IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SECURE THE DRUG COMPONENTS USING NAÏVE BAYES AND SVM

Lavanya T[1] , Bhagya sree C R[2] , Vinmathi M S[3] , Dr Kavitha Subramani[4]

[1,2]Final year students, [3,4]Assistant Professor,

[1,2,3,4]Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India

Abstract: In this paper, we propose a framework to secure drug components in the cloud. Specifically, we design for multiple drug formula providers' to use the cloud securely. In our approach, the analyzer trains the drug formulas using Support Vector Machine (SVM) and naïve Bayes. To perform integer and fraction computations in the cloud server, we designed secure computation protocols. We securely train the SVM to privately refresh the selected SVM parameters using the two protocols, which are SVM parameter selection protocol and sequential minimal optimization protocol. We train NB based on Bayes theorem with an assumption of independence among predictors. To determine whether a drug compound is active or inactive in a cloud, the trained SVM and NB classifier is used. Lastly, we prove that the proposed framework achieves the goal that facilitate drug manufacturers to securely outsource their formulas without privacy leakage to unauthorized parties in the cloud for storage and for SVM and NB training.

Index Terms - Securing drug components, Naïve Bayes, SVM, training datasets

I. INTRODUCTION

The initial phase in the drug discovery process is to find a proper 'drug gable' target, which is either a protein receptor or biomolecule. Once the target is found, the second phase is the validation and confirmation of the target. The next phase involves identifying lead compound of a drug followed by testing the target compound. The library of compounds are screened to identify lead compound in various methods. The various methods are screening high-throughput, isolating the natural products etc. After drug discovery, stringent testing and optimization techniques identify process, in the drug development phase the effectiveness of the drug. To study the properties of the lead compounds, it is tested in cells and in animals. To consider lead candidate successful, it should be non-toxic, absorbed, distributed, metabolized and excreted from the body. The result of a development phase concludes if the drug candidate is best for treatment of disease. All these phases marks an end by submitting the drug components to the specific regulatory authority. There may be a leakage of drug information in this phase. Due to the significant investments and high commercial values involved in drug discovery, privacy is an important factor .To secure the drug components, a data mining tools is used.

Of the data mining tools, Support Vector Machine (SVM) has a relatively high decision rate and has been widely used in recent times to predict ligand-based chemical compounds in drug discovery. In approaches using SVMs,known drug formulas datasets is used to train the SVM classifier, and the new drug compound visual scanning is done by trained SVM.As privacy is prime, how can we minimize the risk of unauthorized disclosure during the SVM training phase? In this context, when a researcher sends some chemical compounds to the cloud for SVM classification, it is important to ensure that the potential new drug compounds will not be leaked to a third party, such as a competing pharmaceutical corporation. Furthermore, to train the SVM, multiple pharmaceutical corporations may collaborate in order to increase the SVM decision rate without revealing their datasets. How to achieve secure SVM training and decision under multiple data sources without compromising, the privacy of each individual party remains a research and operational challenge. Thus, in this paper, we propose secure drug components using SVM and Naïve Bayes for Securing Drug discovery in the cloud environment. Unlike existing drug discovery frameworks, our framework seeks to achieve the following

- Secure Outsourced Data Storage: The drug formula owner can securely outsource the data (e.g. drug formula) to the cloud for storage without leaking the data to unauthorized third parties.
- Secure Multi-Source SVM Training: The POD allows an authorized model provider to use other drug formula owners' encrypted data to train the SVM on the fly. The model provider can decrypt and obtain the trained model without knowing (contents of) the training dataset.
- Secure SVM Drug Decision: An authorized tester can securely upload his/her drug chemical compounds to the cloud and determine whether the compound is active or not in a privacy-preserving way.

Commercialization is the last phase of drug development process. The drug is either marketed or commercialized when it is approved. The drug manufacturers need to submit marketing authorization applications in every country in which they want to sell the drug. As the drug is typically targeted to a very large number of patients, the manufacturer is expected to monitor this stage cautiously and submit reports to the FDA. The reports include evidence for medicine-related problems, e.g., treatment failure, adverse reaction, counterfeit/poor quality medicines, drug interactions, or incorrect use. These reports are significant in terms of generating proof of efficacy that will inspire public confidence and trust.

II. EXISTING SYSTEM:

In the existing system, it is proposed to secure drug components in the cloud. Specifically, it is designed for multiple drug formula providers' to use the cloud securely. In this approach, the analyzer trains the drug formulas using Support Vector Machine (SVM). To perform integer and fraction computations in the cloud server, secure computation protocols is designed. The SVM is trained securely to privately refresh the selected SVM parameters using the two protocols, which are SVM parameter selection protocol and sequential minimal optimization protocol. To determine whether a drug compound is active or inactive in the cloud, the trained SVM classifier is used. The existing datasets of known drug formulas to train the SVM classifier, and the trained SVM classifier can be used for new drug compound visual scanning. Due to the significant investments and high commercial values involved in drug discovery, privacy is an important factor. When a researcher sends some chemical compounds to the cloud for SVM classification, it is important to ensure that the potential new drug compounds will not be leaked to a third party, such as a competing pharmaceutical corporation.

III.PROPESED SYSTEM:

We propose secure drug discovery components in the cloud environment. Unlike existing drug discovery frameworks, our POD seeks to achieve it efficiently. We are not using three real time datasets to check the efficiency of potential new drug component. Instead of using existing datasets, we are using another one data-mining algorithm Naïve Bayes (NB), which is based on Bayes' Theorem .Bayes' Theorem, is a classification technique, which considers independent predictor's assumption. NB classifier considers both the presence of a particular feature and any other feature of a class, which are unrelated. It performs well in multi class prediction since it predicts class of test data sets quickly. These two algorithms such as SVM and NB are used to train the uploaded drug dataset (CSV file).Thus we will get trained data and accuracy for that uploaded dataset. The trained data and accuracy will be sent to the owner from python server. Drug tester will check that new drug component. To check the new drug tester has to send request for accessing the drug component. As privacy is an important factor due to the significant investments and high commercial values involved in drug discovery, even drug tester does not know the contents of that file; they will get the trained data only. Once the Testing is completed, the tester sends result to the admin.If that particular drug component is retained in the cloud then it is assumed that component is still active and passed the testing successfully. If not, then the drug component is removed from the cloud. Finally, admin will approve the drug component.

IV.NAÏVE BAYES:

The simplest solutions are usually the foremost powerful ones, and Naïve Bayes may well be a model of that. Despite the advances in Machine Learning within the past few years, it is well tried to not only be straightforward but also in addition fast, accurate, and reliable. It has been with success used for many functions, but it works notably well with Natural language processing (NLP) issues. Naïve Bayes may well be a probabilistic machine learning formula supported by the Bayes Theorem, utilized in an exceedingly properness of classification tasks. Bayes' Theorem is a classification technique that considers independence among predictors' assumption. Naive mathematician model is simple to make and notably useful for extraordinarily big info sets. Beside simplicity, Naive mathematician is known to defeat even very refined classification ways.

Naïve Bayes uses the formula to predict the accuracy

Likelihood Class Prior Probability

$$P(c\mid x)=\frac{P(x\mid c)P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c\mid \mathrm{X}) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c)$$

Above,

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

V.SUPPORT VECTOR MACHINE:

Support Vector Machine or SVM is one in every of the foremost well-liked supervised Learning algorithms, that is used for Classification equally as Regression issues. However, it is used for clearing Classification issues in Machine Learning.
The SVM formula's goal is to create the foremost effective line or decision boundary that is in a position to segregate n-dimensional house into categories so we will simply place the new information within the correct category among the long-standing run. This best decision boundary is termed as a hyperplane.
The hyperplane is formed with the assistance of acute points/vectors chosen by SVM. These extreme cases are noted as support vectors, thus this algorithmic program is termed as Support Vector Machine or SVM.

VI.ALGORITHM USED:

Instead of using existing datasets, we are using another one data-mining algorithm Naïve Bayes (NB). It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. NB classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature also fast to predict class of test data set and performs well in multi class prediction. These two algorithms such as SVM and NB are used to train the uploaded drug dataset (CSV file).Thus we will get trained data and accuracy for that uploaded dataset.

VII.SYSTEM DESCRIPTION:

This system secure drug components for Secure Drug discovery in the cloud environment. Unlike drug discovery frameworks, the secure drug discovery seeks to achieve it efficiently. We are not using three real time datasets to check the efficiency of potential new drug component. Instead of using existing datasets, we are using another one data-mining algorithm Naïve Bayes (NB). These two algorithms are used to train the uploaded drug dataset (CSV file). In final, we will get trained data and accuracy for that uploaded dataset. Drug tester will check that new drug component. Drug tester does not know the contents of that file; they will get the trained data only. Then they let us know the file was active or not. Finally, admin will approve the drug component.

LIST OF MODULES:

- Drug Owner & Tester Registration
- Drug Component Uploading
- Train dataset
- Drug Testing

DRUG OWNER AND TESTER REGISTRATION:

Drug Owner will register in the service provider platform. MySQL database is used to store the drugs Meta details. This drug owner registration process will involves few entities such as 1.Name(drug owner's)2.Email(drug owner's)3.Contact number(drug owner's)4.lab name(where the drugs has been discovered)5.lab code(unique number of that lab) for this process.
The drug tester will also be registered in the drug tester registration platform. MySQL database is used to store the drugs Meta details. This drug tester registration process will involves few entities such as 1.Name(drug tester's)2.Email(drug tester's)3.Contact number(drug tester's) and 4.TesterID(unique number of that tester) for this process.

DRUG COMPONENT UPLOADING:

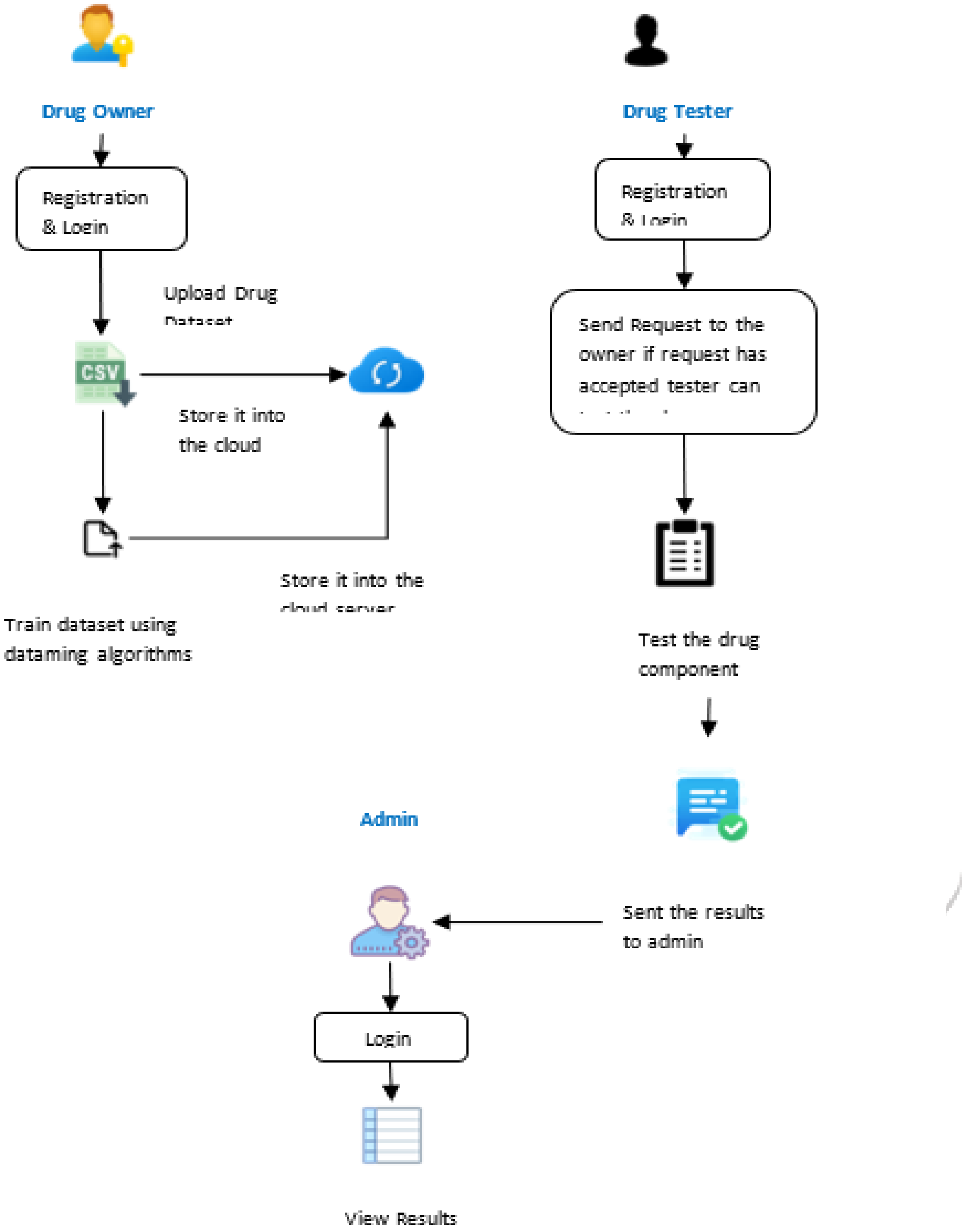
Once the Registration is completed, the drug owner should upload the drug component. For uploading the drug component, the owner must provide the Drug Name, Drug Id and Date of Uploading. Now we have to choose the file that contains the drug data sets and drug components. That data set contains the formula and we have to mention the type of class (Class A, Class B). While uploading the file we will read the content and store into the database and store that .csv file in cloud.

TRAIN DATASET:

As the drug owner has successfully uploaded the drug components, now we will train the uploaded data using python. For this part, we will use two algorithms, SVM and Naïve Bayes. The trained data and accuracy will be sent to the owner from python server.

DRUG TESTING:

As the drug dataset has been trained and uploaded in the cloud, the drug tester now send request to the owner for testing. Only after accepting the request, the test can test the uploaded drug components. Once the Testing is completed, the drug tester sends result to the admin.If that particular drug component is retained in the cloud then it is assumed that component is still active and passed the testing successfully. If not, then the drug component is removed from the cloud.


Drug Owner
Registration & Login
Upload Drug Dataset
CSV
Store it into the cloud
Store it into the cloud server
Train dataset using dataming algorithms
Drug Tester
Registration & Login
Send Request to the owner if request has accepted tester can
Test the drug component
Admin
Sent the results to admin
Login
View Results

VIII. CONCLUSION AND FUTURE ENHANCEMENT:

This project focuses on securing the drug components in the cloud. Privacy is a major factor in drug discovery as it involves significant investment and high commercial values. Drug discovery is a long-term expensive process. Bringing the drugs from the bench to the market involves a lot of threat when stored in the cloud. Thus, we proposed securing the drug components for drug discovery in the cloud. Securing the drug components is designed to facilitate drug manufacturers to securely outsource their formulas to the cloud for storage and SVM and NB training. These two algorithms such as SVM and NB are used to train the uploaded drug dataset (CSV file). As a result, we receive trained data and accuracy for that uploaded dataset. The trained model could be used for authorized client’s compound classification in a privacy-preserving way. Specifically, we designed a secure domain transformation protocol and several basic secure computation components for secure outsourced computation across different parties. We also built two key secure components (i.e. secure parameter selection and secure sequential minimal optimization) to achieve privacy-preserving SVM and NB training in drug discovery. We will be extending our approach to support more sophisticated data mining method in order to support very large dataset in drug discovery.

IX. References

[1] Dibyendu Dana , Satishkumar V. Gadhiya , Luce G. St. Surin , David Li , Farha Naaz ,Quaisar Ali,"Deep Learning in Drug Discovery and Medicine; Scratching the Surface", The British journal of pharmacology, vol. 162, no. 6, pp. 1239–1249, 2018

[2] Yaman Akbulut 1 ID , Abdulkadir ¸ Sengür 1,* ID , Yanhui Guo 2 and Florentin Smarandache, "A Novel Neutrosophic Weighted Extreme Learning Machine for Imbalanced Data Set ",Journal of chemical information and computer sciences, vol. 44, no. 5, pp. 1630–1638, 2017

[3] Richard C. Mohsa, Nigel H. Greig, "Drug discovery and development: Role of basic biological research," IEEE Journal of Biomedical and Health Informatics, vol. 20, pp. 655 – 668, 2017

[4] Aniket Sharma, "Needle Free Drug Delivery Devices Market Size, Industry Outlook and Opportunity Analysis Report 2018-2025"

[5] Shu- Feng Zhou, Wei-Zhu Zhong," Drug Design and Discovery: Principles and Applications", Drug discovery today, vol. 17, no. 19, pp. 1088–1102, 2017.

[6] 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," Biomedical and Health Informatics, IEEE Journal of, vol.18, no. 1, Pp.56-66, 2015.

[7] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine learning classification over encrypted data," in 22nd Annual Network and Distributed System Security Symposium, NDSS 2015, San Diego, California, USA, February 8-11, 2015.

[8] Y. Rahulamathavan, R. C.-W. Phan, S. Veluru, K. Cumanan, and M. Rajarajan, "Privacy-preserving multi-class support vector machine for outsourcing the data classification in cloud," IEEE on Dependable and Secure Computing, vol. 11, no. 5, pp. 467–479, 2014.

[9] J. B. Mitchell, "Machine learning methods in chemoinformatics," Wiley Interdisciplinary Reviews: Computational Molecular Science, . 4, no. 5, pp. 468–481, 2014.

[10] M. A. Lill and M. L. Danielson, "Computer-aided drug design platform using pymol," Journal of computer-aided molecular design, vol. 25, no. 1, pp. 13–19, 2011.

[11] G. Cano, J. Garcia-Rodriguez, A. Garcia-Garcia, H. Perez-Sanchez, J. A. Benediktsson, A. Thapa, and A. Barr, "Automatic selection of descriptors using random forest: Application to drug discovery," Expert Systems with Applications, vol. 72, pp. 151–159,

[12] Liu, K.-K. R. Choo, R. H. Deng, R. Lu, and J. Weng, "Efficient and privacy-preserving outsourced computation of rational numbers," IEEE Journal of Biomedical and Health Informatics, vol. 20, pp. 655 – 668, 2016.