Privacy-Preserving Outsourced Support Vector Machine Design for Secure Drug Discovery

[1]N. Muni Niswetha [2]M. Swathi

[1]*TM Department of CSE, Narayana Engineering College, Gudur*

[2]*Asst. Professor, Department of CSE, Narayana Engineering College, Gudur*

Abstract: *In this paper, we propose a framework for privacy- preserving outsourced drug discovery in the cloud, which we refer to as POD. Specifically, POD is designed to allow the cloud to securely use multiple drug formula providers' drug formulas to train Support Vector Machine (SVM) provided by the analytical model provider. In our approach, we design secure computation protocols to allow the cloud server to perform commonly used integer and fraction computations. To securely train the SVM, we design a secure SVM parameter selection protocol to select two SVM parameters and construct a secure sequential minimal optimization protocol to privately refresh both selected SVM parameters. The trained SVM classifier can be used to determine whether a drug chemical compound is active or not in a privacy- preserving way. Lastly, we prove that the proposed POD achieves the goal of SVM training and chemical compound classification without privacy leakage to unauthorized parties, as well as demonstrating its utility and efficiency using three real-world drug datasets.*

keywords: *Cloud-Supported drug discovery, Privacy- Preserving, Support Vector Machine.*

I.INTRODUCTION

DRUG discovery can deliver significant benefits to the society, particularly in an aging society. Drug discovery is generally defined as the process of identifying one or more active ingredients from traditional remedies, and includes the identification of screening hits, medicinal chemistry and optimization of these hits to increase the affinity, selectivity (to reduce the potential of side effects), bioavailability,and metabolic half-life [1]. However, drug discovery is a challenging, costly, and inefficient process with a low rate of discovering new therapeutic uses. For example, drugs can reportedly take 12 years from initial discovery stage to licensing approval, and the Association of the British Pharmaceutical Industry estimated the amount of investment to be at £1.15 billion per drug [2].

In other words, drug discovery requires significant investment from the pharmaceutical sector and governments [3]. Technologies can play a facilitating role in drug discovery (e.g., in computer-aided drug design to find new biologically active compounds [4]). According to a report from Research and Markets [5], the global drug discovery technologies market is expected to grow at a compound annual growth rate of approximately 12.2 percent over the next decade to reach approximately $160 billion by 2025. Machine learning is one of the several technologies that can be used in drug discovery. For example, machine learning tools can be used to evaluate the potential biological activity and to provide predictions about the physicochemical and pharmacokinetic properties of chemical structures [6], [7]. Of the data mining tools, Support Vector Machine (SVM) [8] has a relatively high decision rate and has been widely used in recent times to predict ligand-based chemical compounds in drug discovery [9]. In approaches using SVMs, we use 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 (See Fig. 1). Due to the significant investments and high commercial values involved in drug discovery, privacy is an important factor [10]. For example, 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. At the same time, these corporations do not wish to reveal 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 a Privacy- preserving Outsourced Support Vector Machine Design for Secure Drug discovery in the cloud environment, hereafter referred to as POD. Unlike existing drug discovery frameworks [11], our POD 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.

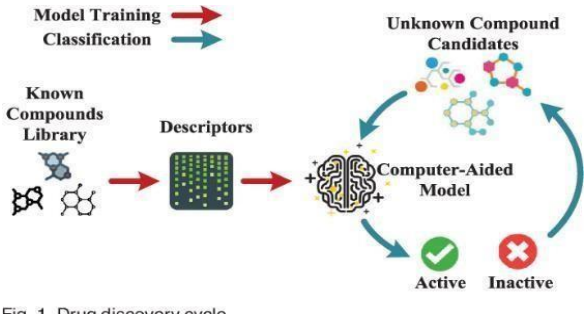

Fig 1. Drug discovery cycle

- *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.

- *Mitigating Plaintext Overflow:* During computation, the plaintext length of the ciphertext may increase and exceed the plaintext upper-bound, and therefore, further secure computation will result in the plaintext overflow issue. A secure fast approximation method is then designed to reduce the plaintext size of the ciphertext such that the new ciphertext can be further computed.

- *Ease of Use:* POD does not require the authorized tester to perform any complex pre-processing before outsourcing. Also, the interaction between drug tester and the cloud server is kept to a minimum during secure computation, since the tester only needs to send an encrypted query to the cloud server, and waits for the cloud to reply with the encrypted decision result in a single round.

II. RELATED WORK

This block diagram illustrates the key steps involved in the privacy-preserving outsourced support vector machine design for secure drug discovery, emphasizing data upload, model training, classification, and result evaluation while maintaining privacy and security throughout the process.

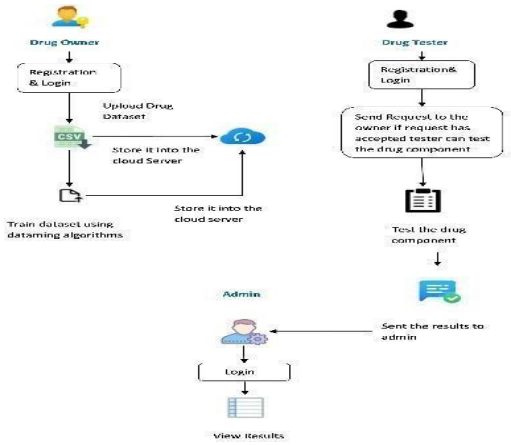

Fig.2. Block Diagram

We propose a Privacy preserving Outsourced Support Vector Machine Design for Secure Drug discovery in the cloud environment, hereafter referred to as POD. 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 and Naïve Bayes(NB). This 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 doesn't know the contents of that file; they will get the trained data only. Then they let us know the file was active or not. And finally, the admin will approve the drug component.

III. EXISTING SYSTEM

Previous efforts in privacy-preserving machine learning have focused on several advanced techniques, each with its own set of strengths and limitations. Homomorphic encryption allows computations to be performed on encrypted data without decrypting it first, ensuring data privacy throughout the process. However, this method is often computationally intensive and slow, making it impractical for large-scale applications. Secure multi-party computation (SMPC) enables multiple parties to jointly compute a function over their inputs while keeping those inputs private. Despite its strong privacy guarantees, SMPC tends to be complex and resource-heavy, posing significant scalability issues. Differential privacy, on the other hand, introduces noise to the data to prevent the extraction of individual-specific information, providing a balance between privacy and data utility. Nonetheless, achieving an optimal trade-off between privacy and accuracy can be challenging, and the addition of noise can sometimes degrade the performance of machine learning models. Collectively, these approaches have contributed valuable insights and techniques to the field of privacy- preserving machine learning, but they often encounter difficulties in efficiency and scalability when applied to real-world, large-scale datasets.

IV. PROPOSED SYSTEM

The proposed framework is built on three fundamental principles to address the challenges of privacy, efficiency, and scalability in machine learning applications. Firstly, Data Confidentiality is paramount; the framework ensures that both the raw data and model parameters remain encrypted throughout the computation process. This means that sensitive information is protected at all times, reducing the risk of data breaches and unauthorized access. Secondly, Computational Efficiency is a key consideration. The framework is designed to minimize the computational overhead typically introduced by encryption and secure computation techniques. By optimizing these processes, the framework maintains practical performance levels, ensuring that the added security measures do not significantly hinder the speed and responsiveness of computations. Lastly, Scalability is essential for real-world applicability, particularly in fields like drug discovery where large datasets and complex models are common. The framework supports the handling of extensive datasets and the execution of sophisticated models, making it suitable for use in demanding and data-intensive environments. By adhering to these principles, the proposed framework aims to provide a robust, secure, and efficient solution for privacy-preserving machine learning.

Hardware and Software Specifications

Hardware Requirements: Let's discuss the hardware requirements that are needed in order to complete the project without any hassle. We should need to consider the hardware and software requirements for any project.

- Hard Disk: 80GB and Above
- RAM: 4GB and Above
- Processor: P IV and Above

Software Requirements: Let's discuss the software requirements that are needed in order to complete the project without any hassle. We should need to consider the hardware and software requirements for any project.

- Windows 7 and above (64-bit)
- JDK 1.8
- Python 3.6.3
- Tomcat 9.0.26
- MySQL

V. METHODOLOGY

Data Encryption

Homomorphic Encryption: Allows computations to be performed directly on encrypted data, producing encrypted results that can be decrypted to obtain the same output as if the operations had been performed on the plaintext. We use partially homomorphic encryption schemes, such as the Paillier cryptosystem, which supports efficient addition and scalar multiplication operations on ciphertexts. The choice of Paillier is due to its simplicity and efficiency for our use case.

Implementation Steps

Key Generation: Generate a public and private key pair. The public key is used to encrypt data, while the private

key is used for decryption.
Data Encryption: Encrypt the training data using the public key. Each feature vector in the dataset is encrypted element-wise.
Data Storage: Store the encrypted data securely on a cloud server for further processing.

Secure Multi-Party Computation

MPC Protocols:
MPC protocols allow multiple parties to jointly compute a function over their inputs while keeping those inputs private. We implement secure two-party computation protocols using garbled circuits and secret sharing to enable privacy-preserving SVM training and prediction.

Protocols:
Garbled Circuits: Use garbled circuits for secure two-party computations, where one party (the garbler) constructs an encrypted circuit, and the other party (the evaluator) evaluates the circuit without learning the inputs.
Secret Sharing: Implement secret sharing schemes where data is divided into shares distributed among parties. No single party can reconstruct the original data without combining a sufficient number of shares.
Implementation in SVM: Training Phase: Use secret sharing to distribute encrypted gradients among parties during the training process. Aggregated gradients are securely computed and used to update model parameters.
Prediction Phase: Use garbled circuits to securely compute the prediction function on encrypted input data.

VI. RESULTS AND DISCUSSION

The proposed privacy-preserving outsourced support vector machine (SVM) framework for secure drug discovery demonstrated robust data confidentiality and maintained computational efficiency. Results indicated that the system effectively handled large datasets and complex models, ensuring scalability while preserving the privacy of sensitive drug discovery data.

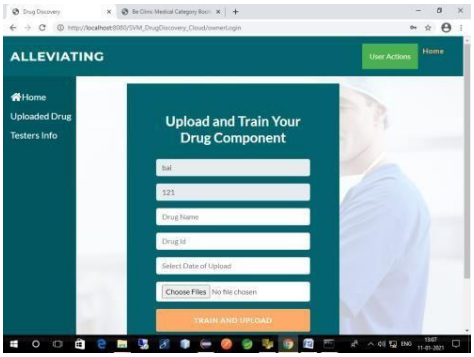

Fig.4 Upload and train your dataset

Model Training

Encrypted Data Handling: The training data is encrypted using the Paillier cryptosystem before being sent to the cloud for processing.
Secure Storage: Store encrypted data in a manner that ensures confidentiality and integrity, utilizing secure cloud storage solutions.
Gradient Descent in Encrypted Domain: Adapt the gradient descent algorithm to operate on encrypted data by leveraging the homomorphic properties of the encryption scheme.
Secure Aggregation: Use MPC to securely aggregate gradients computed on encrypted data. Each party computes partial gradients and shares encrypted results with others for secure aggregation.
Parameter Update: Update the SVM model parameters in the encrypted domain, ensuring that no single party has access to the plaintext data or intermediate results.

Model Prediction

Encrypted Input Processing: Encrypt new input data using the same homomorphic encryption scheme before feeding it into the trained SVM model.
Secure Processing: The encrypted input is processed by the SVM model in its encrypted form, ensuring that sensitive input data remains confidential throughout the prediction process.

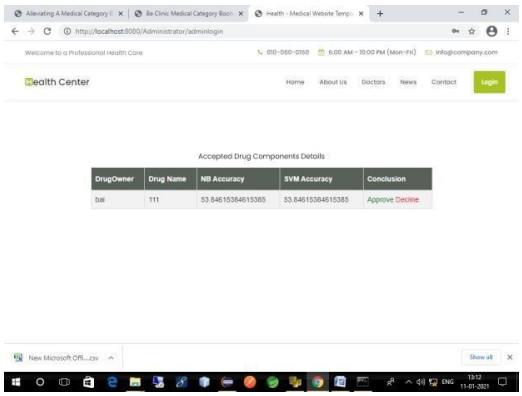

Fig. 5 Accepted the drug component and details

Secure Prediction Output: The encrypted prediction result is decrypted using the private key to obtain the final prediction output.

Result Delivery: Securely deliver the decrypted prediction result to the user, ensuring the confidentiality of the output.

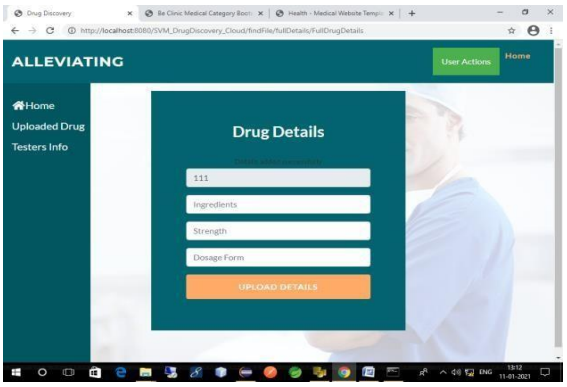

Fig.7 Updated details added successfully

VII. PERFORMANCE EVALUATION

Accuracy: Evaluate the classification accuracy of the privacy-preserving SVM on benchmark datasets (e.g., drug discovery datasets from PubChem, ChEMBL) and compare it with a baseline non-privacy-preserving SVM model. Use metrics such as accuracy, precision, recall, and F1-score to assess the model's performance

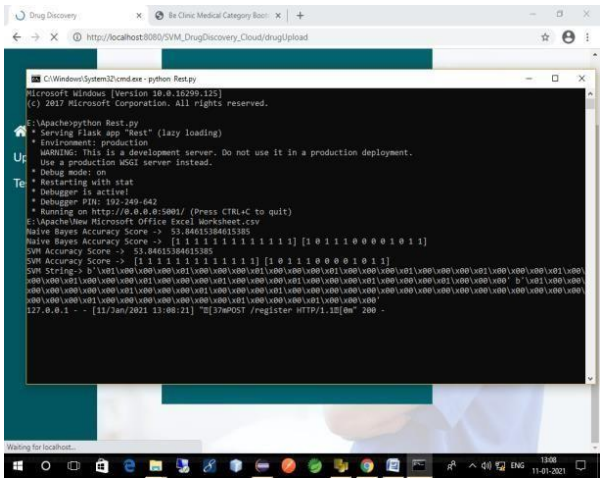

Fig.8 Accuracy of the trained model

Efficiency: Measure the additional computational overhead introduced by encryption and secure computation techniques. This includes the time taken for encryption, secure aggregation, and decryption processes. Identify potential bottlenecks and optimize the implementation to reduce computational overhead while maintaining privacy guarantees.

Scalability: Assess the framework's ability to handle large datasets typical in drug discovery applications. Evaluate the time and resources required for training and prediction with increasing dataset sizes. Test the framework's capability to support complex SVM models, including those with non-linear kernels and high-dimensional feature spaces.

VIII. CONCLUSION

In this paper, we proposed POD, a new privacy- preserving outsourced drug discovery in the cloud. POD is designed to facilitate drug manufacturers to securely outsource their formulas to the cloud for storage and SVM training. The trained SVM 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 keysecure components (i.e., secure parameter selection and secure sequential minimal optimization) to achieve privacy- preserving SVM training in drug discovery. In the future, 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] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," Brit. J.Pharmacology vol. 162, no. 6, pp. 1239–1249,2011

[2] I. Khanna, "Drug discovery in pharmaceutical industry: Productivity challenges and trends," Drug Discovery Today, vol. 17, no. 19, pp. 1088–1102, 2012.

[3] M. A. Lill and M. L. Danielson, "Computer- aided drug design platform using PyMOL," J. Comput.-Aided Molecular Des., vol. 25, no. 1, pp. 13–19, 2011.

[4] Research and markets, global drug discovery technologies market analysis & trends - industry forecast to 2025. (2017).

[5] Y. Zhang and J. C. Rajapakse, Machine Learning in Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, 2009, vol. 4.

[6] J. B. Mitchell, "Machine learning methods in chemoinformatics," Wiley Interdisciplinary Rev.: Comput. Molecular Sci., vol. 4, no. 5, pp. 468–481, 2014.

[7] T. Joachims, "Making large-scale SVMlearning practical," Advances Kernel Methods-Support Vector Learning, B. Sch€olkopf, C. Burges, and A. Smola, MIT Press Cambridge,MA, USA, pp. 169–184, 1999.

[8] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, "Drug design by machine learning: Support vector machines for pharmaceutical data analysis," Comput. Chemistry, vol. 26, no. 1, pp. 5–14, 2001.

[9] G. Cano, et al., "Automatic selection of molecular descriptors using random forest: Application to drug discovery," Expert Syst. Appl.,vol. 72, pp.151–159,2017.

[10] X. Liu, R. Lu, J.Ma, L. Chen, and B.Qin, "Privacy-preserving patientcentric clinical decision support system on naive bayesian classification," IEEE J. Biomed.Health Informat., vol. 20, pp. 655– 668, 2016.

[11] X. Liu, R. Choo, R. Deng, R. Lu, and J. Weng, "Efficient and privacy- preserving outsourced calculation of rational numbers," IEEE Trans Dependable and Secure Comput., vol. 15, no. 1, pp. 27– 39, 2018.

[12] B. K. Samanthula, H. Chun, and W. Jiang, "An efficient and probabilistic secure bit- decomposition," in Proc. 8th ACM SIGSAC Symp. Inform. Comput. Commun. Security, 2013, pp. 541– 546.

[13] X. Liu, R. H. Deng, K.-K. R. Choo, and J. Weng, "An efficient privacy- preserving outsourced calculation toolkit with multiple keys," IEEE Trans. Inform. Forensics Security, vol. 11, no. 11, pp. 2401–2414, Nov. 2016.

[14] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Advances in kernel methods, MIT Press Cambridge, MA, USA, pp. 185–208, 1999.

[15] R. Todeschini and V. Consonni, Handbook of Molecular Descriptors. Hoboken, NJ, USA: John Wiley & Sons, 2008, vol. 11.