International Journal of Multidisciplinary Research and Growth Evaluation.

Secure multiple disease diagnosis system using machine learning in health record management

Malarvizhi A [1*], Reshma S [2], Shanthiya M [3], B Ranjani [4]
[1-3] Computer Science and Engineering, EGS Pillay Engineering College, Nagapattinam, Tamil Nadu, India
[4] Ph.D., Computer Science and Engineering, EGS Pillay Engineering College, Nagapattinam, Tamil Nadu, India

* Corresponding Author: **Malarvizhi A**

Article Info

ISSN (online): 2582-7138
Volume: 04
Issue: 03
May-June 2023
Received: 26-03-2023;
Accepted: 19-04-2023
Page No: 70-74

Abstract
Disease risk assessment system has great eventuality to alleviate the medical treatment problems for the unborn smart megacity and communities, as it can excavate disease risk factors from a large number of case features, give diagnostic references for doctors, and save medical treatment time for cases. The flourish of disease risk assessment service still faces severe challenges including information privacy and security. Naive Bayesian classification techniques have taken over the task of prediction of disease risk assessment scheme over multi sourced vertical datasets, named CARER. With CARER, the e-healthcare provider can securely train a disease risk predication model over vertically distributed medical data from multiple medical centers and provide privacy-preserving disease risk predication services for users. During the model training and disease risk prediction phases, all sensitive data are operated over cipher texts. Finally, the private information of medical centers, e-healthcare provider, and users can be well protected. This security analysis shows that CARER can resist various known security threats. In addition, we evaluate the performance of CARER with real medical datasets, and the experimental results demonstrate that CARER is efficient and it improves prediction accuracy, privacy, and security compared to the existing methods.

Keywords: Disease risk assessment, privacy-preserving, secure data training, naive bayesian classification

1. Introduction

Under big data-driven society, a large amount of data is already being collected in an e-healthcare system to generate insights on disease prediction and to improve patient care [1, 2]. As one of the most popular applications in e- healthcare, online disease risk assessment is revolutioning traditional medical, as it can detect a risk condition before it becomes an illness or disorder, and the cost of intervention is far less than the eventual cost of treatment [3, 4]. It consists of two phases, i.e., model training and disease risk prediction. In the model training phase, the e-healthcare provider collects and aggregates local medical data from multiple medical centers, and trains a disease risk prediction model based on machine learning algorithms [5, 6]. While in the disease risk prediction phase, the e-healthcare provider can offer online disease risk prediction services for users with the trained prediction model, which will significantly improve the medical treatment efficiency and the quality of people's life.

Unfortunately, owing to the sensitivity of medical information, the flourish of online disease risk assessment is still confronted with severe hassles including data privacy and security [7, 8, 9]. Firstly, local medical data generally contain massive patients' treatment records and statistical data of medical centers, which may disclose patients individual information and medical centers clinical treatment programs when outsourcing them to the e- healthcare provider. Secondly, the trained disease risk prediction model is commonly regarded as valuable business assets. Leakage of the prediction model may directly result in an economic loss of the e-healthcare provider. Thirdly, users' disease risk query requests and corresponding query results are also high sensitive, since they may reveal users' private information, such as health conditions, illness, and medication situation during the disease risk prediction.

Therefore, in online disease risk assessment, these sensitive data of medical centers, e- healthcare provider and users cannot be leaked to each other.

To address the above-mentioned challenges, plenty of privacy-preserving medical data processing schemes have been proposed, which mainly rely on homomorphic encryption [10, 11] and secure multi-party computation (SMC) technique [12, 13]. Specifically, homomorphic encryption supports arithmetical operations over ciphertexts, which can well protect sensitive medical data during disease risk assessment. However, most homomorphic encryption based schemes bring heavy computation overhead since they contain massive time-consuming operations such as bilinear pairing. SMC-based schemes can achieve privacy preserving model training or disease risk prediction, but most of them require multiple interactions to complete a specific operation over ciphertexts, which will bring massive extra communication overhead in distributed scenarios. Moreover, few of existing privacy-preserving medical data processing schemes work on vertically distributed datasets.

Naive Bayesian classification technique have taken over the task of prediction of disease risk assessment scheme over multi-sourced vertical datasets, named CARER

.With CARER, the e-healthcare provider can securely train a disease risk predication model over vertically distributed medical data from multiple medical centers, and provide privacy-preserving disease risk prediction services for users.

2. Related Works

Afshar *et al*. [14] did not compare the performance of capsule network with other methods. Therefore, the contributions of this report include Performance comparison of three different classification methods: SVM classifier with oriented fast and rotated binary robust independent elementary features (ORB), transfer learning of VGG16 and InceptionV3, and training capsule network from scratch. An analysis of the effects of data augmentation, network complexity, fine-tuned convolutional layer, and other preventing over fitting mechanics on the classification of small chest X-ray dataset by transfer learning of CNN.

Li *et al*. [15] introduced a privacy-preserving outsourced classification framework based on fully homomorphic encryption, where the evaluator and crypto service provider can jointly train a naive bayesian classification model over multi-party outsourced encrypted data. Based on additive homomorphic encryption, Mandal *et al*.

Based on additive homomorphic encryption, Mandal *et al*. [16] designed a method to securely execute gradient descent for data owners and the cloud server, and further achieve the privacy-preserving linear and logistic regression model training. Gasc on *et al*.

Haq *et al*. [17] have developed a machine-learning- based diagnosis system for heart disease prediction by using heart disease dataset. They used seven popular machine learning algorithms, three feature selection algorithms, the cross-validation method, and seven classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity, Matthews’ correlation coefficient, and execution time. The proposed system can easily identify and classify people with heart disease from healthy people.

Zhou *et al*. [18] proposed a novel secure data processing protocol, which supports both homomorphic addition and multiplication operations over ciphertexts. Based on the proposed protocol, an efficient and privacy- preserving dynamic medical text mining and image feature extraction scheme was proposed. Most above-mentioned schemes only achieve the privacy-preserving data training. Besides, massive interactions are necessary between data providers and cloud servers, which bring heavy communication overhead in practice.

3. Working Methodology

The system consists of four parts: 1) trusted authority (TA); 2) medical centers (MCs); 3) e-healthcare provider (EP) and 4) users. TA is a trusted authority (i.e., a government organization), which bootstraps the whole system through generating system parameters, and distributing keys for medical centers, e-healthcare provider, and users. EP is the e-healthcare provider, which is an online healthcare organization offering disease risk assessment. EP is responsible for aggregating the encrypted local training data from multiple medical center-s, training the disease risk prediction model, and offering privacy- preserving disease risk prediction services for users user can compute her/his encrypted disease risk query request, and further access disease risk prediction services from EP.

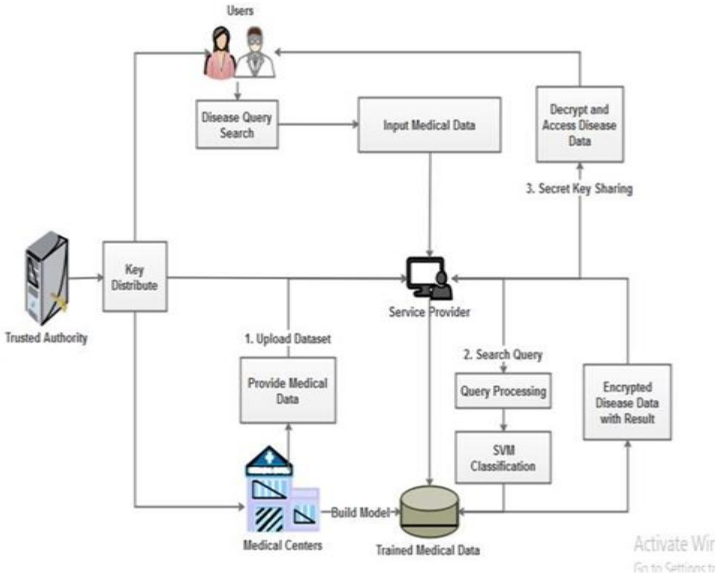


Fig 1: Outline of Proposed System

Implement the disease risk prediction system that includes the disease model training and remote disease prediction. Specifically, in the disease model training, we use the historical medical data collected from confirmed patients to train using Support Vector Machine algorithm Providing privacy and confidentiality of data using encryption techniques.

4. Implementation

A. Support Vector Machine (SVM)

SVM means supervised machine learning algorithm which is very useful technique for data classification. However, this learning algorithm can also be used for regression challenges. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e. the class labels) and several attributes.

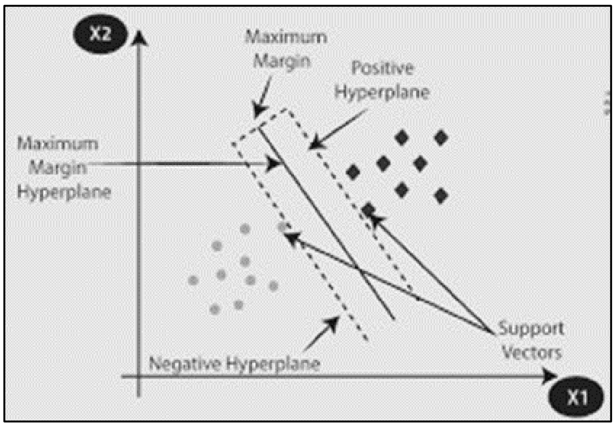


Fig 2: Concept of SVM

SVM works well on small data sets but is more efficient with large data sets. Given a dataset with n features, SVM initiates with plotting all points in the dataset in n- dimensional space, and each point is assigned a coordinate classification process is conducted by determining a suitable hyper plane which to the furthest extent, differentiates the points into two distinct classes. Support vectors are essentially the points that are located close to the hyper plane and determine its position and orientation. The distance between the support vectors and the hyper plane is called the margin and to generate the most accurate hyper plane, the margin needs to be maximized as far as possible.

The main advantages of the SVM algorithm are, primarily it is very effective in analysing high dimensional datasets. It is of great use in cases where the number of dimensions is greater than the number of samples. SVM utilizes the support vectors for training and therefore consumes less memory.

The disadvantages of the SVM algorithm are, it is not suitable for very large datasets as the time required to train the model increases. It also gives inaccuracies when the target classes overlap with each other. Moreover, the SVM algorithm cannot account for probability. SVM is used to classify agricultural data to allow for better decision-making. In a comparative study of classification techniques used for agricultural data, SVM was able to outperform Naïve Bayes and Artificial Neural Network methods.

Algorithm

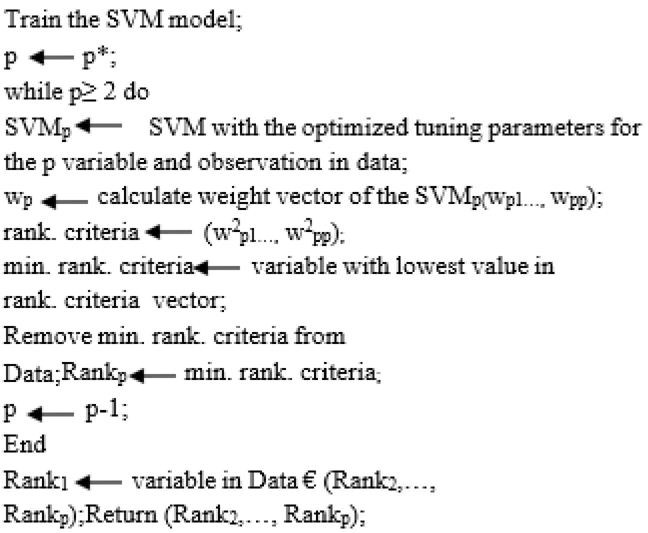
Step 1: Load the important libraries.

Step 2: Import dataset and extract the X variables and Y separately.

Step 3: Divide the dataset into train and test. Step 4: Initializing the SVM classifier model. Step 5: Fitting the SVM classifier model.

Step 6: Coming up with predictions.

Data: Dataset with p* variables and binary outcome. Output: Ranked list of variables according to their relevance.



```
Train the SVM model;
p ← p*;
while p≥ 2 do
SVM$_p$ ←   SVM with the optimized tuning parameters for
the p variable and observation in data;
w$_p$ ← calculate weight vector of the SVM$_{p(wp1..., wpp)}$;
rank. criteria ← (w$^2$$_{p1}$..., w$^2$$_{pp}$);
min. rank. criteria← variable with lowest value in
rank. criteria  vector;
Remove min. rank. criteria from
Data;Rank$_p$← min. rank. criteria;
p ← p-1;
End
Rank$_1$ ← variable in Data € (Rank$_2$,…,
Rank$_p$);Return (Rank$_2$,…, Rank$_p$);
```

The reasons we use SVMs in machine learning is,it can handle both classification and regression on linear and non-linear data. Another reason we use SVMs is because they can find complex relationships between your data without you needing to do a lot of transformations on your own. It's a great option when you are working with smaller datasets that have tens to hundreds of thousands of features. They typically find more accurate results when compared to other algorithms because of their ability to handle small, complex datasets.

B. Naive Bayes Classifier Algorithm

It is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute in dependently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

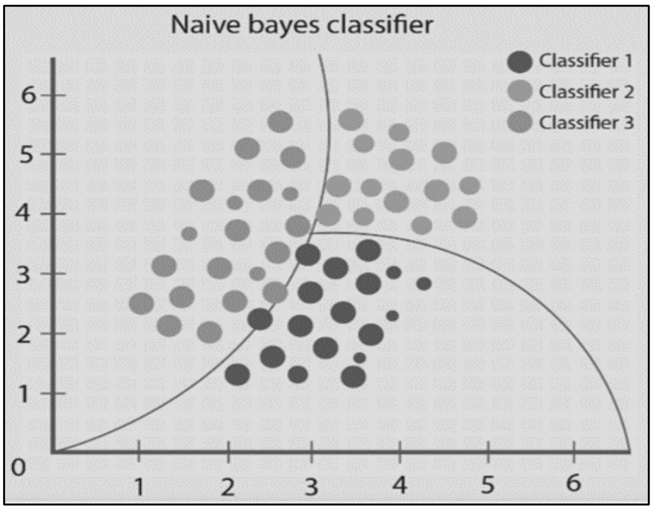


Fig 3: Naive Bayes classifier

The advantages are, it is easy and fast to predict class of test data set. It also perform well in multi class prediction When assumption of independence holds, the classifier performs better compared to other machine learning algorithm like logistic regression or decision tree, and requires less training data. It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed.

C. AES Algorithm

It is also known as the block cipher. No successful attack has been reported on AES. Some advantages of AES are easy to implement on 8-bit architecture processors and effective implementation on 32-bit architecture processors. This algorithm is based on a substitution-permutation network, also known as an SP network. It consists of a series of linked operations, including replacing inputs with specific outputs (substitutions) and others involving bit shuffling.

Algorithm

Step 1: Derive the set of round keys from the cipher key.
Step 2: Initialize the state array with the block data (plaintext).
Step 3: Add the initial round key to the starting state array.
Step 4: Perform nine rounds of state manipulation.
Step 5: Perform the tenth and final round of state manipulation.
Step 6: Copy the final state array out as the encrypted data (cipher text).

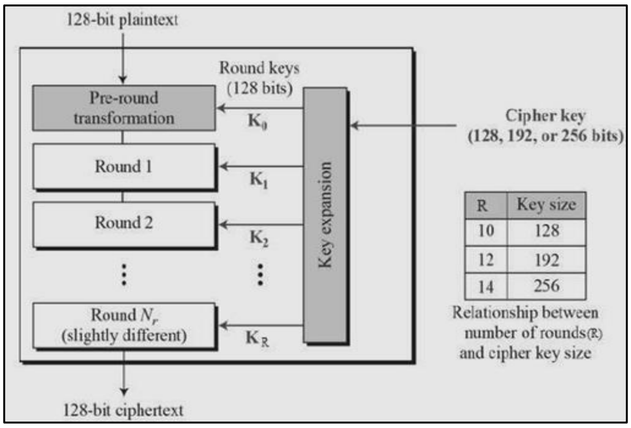


Fig 3: Block diagram for AES algorithm

The AES algorithm uses a substitution- permutation, or SP network, with multiple rounds to produce cipher text. The number of rounds depends on the key size being used. A 128-bit key size dictates ten rounds, a 192-bit key size dictates 12 rounds, and a 256- bit key size has 14 rounds. Each of these rounds requires a round key, but since only one key is inputted into the algorithm, this key needs to be expanded to get keys for each round, including round 0.

5. Result Analysis

The major part behind this project is the data classification which classifies the data as normal or disease-affected with inflexibility. At the same time, security is pivotal for securely transferring the data. In python, this system is executed. The trials were done by exercising the medical datasets to examine the proposed work concerning the parameters, namely specificity, accuracy, perfection, and sensitivity. Likewise, the proposed system's security is computed regarding security position analysis, decryption time, and encryption time. Utilizing the SVM classification algorithm, various trails are done on classifying the disease affected. The experimental outgrowth is also tested to estimate the proposed technique's performance.

6. Conclusion

The proposed solution for efficient and privacy preserving disease risk assessment scheme over multioutsourced vertical datasets, called CARER. Using a modified Paillier cryptosystem and random masking techniques, in CARER, EP can securely train a disease risk prediction model over vertically distributed medical data from multiple medical centers, and provide privacy- preserving disease risk prediction services. In this process, all sensitive data of medical centers, e- healthcare provider, and users were well protected. This system greatly improved the efficiency of privacy- preserving disease risk assessment through data preprocessing and operation transformation.

7. References

1. G Manogaran, R Varatharajan, D Lopez, PM Kumar, R Sundarasekar, C Thota. A new architecture of internet of things and big data ecosystem for secured smart healthcare monitoring and alerting system, Future Generation Computeer Systems. 2018; 82:375-387.
2. J Qiu, X Liang, S Shetty, D Bowden. Towards secure and smart healthcare in smart cities using block chain, in IEEE International Smart Cities Conference. IEEE, 2018, 1-4.
3. JS Lin, CV Evans, E Johnson, N Redmond, EL Coppola, N Smith. Nontraditional risk factors in cardiovascular disease risk assessment: Updated evidence report and systematic review for the us preventive servicestask force, Journal of the American Medical Association. 2018; 320(3):281-297.
4. A Abbas, M Ali, MUS Khan, SU Khan. Personalized healthcare cloud services for disease risk assessment and wellness management using social media, Pervasive and Mobile Computing. 2016; 28:81-99.
5. S Perveen, M Shahbaz, K Keshavjee, A Guergachi. A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression, Scientific reports. 2018; 8(1):1-12.
6. L Jena, S Nayak, R Swain. Chronic disease risk (cdr) prediction in biomedical data using machine learning

approach, in Advances in Intelligent Computing and Communication, 2020, 232-239.

7. C Xu, N Wang, L Zhu, K Sharif, C Zhang. Achieving searchable and privacy-preserving data sharing for cloud-assisted e- healthcare system, IEEE Internet of Things Journal. 2019; 6(5):8345-8356.
8. W Tang, J Ren, K Deng, Y Zhang. Secure data aggregation of lightweight e-healthcare iot devices with fair incentives, IEEE Internet of Things Journal. 2019; 6(5):8714–8726.
9. L Yang, Q Zheng, X Fan. RSPP: A reliable, searchable and privacy-preserving e-healthcare system for cloud-assisted body area networks, in 2017 IEEE Conference on Computer Communications. IEEE, 2017, 1-9.
10. R Bocu, C Costache. A homomorphic encryption-based system for securely managing personal health metrics data," IBM Journal of Research and Development. 2018; 62(1-1):1–1:10.
11. X Liu, R Lu, J Ma, L Chen, B Qin. Privacy-preserving patient-centric clinical decision support system on na¨ıve bayesian classification," IEEE Joutnal of Biomedical and Health Informatics. 2016; 20(2):655-668.
12. KY Yigzaw, JG Bellika. Evaluation of secure multi-party omputation for reuse of distributed electronic health data, in Proceedings of IEEE-EMBS International Conference on Biomedical and Health Informatics. IEEE, 2014, 219-222.
13. D Zhu, H Zhu, X Liu, H Li, F Wang, H Li D Feng. CREDO: Efficient and privacy-preserving multi-level medical pre- diagnosis based on ml-KNN," Information Sciences, 2016, 5-14.
14. Afshar P, Mohammadi A, Plataniotis KN. Brain tumor type classification via capsule networks. In: 25th IEEE international conference on image processing (ICIP). IEEE, 2018, 3129-33.P. Li,
15. J Li, Z Huang, C Gao, W Chen, K Chen. Privacy preserving outsourced classification in cloud computing, Cluster Computing. 2018; 21(1):277-286.
16. K Mandal, G Gong. Privfl: Practical privacy-preserving federated regressions on high-dimensional data over mobile networks, in Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop. ACM, 2019, 57-68.
17. Haq Amin, Li Jianping, Memon Muhammad, Memon Muhammad, Khan Jalaluddin, Marium Syeda. Heart Disease Prediction System Using Model of Machine Learning and Sequential Backward Selection Algorithm for Features Selection, 2019, 1-4. 10.1109/I2CT45611.2019.9033683.
18. J Zhou, Z Cao, X Dong, X Lin. PPDM: A privacy preserving protocol for cloud-assisted e-healthcare systems, IEEE Journal of Selected Topics in Signal Processing. 2015; 9(7):1332-1344.
19. J Zhou, Z Cao, X Dong, X Lin. PPDM: A privacy preserving protocol for cloud-assisted e-healthcare systems, IEEE Journal of Selected Topics in Signal Processing. 2015; 9(7):1332-1344.