International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-9, July 2020

Privacy Preserving Analytics in Outsourced Healthcare System

D.Sudha Devi, S.Sudendar

Abstract*: *The most data intensive industry today is the healthcare system. The advancement in technology has revolutionized the traditional healthcare practices and led to enhanced E-Healthcare System. Modern healthcare systems generate voluminous amount of digital health data. These E-Health data are shared between patients and among groups of physicians and medical technicians for processing. Due to the demand for continuous availability and handling of these massive E-Health data, mostly these data are outsourced to cloud storage. Being cloud-based computing, the sensitive patient data is stored in a third-party server where data analytics are performed, hence more concern about security raises. This paper proposes a secure analytics system which preserves the privacy of patients' data. In this system, before outsourcing, the data are encrypted using Paillier homomorphic encryption which allows computations to be performed over encrypted dataset. Then Decision Tree Machine Learning algorithm is used over this encrypted dataset to build the classifier model. This encrypted model is outsourced to cloud server and the predictions about patient's health status is displayed to the user on request. In this system nowhere the data is decrypted throughout the process which ensures the privacy of patients' sensitive data.

***Keywords* : e-healthcare, homomorphic encryption, decision tree classifier, cloud server, privacy preserving, machine learning.**

I. INTRODUCTION

E-health grew out to keep track of patients' health records in a much comfortable and better way. Today the e-healthcare system generates a massive amount of sensitive data. Due to its voluminous amount these data are being outsourced to manage storage capability. Cloud computing being an essential paradigm today, most of the e-health providers store health data on cloud storage for various purposes. This leads to the necessity of efficiently addressing the security of sensitive data stored in cloud. Since opportunities are there for the attackers and the privileged users on the cloud site to leak the data intentionally or accidently, it is good to encrypt the health data before outsourcing.

Revised Manuscript Received on July 30, 2020.
***** Correspondence Author
Dr. D. Sudha Devi*, Department of Computing(Data Science), Coimbatore Institute of Technology, Coimbatore, India. E-mail:sudhadevi.cit@gmail.com
S. Sudendar, Department of Computing(Data Science), Coimbatore Institute of Technology, Coimbatore, India. E-mail: sudendarsude@gmail.com

To process and analyze, the data has to be decrypted and if the data is decrypted the sensitive information may be revealed. Hence the aim of this paper is to preserve the privacy of health data that are outsourced for analysis. The health service provider must establish a methodology for the users to query about their health status. The challenges for protecting the privacy of the sensitive data need to be addressed effectively. In this paper, we propose a scheme for the health providers to delegate a cloud server to provide the privacy-preserving analysis service for users. In this system, the data is encrypted using Homomorphic encryption before outsourcing which preserves the privacy of patient's health data. And computations and analysis are performed over this encrypted data using the Decision Tree Machine Learning algorithm and prediction results are displayed to the user. We implemented the proposed system and deployed in Amazon EC2 service. The results are shown as screenshots which shows that the system is practical. The rest of the sections are organized as follows: In Section II, related works are presented. Section III briefly explains the preliminary technologies used in this paper. Section IV describes the architecture of the proposed system and how it works. Section V deals with the results and discussions of the implemented system. Section VI draws the conclusions of the proposed system.

II. RELATED WORK

Wu and Haven [1] analyzed the overhead in implementing homomorphic computation for statistical analysis. They have demonstrated using homomorphic encryption to compute on massive datasets. They computed the mean and covariance of multivariate data to perform linear regression over encrypted datasets. Bhuvaneshwari and Kalaiselvi [2] used the Naïve Bayes classifier and back propagation neural network methods to calculate the probability model for better prediction. Nikolaenko et al [3] presented a scheme for ridge regression using garbled circuits and homomorphic encryption algorithm to preserve privacy. Bost et al [4] constructed classification protocols to accomplish privacy constraints. They have designed a new library of building blocks for developing classifiers securely. Mudasir and Syed [5] discussed about data mining techniques in discovering patterns. And they have investigated the results of applying various types of decision trees in diagnosing heart disease of patients. Deepa et al [6] intended to design a diagnostic model for the liver cancer using data mining techniques. They used Random forest and Naive Bayes algorithms for health care diagnosis and analyzed the performances of the classifiers.

Mohammad and Morris [7] analyzed the knowledge gap between privacy concerns and machine learning that need to be bridged. In this article they discussed about various threats that are associated with the sharing data and the possible solutions used to protect the data. In [8], authors proposed a method for the classifier owner to outsource a server to preserve the privacy of classification service. They designed classification protocols for classifiers and implemented the prototype of the scheme. In [9], authors proposed a scheme based on fully homomorphism to preserve privacy and to process e-Health data. Mai et al [10] demonstrated TPA authentication and proposed an auditing scheme to preserve confidentiality and integrity of the data. Chen et al [11] proposed a secure Exclusive OR protocol to preserve privacy of the outsourced classification model. This protocol computes the XOR over encrypted data to produce encrypted result.

III. TECHNOLOGY ASPECTS

A. Machine Learning Decision Tree Classifier

Machine Learning (ML) makes the systems to learn automatically and improve from experience without being programmed explicitly. The learning process starts with observations, data, and experience to find patterns in data and make better decisions in the future. ML allows systems to learn automatically without any human intervention and perform actions accordingly. The most widely adopted machine learning methods are supervised and unsupervised learning. Supervised learning trains algorithms based on input and output data that is labeled. Unsupervised learning provides the algorithm with no labeled data in order to allow it to find pattern within its input data. One of the important applications of machine learning in healthcare is to identify, diagnose, or predict disease based on the trained data.

Decision Tree Classifier a ML algorithm is a flowchart-like tree structure which helps in decision making. In this tree each internal node represents a test on the feature, the branch represents a decision rule, and each leaf node represents the outcome. The topmost node is known as the root node which learns to partition tree on the basis of the attribute value. At each level to identify the attribute for the root node two attribute selection methods namely Information Gain or Gini Index is being used. In the healthcare system decision trees are leveraged mainly in the diagnosis of disease.

B. Paillier Homomorphic Encryption

Homomorphic Encryption (HE) technique helps to perform computation over cipher texts and generates an encrypted result and if decrypted the result resembles the operations as they had been performed on the plaintext. HE can be used where privacy-preserving outsourced computation and storage are required. HE can be used in highly regulated industries, such as health care to overcome privacy problems.

IV. PROPOSED SYSTEM

The proposed system is developed using Python Django framework with MySQL and is deployed in AWS EC2 instance. The health data considered in this paper is the Cleveland Heart Disease dataset obtained from the UCI repository[12]. The attributes in this data set are the details of patients test report related to heart disease such as age, gender, chest pain, resting blood pressure, cholesterol, hereditary, fasting blood sugar, smoking, thal and target result. This dataset is encrypted using paillier homomorphic encryption in order to preserve the privacy of patients' sensitive data. The analysis model is built over this encrypted dataset using Decision Tree Classifier ML algorithm. This analysis model is hosted in cloud. The users can create a valid login through the portal and can give their test report. The user input is also encrypted with paillier encryption and is sent to the model for processing. The model analyses the user input and the prediction result is displayed through the dashboard. Here, users data are nowhere stored in plaintext on the cloud and all computations are performed over encrypted data and model which preserves privacy of patients sensitive information.

A. Paillier Encryption Algorithm

KEY GENERATION

1. Choose two large prime numbers p and q independent such that GCD (p q, (p-1)(q-1))= 1.
2. Compute n= pq and l = LCM (p-1, q-1), where l is lambda.
3. Select a random integer g. Ensure n divides the order of g by checking the existence of the following modular multiplicative inverse: u = (L (g^l mod n^2))^-1 mod n, where function L is defined as L(x)=x-1/n.
4. The public key is (n,g) and the private key is (l,u).

ENCRYPTION

1. Let m be a message to be encrypted where 0 <= m < n.
2. Select random r where 0< r < n and ensure gcd(r,n)=1.
3. Compute cipher text as : c= g^m * r^n mod n^2.

B. Decision Tree Classifier

1. It begins with the original set S as the root node.
2. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy and Information Gain of this attribute.
3. It then selects the attribute which has the smallest Entropy or Largest Information gain.
4. The set S is then split by the selected attribute to produce a subset of the data.
5. The algorithm continues to recur on each subset, considering only attributes never selected before.

C. System Modules

The system has the following modules:

- Registration
- Login
- Data Encryption
- Analysis over encrypted dataset
- Output

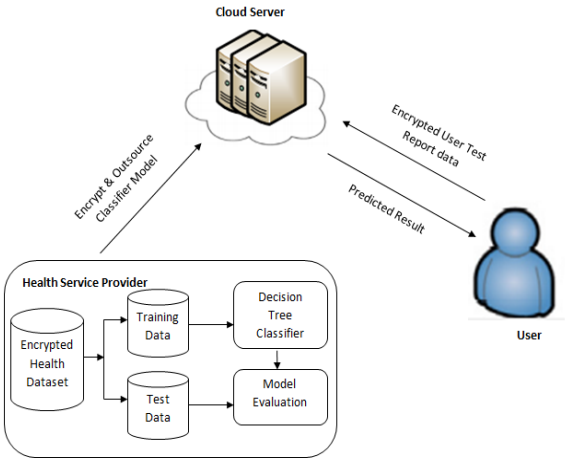

Fig. 1.Proposed System Model

REGISTRATION

This module is used to create login credential for new user. User who wants to check their test report status can create credential and give their test report data to predict the disease status.

LOGIN

This module is used to sign in to the prediction system hosted in cloud. User can enter the valid login credential and can enter into the system to check their disease status.

DATA ENCRYPTION

The heart disease data set is encrypted using Pailier homomorphic encryption algorithm. The encrypted data set is used by the classifier to build the model and prediction is done based on it.

ANALYSIS OVER ENCRYPTED DATASET

This module uses Decision Tree Classifier Algorithm and builds the model based on gini index of the features. This algorithm splits the dataset into training and testing data set and builds the model. Once the model is built, the prediction function receives the test report data in encrypted form and analyses with the classifier model and produces the prediction result which is displayed to the user. Following is the sample data set which is encrypted and used to build the model. A sample of Cleveland Heart Disease dataset from the UCI repository is shown below.

Table- I: Heart Disease sample dataset

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
67	1	0	160	286	0	0	108	1	1.5	1	3	2	0
67	1	0	120	229	0	0	129	1	2.6	1	2	3	0
62	0	0	140	268	0	0	160	0	3.6	0	2	2	0
63	1	0	130	254	0	0	147	0	1.4	1	1	3	0

OUTPUT

The output module displays the disease prediction status to the user through the dashboard.

V. RESULTS AND DISCUSSIONS

The following screenshots depicts the practical implementation of the proposed sytem with results. The heart disease dataset is encrypted using Paillier encryption and the Decision Tree Classifier is used to build the model over this encrypted dataset. This classifier model is hosted in AWS EC2 instance. The patient/user can enter their test report throuth the user interface on the client side. This data is encrypted and is analysed with the classifier model in the cloud server and results are predicted and displayed through the dashboard. Nowhere patient's sensitive data is stored or processed as clear text which implies the privacy preservation of sensitive data in the proposed system.

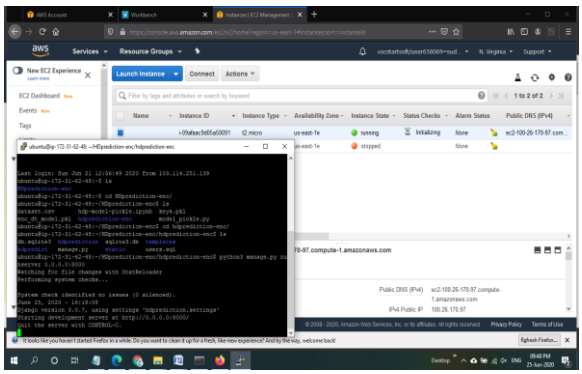

Fig. 2. Connected to AWS EC2 Instance

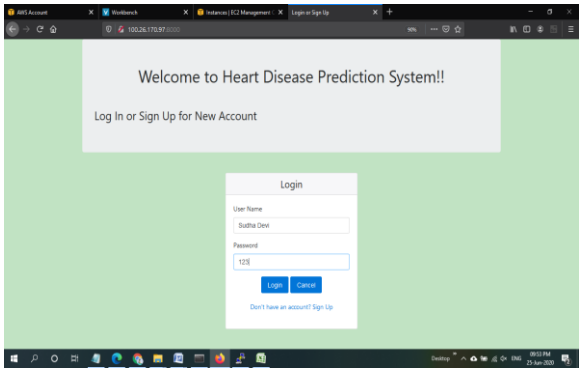

Fig. 3.Django Server running in Browser – Login Page

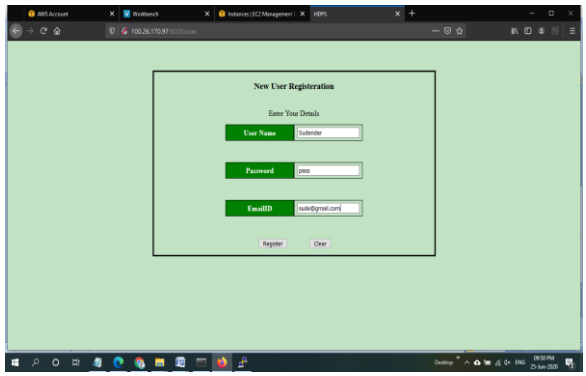

Fig. 4. New User Registration Page

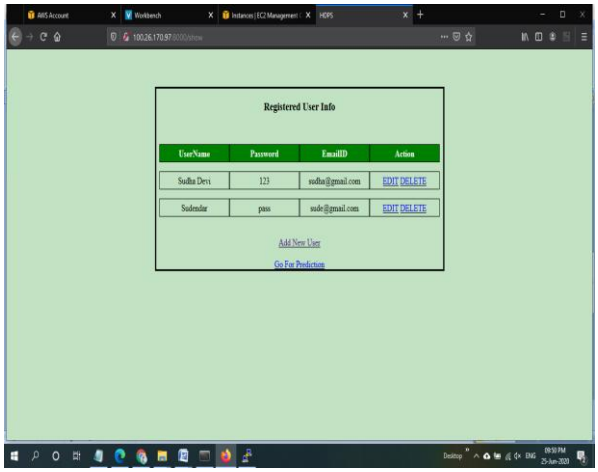

Fig. 5. List of registered users

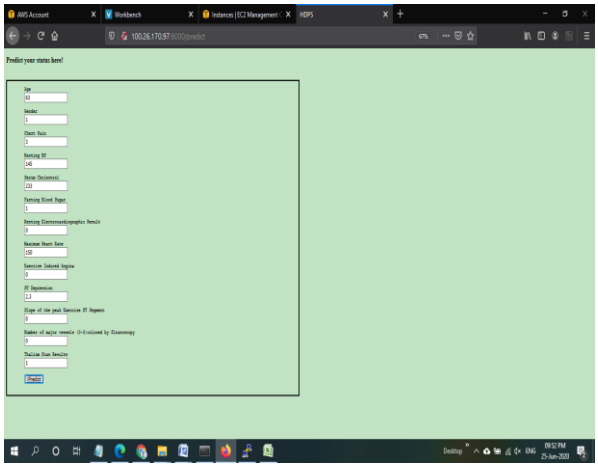

Fig. 6. Patient/User entering test report

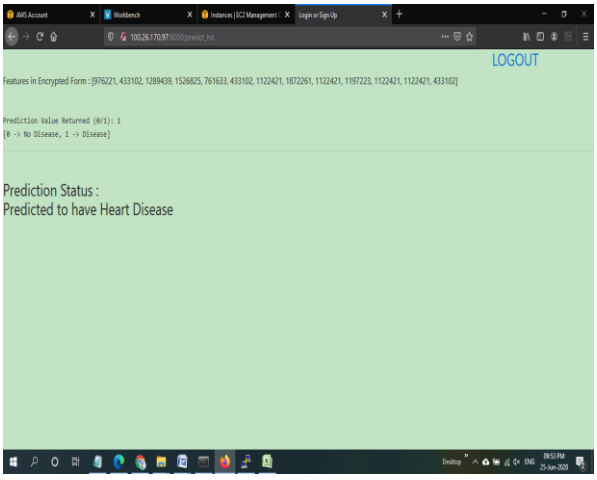

Fig. 7. Encrypted Test report data is analyzed and the predicted result is displayed

The above screenshots shows the working of the proposed system. Test results gives the accuracy percentage of the system as 86%. As the dataset size increases, the proposed model could train still better and even more reliable prediction can be guaranteed.

VI. CONCLUSION

The Secure Health Care Analytics System is implemented and tested. Whenever the test report data is given as input to the system, the data is encrypted and is analyzed with the built model which also uses the encrypted data set. The input data is used by the application to provide reliable prediction regarding the disease status of the patients. Since nowhere the patient details are stored and processed as plain text, the system ensures that the privacy of patient is preserved.

Today, preserving privacy of patient details in a Health Care System is the most important issue which needs to be concentrated. This proposed system tries to train the model using the encrypted data set and tests the data with the same. Hence this system ensures the privacy of patient details in an efficient manner.

REFERENCES

1. D. Wu and J. Haven, “Using homomorphic encryption for large scale statistical analysis”, 2012.
2. R.Bhuvaneshwari, K .Kalaiselvi,”Naïve Bayesian Classification approach In Healthcare Applications”, International Journal of Computer Science and Telecommunication, 2012(vol.3,no.1), pp.106-112.
3. V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, and N. Taft, “Privacy-preserving ridge regression on hundreds of millions of record”, IEEE Computer Society,2013, pp. 334–348.
4. R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, “ Machine learning classification over encrypted data, Ineternet Society, NDSS, 2015.
5. Mudasir Manzoor Kirmani, Syed Immamul Ansarullah, “Prediction of Heart Disease using Decision Tree a Data Mining Technique”, International Journal of Computer Science and Network, 2016(vol.5, no.6), pp. 885-892.
6. Deepa, Karthik Kumar, Dharneshawar, Rohith, “Health Care Analysis Using Random Forest Algorithm”, Journal of Chemical and Pharmaceutical Sciences, 2017(vol.10, no.3)
7. Mohammad Al-Rubaie , J. Morris Chang, "Privacy Preserving Machine Learning: Threats and Solution”, IEEE Security and Privacy Magazine, 2018.
8. Li, T., Huang, Z., Li, P., Liu, Z., Jia, C., “Outsourced privacy-preserving classification service over encrypted data”, Journal of Network and Computer Applications (2018).
9. Xiaoliang Wanga, Liang Bai, Qing Yanga, Liu Wanga and Frank Jianga, “A dual privacy-preservation scheme for cloud-based eHealth systems”, Journal of Information Security and Applications, 2019(vol. 47), pp.132–138.
10. Mai Rady, Tamer Abdelkader and Rasha Ismail, “Integrity and Confidentiality in Cloud Outsourced Data”, Ain Shams Engineering Journal, 2019(vol.10).
11. Chen Wang, Andi Wang, Jian Xu∗ , Qiang Wang, Fucai Zhou , “Outsourced privacy-preserving decision tree classification service over encrypted data”, Journal of Information Security and Applications, 2020(vol.53).
12. Cleveland Heart Disease Data Set http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

AUTHORS PROFILE

Dr. D. Sudha Devi is currently working as Associate Professor in the department of Computing - Data Science at Coimbatore Institute of Technology, Coimbatore, India. She has 16 years of teaching experience and her primary research interests include Data security in Cloud computing, Big Data Computing, web security and Privacy preserving Machine Learning. She has published several papers in the area of cloud computing and data security in various National and International Journals and Conferences.

S.Sudendar is pursuing M.Sc(Data Science) at Coimbatore Institute of Technology. He has carried out projects on various technologies and domains like Python, Flask, Django, Tensorflow, ReactJS, Flutter/Dart, Blockchain, Malware Analysis and Security. He has published papers in the area of Data Analytics and Blockchain in National and International Conferences.