Secure and Efficient Outsourced Clustering using k-Means with Fully Homomorphic Encryption by Ciphertext Packing Technique

K. BALAKRISHNA MARUTHIRAM[1], DR. G. VENKATA RAMI REDDY[2], M. RAGHAVENDRA[3]
[1] *Assistant Professor of CSE, Department of IT, JNTU Hyderabad, Hyderabad, India*
[2] *Professor of IT, Department of IT, JNTU Hyderabad, Hyderabad, India*
[3] *Student, M. Tech (Computer Networks and Information Security), Department of Information Technology, JNTU Hyderabad, Hyderabad, India*

Abstract*— In our current reality where enormous organizations use cloud administrations to run applications, safeguarding information security and privacy is basic. This is especially valid for sensitive enterprises like medical care, money, and AI-driven applications where independent direction depends on ML calculations. At the point when cloud servers are utilized for ML exercises, there are stresses with respect to information openness and conceivable abuse. This examination proposes an inventive procedure to beat these troubles. The scholars give a clever thought: Secure and Re-appropriate KMEANS ML Calculation to safeguard information protection and save the mystery of delicate data. Utilizing Fully Homomorphic Encryption (FHE), which grants performing expansion and augmentation straightforwardly on encoded information without unveiling the first information, this clever methodology utilizes scrambled information. Secure estimations on the encoded information are made conceivable by the mystery key produced by the FHE encryption, which kills the requirement for unscrambling. The improvement of YASCHE Completely Homomorphic encryption, which joins numerous plaintexts into a solitary ciphertext, is the premise of this imaginative system. This pressing system makes it more straightforward to register in equal, which permits information to be moved to a few mists and enormously builds the plan's processing effectiveness. Since FHE depends on numerical polynomial computation, it is feasible to perform ML exercises on encoded information in a solid and successful way while making preparations for unlawful access and cloud suppliers abusing the information. Basically, the proposed approach empowers organizations to involve the cloud's handling limit with respect to ML tasks while maintaining the best principles of information security and protection. It guarantees the security of delicate information in a period of expanding interest for cloud administrations by making ready for protected and powerful AI like facial acknowledgment, wellbeing expectation, and qualification assessments for advances or Visas in a cloud-based climate.

Index Terms- Cloud Computing, Fully Homomorphic Encryption, K-Means Clustering, Privacy-Preserving.

I. INTRODUCTION

Machine Learning as a Service (MLaaS) has been more and more popular in the last few years because of its affordability and scalability. Cloud-based machine learning systems, such AWS Machine Learning [1], Microsoft Machine Learning Studio [2], and Google AI Platform [3], are offered by well-known tech companies like Amazon, Microsoft, and Google. In order to take advantage of high-performance cloud computing resources and lower implementation costs locally, more customers are prepared to experiment with machine learning as a service (MLaaS), outsourcing their data and machine learning workloads to cloud service providers (RightScale, 2014). However, both the original data and the learning outcomes can be extremely sensitive for users in delicate fields like banking, government, and healthcare. These consumers are frequently reluctant or unable to employ MLaaS in cloud settings because to privacy and security concerns.

MLaaS data privacy may be maintained, for example, by using homomorphic encryption (HE), which enables arithmetic operations on ciphertexts without the need for decryption. In order to create interactive protocols between two non-colluding clouds and accomplish addition, multiplication, and other complicated operations, previous research [5-8] has concentrated on partly homomorphic encryptions

(PHE). Nevertheless, these approaches are not appropriate for huge datasets because to their high computational and communication costs. Neural network prediction over encrypted data has recently made use of the fully homomorphic encryption (FHE) known as YASHE (Yet Another Somewhat Homomorphic Encryption) [10] [11]. By combining many plaintexts into a single ciphertext, this technique offers Single Instruction Multiple Data (SIMD) operations [12], enabling parallel computing and greatly increasing efficiency. Although YASHE encryption has its benefits, it is better appropriate for privacy-preserving machine learning methods that need fewer ciphertext multiplications since it is restricted to arithmetic circuits with a predefined multiplicative depth.

With an emphasis on k-means clustering, this work attempts to effectively construct privacy-preserving MLaaS. We suggest utilizing YASHE encryption to create a novel, safe, and effective outsourced k-means clustering (SEOKC) technique. The maximum number of successive ciphertext multiplications in privacy-preserving k-means clustering is two (see Section 4 for details), which allows smaller parameters to be used for better results. Our method uses ciphertext packing to handle huge encrypted databases effectively, unlike earlier PHE-based systems [5–9]. To create secure interactive protocols, we use a concept of two non-colluding cloud servers [5–9], as YASHE encryption does not directly provide ciphertext comparison. Given the widespread usage of k-means clustering in many different disciplines, this technique is crucial for real-world applications [13, 14]. Our approach may also be modified for additional privacy-preserving machine learning applications including association rule mining and categorization.

II. LITERATURE SURVEY

A lot of attention has been paid to privacy-preserving data mining as worries about data security and privacy have grown. Specifically, a great deal of research has been done on privacy-preserving clustering techniques to help with the problem of handling sensitive data for data mining activities without sacrificing privacy. A privacy-preserving multi-user k-means clustering technique was first developed by Fang-Yu Rao et al. [5] in their foundational work. This approach allows several users to work together to perform clustering on their combined datasets while maintaining the privacy of individual data points. Using secure multi-party computation methods, this approach maintains privacy in an outsourced environment.

Vadlana Baby and N. Subhash Chandra [20] presented a distributed threshold k-means clustering method for privacy-preserving data mining in the field of distributed data mining. By securing the data throughout the clustering process with threshold cryptography, their method solves privacy problems in dispersed situations. With the use of this technique, several data owners can work together to compute clusters without disclosing any of their personal information to outside parties.

The study of privacy-preserving clustering across horizontally and vertically partitioned data conducted by Mina Sheikhalishahi and Fabio Martinelli [21] is another important contribution to the field. With the goal of protecting data privacy throughout the clustering process, they created methods for performing clustering on data that is divided either vertically or horizontally over several locations. Their techniques are especially useful in situations when data is dispersed among several places and data owners are hesitant to release their unprocessed data.
Dongxi Liu et al.'s research [22] delves more into the subject of outsourced k-means clustering privacy. They put out a plan that contracts out the clustering computation to an outside service provider while maintaining data privacy. In order to guarantee that the service provider cannot access the raw data or the intermediate outcomes, this method uses cryptographic methods to safeguard the secrecy of the data during the clustering process.

Abdulatif Alabdulatif et al. [29] introduced a privacy-preserving data clustering technique based on completely homomorphic encryption in the context of cloud computing. With this method, encrypted data can be kept in the cloud and clustering can be done on it without having to first decrypt it. This technique makes sure that the information is kept private even when there are possible security risks in the cloud environment. Fully homomorphic encryption offers a high degree of security and is therefore a suitable

solution for cloud computing privacy-preserving clustering.

All things considered, these studies demonstrate the several methods that may be used to accomplish privacy-preserving clustering, each of which tackles a distinct facet of the issue. From threshold cryptography and safe multi-party computing to completely homomorphic encryption, each technique offers special benefits in terms of efficiency and security. In order to ensure that data mining activities may be completed without jeopardizing the privacy of sensitive data, creative solutions are still being developed via continuing research in this subject.

III. METHODOLOGY

i) Proposed Work:

Using Fully Homomorphic Encryption (FHE), the suggested system presents a revolutionary approach: Secure and Outsource KMEANS Machine Learning Algorithm. This novel method allows for mathematical operations on encrypted data without the need for decryption, ensuring data privacy and secrecy. The secret to this strategy is YASCHE Fully Homomorphic encryption, which improves computational performance by allowing data outsourcing to several clouds and parallel processing. Because FHE uses polynomial computing, machine learning is safe and effective while limiting unwanted access. With the greatest degree of data privacy and security, this technology enables businesses to use cloud computing for AI applications. It allows for safe face recognition, health prediction, and eligibility evaluations for credit cards or loans in a cloud-based setting.

ii) System Architecture:

Diagram of a cloud-based system that clusters encrypted data using k-means using fully homomorphic encryption (FHE). A data analysis method called K-means clustering is used to put related data points in one category. With FHE encryption, calculations may be done on encrypted data without having to first decode it. Because it enables users to send their data to the cloud for processing without worrying about it being exposed to the cloud provider, this is significant for cloud computing.

The data, denoted by D in the figure, is encrypted with a private key. After that, the encrypted data is sent to the cloud and processed there using FHE. The user is subsequently given an encrypted copy of the clustering findings.

With the help of this technology, users may safely send their data to be processed on the cloud. It's crucial to remember that FHE requires a lot of computing power, thus not all applications will benefit from it.

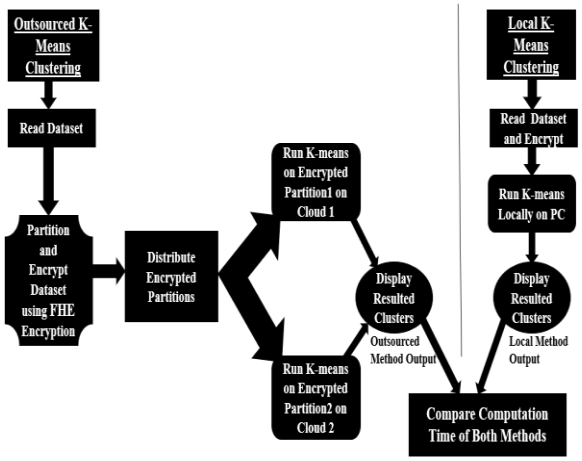

Fig 1 Proposed Architecture

iii) Modules:

The following modules were utilized in the project's implementation. The following is a description of these modules:

Cloud1: Since we lack cloud servers, we have built a fake Python cloud that will first undergo clustering and an encrypted partition.

Cloud 2: This is the second cloud that underwent clustering after receiving a second encrypted partition. The modules for data owners are listed below.

Reaction Network Dataset Upload: The clustering effort in this study was carried out by the authors using the "Reaction Network Dataset". We also used the same dataset in order to preserve consistency and enable comparison analysis. This decision guarantees the comparability of our findings with earlier studies, allowing for a more precise assessment of the clustering techniques employed. Using the "Reaction Network Dataset," we want to verify the efficacy and

efficiency of the privacy-preserving clustering methods we have suggested.

Divide the dataset and use YASHE encryption to encrypt it: To securely encrypt the dataset, we use the YASHE Fully Homomorphic Encryption (FHE) technique. To protect data privacy during further processing, we first encrypt the complete dataset using our encryption module. We divide the dataset into digestible chunks once it has been encrypted. This method enables effective processing and analysis while protecting the privacy of the data. We guarantee data security during partitioning by utilizing YASHE FHE, which allows for privacy-preserving clustering without jeopardizing the integrity and confidentiality of the original dataset.

Access Secured Partition: We may examine the encrypted entries inside each partitioned dataset segment using the "View Encrypted Partition" module. We may confirm that the dataset has been appropriately encrypted and partitioned without jeopardizing data security by using this module. Before moving further with additional research, this feature is essential for guaranteeing the security and integrity of the encrypted data. It offers a mechanism to verify the encryption procedure, guaranteeing that all records stay safe and encrypted during the partitioning process, protecting critical data privacy while clustering activities are being completed.

Distribute Encrypted Partition: Distribution of encrypted dataset partitions to various cloud services is made easier with the help of the "Distribute Encrypted Partition" module. By utilizing various clouds, this module makes sure that the dataset is handled efficiently, which improves processing speed and resource consumption overall. Our goal for the clustering challenge is to produce four clusters. After processing the encrypted data, the cloud services provide cluster labels that show which cluster each record is assigned to. We record the beginning and ending times of the clustering process in order to gauge the effectiveness of the outsourced k-means algorithm. This method takes advantage of the dispersed nature of cloud computing to boost processing speed and scalability while ensuring effective, privacy-preserving clustering.

Run KMeans Locally: We can run the k-means clustering method locally on the same dataset by utilizing the "Run KMeans Locally" module. When compared to distributed approaches, this module may be less efficient because it processes the full dataset in-house. We determine the cluster computation time and record the start and finish timings of the procedure by executing the k-means algorithm locally. This local execution highlights the benefits of leveraging dispersed cloud resources for this work by acting as a benchmark to assess the effectiveness and performance of the outsourced k-means clustering.

Comparison Graph: The proposed outsourced k-means clustering (using multiple clouds) and the current local k-means algorithm are compared in terms of execution time using a visual representation produced by the "Comparison Graph" module. We show the differences in efficiency between the two methods by graphing these timings. The graph illustrates the differences in performance between performing k-means clustering locally and outsourcing it to dispersed cloud environments. This comparison aids in assessing the speed and scalability gains made possible by outsourcing and highlights the possible advantages of using cloud resources for privacy-preserving data clustering applications

iv) Algorithms:

K-Means Algorithm: Information focuses are isolated into K clusters using the iterative K-means clustering grouping strategy as indicated by how comparable their elements are. The cycle begins with the irregular choice of K centroids, appoints every information highlight the nearest centroid, figures new centroids as the mean of the data of interest in each bunch, and afterward proceeds with the interaction until the quantity of emphasess is reached or the centroids settle.

YASHE Fully Homomorphic Encryption (FHE) Algorithm: YASHE FHE empowers encoded information handling without the requirement for unscrambling. To safeguard the security of the scrambled information and permit computations on it, it involves key creation, encryption, and homomorphic operations (addition and multiplication). Applications like clustering that are

delicate to security can profit from this capacity's protected information handling.

IV. EXPERIMENTAL RESULTS

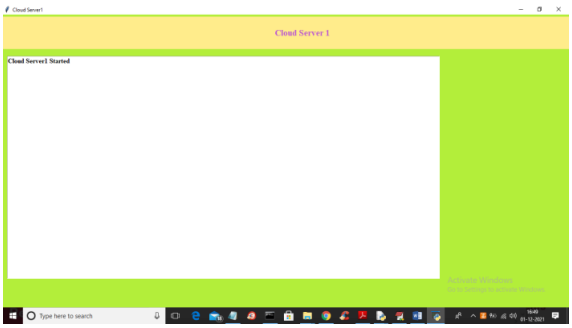

Fig 2 Cloud Server

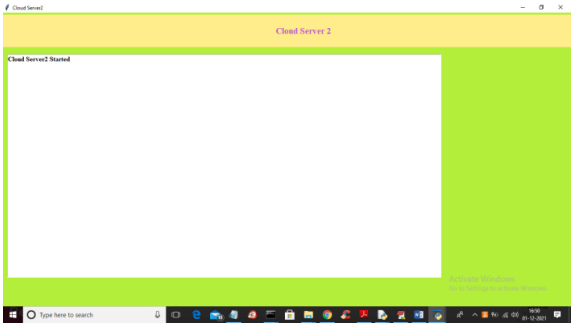

Fig 3 Cloud Server 2

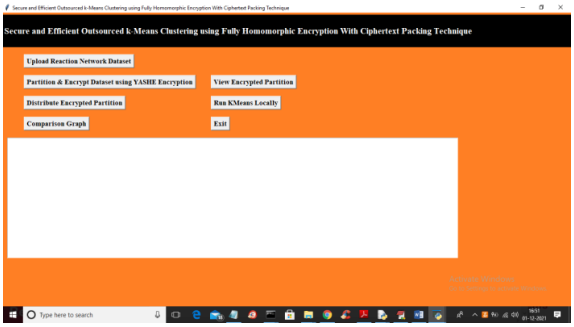

Fig 4 Home Page

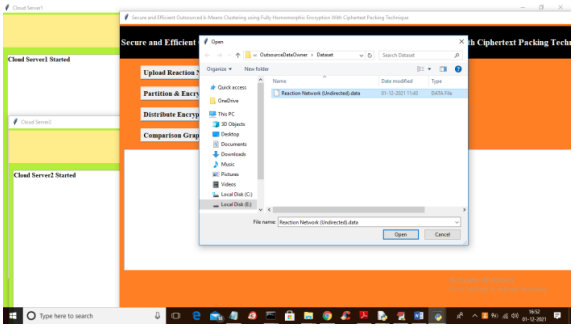

Fig 5 Uploading Dataset Page

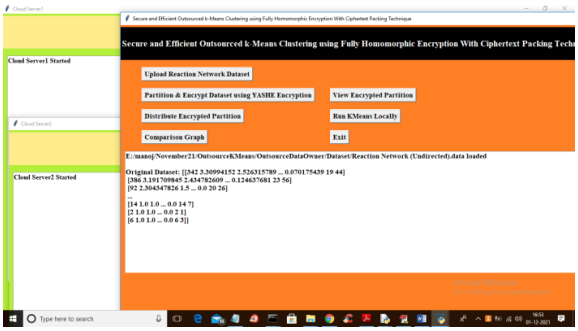

Fig 6 Dataset Uploaded

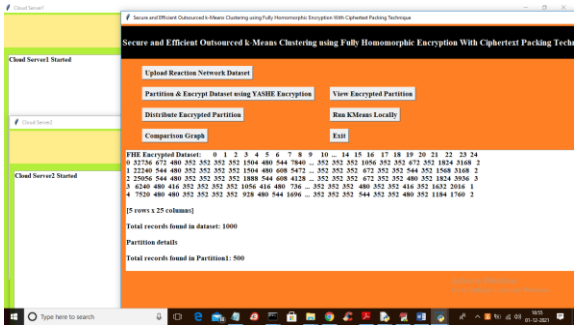

Fig 7 Partition & Encrypt Dataset using YASHE Encryption

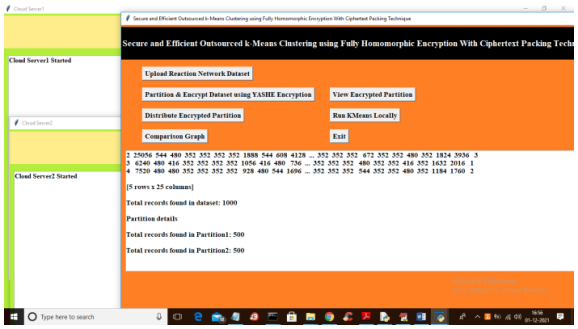

Fig 8 View Encrypted Partition

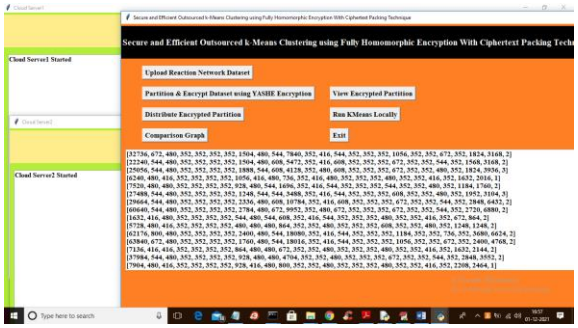

Fig 9 Distributed Encrypted Partition

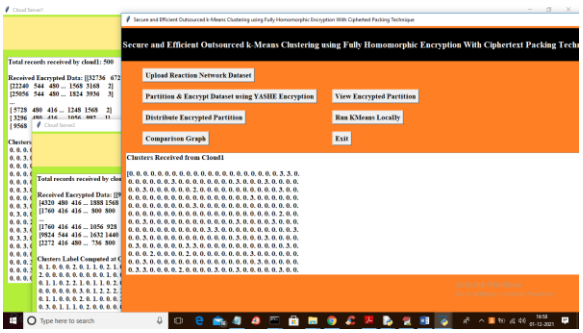

Fig 10 Clusters Received from Cloud 1- Run K-Means Clustering

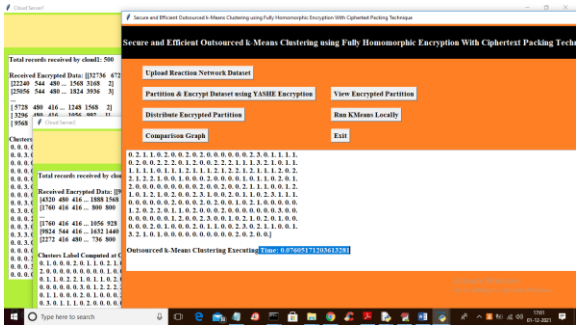

Fig 11 Outsourced K-Means Clustering Executing Time

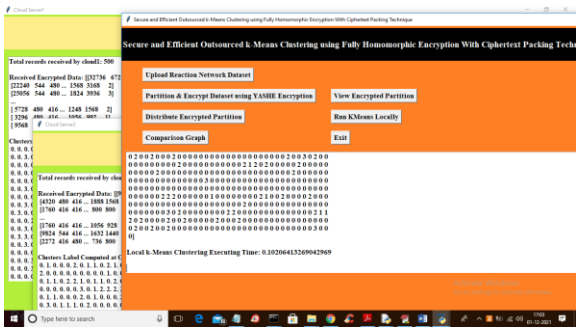

Fig 12 Local K-Means Clustering Executing Time

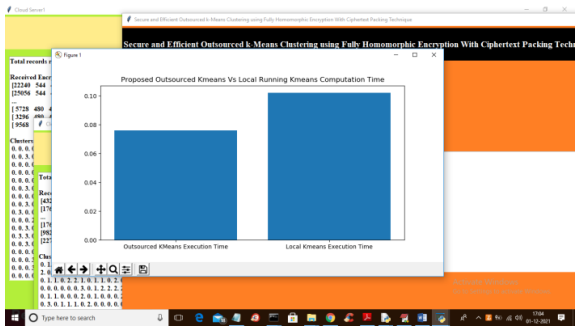

Fig 13 Comparison Graph

CONCLUSION

In synopsis, a solid obligation to information security and protection is required given the developing reliance on cloud administrations for application execution, particularly in delicate businesses like banking, medical care, and computer based intelligence driven ML. The standard use of cloud servers has started serious concerns over the conceivable revelation and double-dealing of secret information. This work presents an original thought that utilizes Fully Homomorphic Encryption (FHE) to beat these troubles: the Solid and Rethink KMEANS ML Calculation. FHE offers an unmatched level of information security by empowering numerical procedure on encoded information without unveiling the first information. YASCHE Completely Homomorphic Encryption, which empowers equal registering and information moving to a few cloud suppliers, is the cornerstone of this clever procedure. Since it increments computational proficiency, this is a functional option for organizations seeking use distributed computing limit with regards to ML projects while keeping up with information classification and protection. By executing FHE and the recommended procedure, organizations can utilize cloud-based AI applications like face acknowledgment, wellbeing forecast, and charge card or advance qualification checks with certainty, realizing that private data is safeguarded from misuse and unlawful access by cloud suppliers. This strategy not just fulfills the market's ongoing requirements for cloud administrations, however it likewise lays out another benchmark for keeping up with information mystery when it is significant to safeguard information protection. The Solid and Re-appropriate KMEANS ML Calculation gives an exploring arrangement in this present reality where the computerized scene is continuously evolving. It stays aware of the times, yet in addition continues onward with a more proficient, secure, and protection cognizant future for cloud-based AI applications and taking care of touchy information.

FUTURE SCOPE

Our future review will focus on carrying out security protecting k-means clustering across a few cloud settings, determined to grow our discoveries into

additional practical conditions. To further develop versatility and productivity, this adjustment will enhance the dispersion and coordination of encoded information allotments among a few cloud suppliers. We likewise plan to explore mixture cloud structures and state of the art cryptography techniques to improve the security and usefulness of our grouping calculations. By resolving these issues, frameworks that are strong and ready to oversee huge scope datasets while meeting the severe security and protection prerequisites of certifiable applications will be created.

REFERENCES

[1] Machine Learning on AWS. URL https://amazonaws-china.com/machine-learning/?nc1=h ls.

[2] Microsoft Azure Machine Learning Studio. URLhttps://azure.microsoft.com/en-us/services/machine-learning-studio/.

[3] AI Platform. URL https://cloud.google.com/ai-platform/.

[4] 2019 State of the Cloud Report: See the LatestCloud Trends. URL https://info.flexerasoftware.com/SLO-WP-State-of-the-Cloud-2019?id=FLX HP-SOTC2019.

[5] Fang-Yu Rao, Bharath K Samanthula, Elisa Bertino, XunYi, and Dongxi Liu. Privacy-preserving and outsourced multi-user k-means clustering. In Collaboration and Internet Computing (CIC), 2015 IEEE Conference on, pages 80–89. IEEE, 2015.

[6] Bharath K Samanthula, Yousef Elmehdwi, and Wei Jiang. K-nearest neighbor classification over semantically secure encrypted relational data. IEEE Transactions on Knowledge and Data Engineering, 27(5):1261–1273, 2015.

[7] Hong Rong, Hui-Mei Wang, Jian Liu, and Ming Xian. Privacy-preserving k-nearest neighbor computation in multiple cloud environments. IEEE Access, 4:9589–9603, 2016.

[8] Hong Rong, Huimei Wang, Jian Liu, Jialu Hao, and Ming Xian. Privacy-preserving k-means clustering under multiowner setting in distributed cloud environments. Security and Communication Networks, 2017, 2017.

[9] Wei Wu, Jian Liu, Hong Rong, Huimei Wang, and Ming Xian. Efficient k-nearest neighbor classification over semantically secure hybrid encrypted cloud database. IEEE Access, 6:41771–41784, 2018.

[10] Joppe W Bos, Kristin Lauter, Jake Loftus, and Michael Naehrig. Improved security for a ring-based fully homomorphic encryption scheme. In IMA International Conference on Cryptography and Coding, pages 45–64. Springer, 2013.

[11] Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In International Conference on Machine Learning, pages 201–210, 2016.

[12] Nigel P Smart and Frederik Vercauteren. Fully homomorphic simd operations. Designs, codes and cryptography, 71(1):57–81, 2014.

[13] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. Image segmentation using k-means clustering algorithm and subtractive clustering.

[14] Zeyad Safaa Younus, Dzulkifli Mohamad, Tanzila Saba, Mohammed Hazim Alkawaz, Amjad Rehman, Mznah Al-Rodhaan, and Abdullah Al-Dhelaan. Content-based image retrieval using pso and k-means clustering algorithm. Arabian Journal of Geosciences, 8(8):6211–6224, 2015.

[15] Oded Goldreich. Encryption schemes. The Foundations of Cryptography, 2:373–470, 2004.

[16] Stuart Lloyd. Least squares quantization in pcm. IEEE transactions on information theory, 28(2):129–137, 1982.

[17] Vadim Lyubashevsky, Chris Peikert, and Oded Regev. On ideal lattices and learning with errors over rings. In Annual International Conference on the Theory and Applications of Cryptographic Techniques, pages 1–23. Springer, 2010.

[18] Hao Chen, Kim Laine, and Rachel Player. Simple encrypted arithmetic library-seal v2.3.0-4. 2017.

[19] Xun Yi and Yanchun Zhang. Equally contributory privacy-preserving k-means clustering over vertically partitioned data. Information systems, 38(1):97–107, 2013.

[20] Vadlana Baby and N Subhash Chandra. Distributed threshold k-means clustering for privacy preserving data mining. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 2286–2289. IEEE, 2016.

[21] Mina Sheikhalishahi and Fabio Martinelli. Privacy preserving clustering over horizontal and vertical partitioned data. In 2017 IEEE Symposium on Computers and Communications (ISCC), pages 1237–1244. IEEE, 2017.

[22] Dongxi Liu, Elisa Bertino, and Xun Yi. Privacy of outsourced k-means clustering. In Proceedings of the 9th ACM symposium on Information, computer and communications security, pages 123–134. ACM, 2014.

[23] Nawal Almutairi, Frans Coenen, and Keith Dures. Kmeans clustering using homomorphic encryption and an updatable distance matrix: Secure third party data clustering with limited data owner interaction. In International Conference on Big Data Analytics and Knowledge Discovery, pages 274–285. Springer, 2017.

[24] Yongge Wang. Notes on two fully homomorphic encryption schemes without bootstrapping. IACR Cryptology ePrint Archive, 2015:519, 2015.

[25] Keng-Pei Lin. Privacy-preserving kernel k-means clustering outsourcing with random transformation. Knowledge and Information Systems, 49(3):885–908, 2016.

[26] Rakesh Agrawal and Ramakrishnan Srikant. Privacypreserving data mining, volume 29. ACM, 2000.

[27] Wai Kit Wong, David Wai-lok Cheung, Ben Kao, and Nikos Mamoulis. Secure knn computation on encrypted databases. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, pages 139–152. ACM, 2009.

[28] Yila Huang, Qiwei Lu, and Yan Xiong. Collaborative outsourced data mining for secure cloud computing. Journal of Networks, 9(10):2655, 2014.

[29] Abdulatif Alabdulatif, Ibrahim Khalil, Mark Reynolds, Heshan Kumarage, and Xun Yi. Privacy-preserving data clustering in cloud computing based on fully homomorphic encryption. In PACIS, page 289, 2017.

[30] Naeem Muhammad and Asghar Sohail. KEGG Metabolic Reaction Network Data Set, 2011. URL https://archive.ics.uci.edu/ml/datasets/KEGG+Metabolic+Reaction+Network+(Undirected).