-medical-device-market-grow-4-5/ April 2018. [Accessed: Dec. 21, 2023].
3. A.H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal, R. Kumar and R. A. Khan, "Healthcare Data Breaches: Insights and Implications," Healthcare, vol. 8, no. 2, pp. 133, 2020.
4. Verizon Company "2018 Data Breach Investigations Report- 11th Editions". [Online]. Available: https://enterprise.verizon.com/resources/reports/DBIR_2018_Report.pdf [Accessed: Dec. 24, 2023].
5. HIPAA Journal "December 2019 Healthcare Data Breach Report". [Online]. Available: https://www. hipaajournal.com/december-2019-healthcare-data-breach-report/ Accessed: Dec. 24, 2023].
6. The Wall Street Journal "Anthem: Hacked Database Included 78.8 Million People". [Online]. Available: https://www.wsj.com/articles/anthem-hacked-database-included-78-8-million-people-1424807364?cb=logged0.7172621067147702 [Accessed: Dec. 24, 2023].
7. T. Kanwal, A. Anjum and A. Khan, "Privacy preservation in e-health cloud: taxonomy, privacy requirements, feasibility analysis, and opportunities," Cluster Computing, pp. 1-25, 2020.
8. B. Tiwari and A. Kumar, "Role-based access control through on-demand classification of electronic health record," International Journal of Electronic Healthcare, vol. 8, no. 1, pp. 9-24, 2015.
9. J. Heurix and T. Neubauer, "Privacy-Preserving Storage and Access of Medical Data through Pseudonymization and Encryption," In International Conference on Trust, Privacy and Security in Digital Business, Berlin, Heidelberg, pp. 186-197, 2011.
10. T. Neubauer and J. Heurix, "A methodology for pseudonymization of medical data," International Journal of Medical Informatics, vol. 80, no. 3, pp. 190-204, 2011.
11. T. Tsegaye and S. Flowerday, "A Clark-Wilson and ANSI role-based access control model," Information and Computer Security, vol. 28, no. 3, pp. 373-395, 2020.
12. S. Chenthara, K. Ahmed, H. Wang and F. Whittaker, "Security and Privacy-Preserving Challenges of e-Health Solutions in Cloud Computing," IEEE Access, vol. 7, pp. 74361- 74382, 2019.
13. P. Vimalachandran, H. Wang and Y. Zhang, "Securing Electronic Medical Record and Electronic Health Records Systems Through an Improved Access Control," In International Conference on Health Information Science, Cham, pp. 17-30, 2015.
14. H. Zhong, Y. Zhou, Q. Zhang, Y. Xu and J. Cui, "An efficient and outsourcing-supported attribute-based access control scheme for edge-enabled smart healthcare," Future Generation Computer Systems, vol. 115, pp. 486-496, 2020.
15. Y. Elmehdwi, B. K. Samanthula and W. Jiang, "Secure k-Nearest Neighbor Query over Encrypted Data in Outsourced Environments," In IEEE 30th International Conference on Data Engineering, pp. 664-675, 2014.
16. M. Li, S. Yu, Y. Zheng, K. Ren and w. Lou, "Scalable and Secure Sharing of Personal records in Cloud Computing Using Attribute Based Encryption," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 1, pp. 131-143, 2012.
17. O. Kocabas and T. Soyata, "Towards Privacy-Preserving Medical Cloud Computing Using Homomorphic Encryption," Virtual and Mobile Healthcare: Breakthroughs in Research and Practice, pp. 93-125, 2020.
18. I. Priyadharshini S and V. Prem M, "Secure e-health cloud framework for patients' EHR storage and sharing for Indian Government healthcare model," In Proceedings of the Estonian Academy of Sciences, vol. 69, no. 3, pp. 266-276, 2020.
19. R. Zhang, R. Xue and L. Liu, "Searchable Encryption for Healthcare Clouds: A Survey," IEEE Transactions on Services Computing, vol. 11, no. 6, pp. 978-996, 2017.
20. V. Vats, L. Zhang, S. Chatterjee, S. Ahmed, E. Enziama and K. Tepe, "A Comparative Analysis of Unsupervised Machine Learning Techniques for Liver Disease Prediction," In IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 303-307, 2018.
21. S. Gupta, V. Bharti and A. Kumar, "A Survey on various Machine Learning Algorithms for Disease Prediction," International Journal of Recent Technology and Engineering(IJRTE), vol. 7, pp. 84-87, 2019.

22. A. Dhillon and A. Singh, "Machine Learning in Healthcare Data Analysis: A Survey," Journal of Biology and Today's World, vol. 8, no. 2, pp. 1-10, 2019.
23. M. A. Myszczynska, P. N. Ojamies, A. M. B. Lacoste, D. Neil, A. Saffari, R. Mead, G. M. Hautbergue, J. D. Holbrook and L. Ferraiuolo, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," Nature reviews Neurology, vol.16, no. 8, pp. 440-456, 2020.
24. S. Bakyarani E, H. Srimathi and M. Bagavandas, "A Survey of Machine Learning Algorithms in Health Care," International Journal of Scientific & Technology Research (IJSTR), vol. 8, no.11, pp. 2288-2292, 2019.
25. Y. J. Chauhan, "Cardiovascular Disease Prediction using Classification Algorithms of Machine Learning," International Journal of Science and Research (IJSR), vol. 9, no. 5, pp. 194-200, 2020.
26. K. Harimoorthy and M. Thangavelu, "Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system," Journal of Ambient Intelligence and Humanized Computing, pp. 1-9, 2020
27. H. Ma, X. Guo, Y. Ping, B. Wang, Y. Yang, Z. Zhang and J. Zhou, "PPCD: Privacy-preserving clinical decision with cloud support," Plos one, vol. 14, no. 5, pp. 1-17, 2019.
28. X. Yang, R. Lu, J. Shao, X. Tang and H. Yang, "An Efficient and Privacy-Preserving Disease Risk Prediction Scheme for E-Healthcare," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 3284-3297, 2018.
29. Y. Rahulamathavan, S. Veluru, R.C.W. Phan, J. A. Chambers and M. Rajarajan, "Privacy-Preserving Clinical Decision Support System Using Gaussian Kernel-Based Classification," IEEE Journal of Biomedical and Health Informatics, vo. 18, no. 1, pp. 56-66, 2013.
30. X. Liu, R. Lu, J. Ma, L. Chen and B. Qin, "Privacy-Preserving Patient-Centric Clinical Decision Support System on Naïve Bayesian Classification," IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 2, pp. 655-668, 2015.
31. G. Wang, R. Lu and C. Huang, "PSLP: Privacy-Preserving Single-Layer Perceptron Learning for e-Healthcare," In IEEE 10th International Conference on Information, Communications and Signal Processing (ICICS), pp. 1-5, 2015.
32. G. Wang, R. Lu and C. Huang, "PGuide: An Efficient and Privacy-Preserving Smartphone-Based Pre-Clinical Guidance Scheme," In IEEE Global Communications Conference (GLOBECOM), pp. 1-6, 2015.
33. P.M. Lavanya and P. Valarmathie, "Big Data in Healthcare Using Cloud Database with Enhanced Privacy," International Journal of Advanced Research in Computer Engineering and Technology (IJARCET), vol. 5, no. 5, 2016.
34. Q. Xue, M. C. Chuah and Y. Chen, "Privacy Preserving Disease Treatment Complication Prediction System (PDTCPS)," In Proceedings of the 11th ACM on Asia Conference on Computer and Communication Security, pp. 841-852, 2016.
35. H. Zhu, X. Liu, R. Lu and H. Li, "Efficient and Privacy-Preserving Online Medical Prediagnosis Framework Using Nonlinear SVM," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 3, pp. 838-850, 2016.
36. J. Lin, J. Niu and H. Li, "PCD: A Privacy-preserving Predictive Clinical Decision Scheme with E-health Big Data Based on RNN," In IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 808-813, 2017.
37. X. Liu, R. H. Deng, K. K. R. Choo and Y. Yang, "Privacy-Preserving Outsourced Clinical Decision Support System in the Cloud," IEEE Transactions on Services Computing, 2017.
38. X. Liu, H. Zhu, R. Lu and H. Li, "Efficient privacy-preserving online medical primary diagnosis scheme on naive bayesian classification," Peer-to-Peer Networking and Applications, vol.11, no.2, pp. 334-347, 2018.
39. C. Zhang, L. Zhu, C. Xu and R. Lu, "PPDP: An efficient and privacy-preserving disease prediction scheme in cloud-based e-Healthcare system," Future Generation Computer Systems, vol. 79, pp. 16-25, 2018.
40. A. Alabdulkarim, M. Al-Rodhaan and Y. Tian, "Privacy-Preserving Healthcare System for Clinical Decision-Support and Emergency Call Systems," Communications and Networks, vol. 9, no. 4, 2017.
41. J. Park and D. H. Lee, "Privacy Preserving k-Nearest Neighbor for Medical Diagnosis in e-Health Cloud," Journal of Healthcare Engineering, 2018.
42. W. Guo, J. Shao, R. Lu, Y. Liu and A. A. Ghorbani, "A Privacy-Preserving Online Medical Prediagnosis Scheme for Cloud Environment," IEEE Access, vol. 6, pp. 48946-48957, 2018.
43. X. Liu, R. H. Deng, Y. Yang, H. N. Tran and S. Zhong, "Hybrid privacy-preserving clinical decision support system in fog-cloud computing," Future Generation Computer Systems, vol. 78, pp. 825-837, 2018.
44. A. Alabdulkarim, M. Al-Rodhaan, Y. Tian and A. Al-Dhelaan, "A Privacy-Preserving Algorithm for Clinical Decision-Support Systems Using Random Forest," CMC- Computers, Materials and Continua, vol. 58, pp. 585-601, 2019.
45. D. Malathi, R. Logesh, V. Subramaniyaswamy, V. Vijayakumar and A. K. Sangaiah, "Hybrid Reasoning-based Privacy-Aware Disease Prediction Support System," Computers and Electrical Engineering, vol. 73, pp. 114-127, 2019.
46. Z. Ma, J. Ma, Y. Miao and X. Liu, "Privacy-preserving and high-accurate outsourced disease predictor on random forest," Information Sciences, vol. 496, pp. 225-241, 2019.

47. A. Alabdulkarim, M. Al-Rodhaan, T. Ma and Y. Tian, "PPSDT: A Novel Privacy-Preserving Single Decision Tree Algorithm for Clinical Decision-Support Systems Using IoT Devices," Sensors, vol. 19, no. 1, pp. 142, 2019.
48. D. Zhu, H. Zhu, X. Liu, H. Li, F. Wang, H. Li and D. Feng, "CREDO: Efficient and privacy-preserving multi-level medical pre-diagnosis based on ML- kNN," Information Sciences, vol. 514, 2020.
49. F. Wang, H. Zhu, R. Lu, Y. Zheng and H. Li, "Achieve Efficient and Privacy-preserving Disease Risk Assessment over Multi-Outsourced Vertical Datasets," IEEE Transactions on Dependable and Secure Computing, 2020.
50. T. Munirathinam, S. Ganapathy and A. Kannan, "Cloud and IoT based privacy preserved e-Healthcare system using secured storage algorithm and deep learning," Journal of Intelligent and Fuzzy Systems, pp. 1-13, 2020.
51. M. Zhang, W. Song and J. Zhang, "A Secure Clinical Diagnosis with Privacy-Preserving Multiclass Support Vector Machine in Clouds," IEEE Systems Journal, 2020.
52. Z. Ma, J. Ma, Y. Miao, X. Liu, K. K. R. Choo, R. Yang and X. Wang, "Lightweight Privacy-Preserving Medical Diagnosis in Edge Computing," IEEE Transactions on Services Computing, vol. 15, no. 3, pp. 1606-1618, 2020.
53. J. Liang, Z. Lin, L. Xue, X. Lin and X. Shen, "Efficient and Privacy-Preserving Decision Tree Classification for Health Monitoring Systems," IEEE Internet of Things Journal, vol. 8, no. 16, pp. 12528-12539, 2021.
54. B. Xie, T. Xiang, X. Liao and J. Wu, "Achieving Privacy-Preserving Online Diagnosis with Outsourced SVM in Internet of Medical Things Environment," IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 6, pp. 4113-4126, 2022.
55. S.P. Gopalan, C.L. Chowdhary, C. Iwndi, M.A.Farid and L. K. Ramasamy, " An Efficient and Privacy-Preserving Scheme for Disease prediction in Modern Healthcare Systems," Sensors, vol. 22, no. 15, pp. 5574, 2022.
56. M. U. Rehman, A. Shafique, K. H. Khan, S. Khalid, A. A. Alotaibi, T. Althobaiti, N. Ramzan, J. Ahmad, S. A. Shah and Q. H. Abbasi, "Novel Privacy Preserving Non-Invasive Sensing-Based Diagnoses of Pneumonia Disease Leveraging Deep Network Model," Sensors, vol. 22, no. 2, pp. 461, 2022.
57. R. Zhao, Y. Xie, X. Jia, H. Wang and N. Kumar, "Practical Privacy Preserving-Aided Disease Diagnosis with Multiclass SVM in an Outsourced Environment," Security and Communication Networks, 2022.
58. Y. Chen, Q. Mao, B. Wang, P. Duan, B. Zhang and Z. Hong, "Privacy-Preserving Multi-Class Support Vector Machine Model on Medical Diagnosis," IEEE journal of Biomedical and Health Informatics, vol. 26, no. 7, pp. 3342-3353, 2022.
59. Y. Zhou, L. Song, Y. Liu, P. Vijayakumar, B. B. Gupta, W. Alhalabi and H. Alsharif, "A privacy-preserving logistic regression-based diagnosis scheme for digital healthcare," Future Generation Computer Systems, vol. 144, pp. 63-73, 2023.
60. G. Shen, Z. Fu, Y. Gui, W. Susilo and M. Zhang, "Efficient and privacy-preserving online diagnosis scheme based on federated learning in e-healthcare system," Information Sciences, vol. 647, pp. 119261, 2023.