Privacy-Preserving and Efficient Outsourced k-Means Clustering via Fully Homomorphic Encryption and Ciphertext Packing

P.Arun Reddy[1], Uppada Bhagavatha Reddy[2], Putkapu Manideep[3], Sarla Harish[4], Jutla Vijay[5]

[1] Assistant Professor, Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

[2,3,4,5]BTech Students ,Department of Computer Science and Engineering(AI & ML), Samskruthi College of Engineering And Technology , Kondapur(V), Ghatkesar(M), Medchal(D),Telangana

***Abstract*—Outsourcing data processing to the cloud offers scalability and cost benefits, but it raises serious concerns about data privacy. This work presents a secure and efficient approach to performing k-means clustering on encrypted data using fully homomorphic encryption . By leveraging ciphertext packing techniques, multiple data values are processed simultaneously, significantly improving computational efficiency compared to traditional FHE methods. The proposed method ensures that sensitive information remains encrypted throughout the entire clustering process, preventing unauthorized access even from the service provider. At the same time, it maintains accuracy comparable to standard k-means clustering on plaintext data. Experimental results demonstrate that the approach reduces computation time and resource usage while preserving strong security guarantees. This makes it a practical solution for privacy-preserving data analytics in cloud environments, particularly in applications involving sensitive datasets such as healthcare, finance, and user behaviour analysis.**

***Keywords*—Privacy-preserving data analysis, Outsourced k-means clustering, Fully Homomorphic Encryption, Ciphertext packing, Secure cloud computing, Encrypted data processing, Computational efficiency, Data confidentiality.**

I. INTRODUCTION

In recent years, the rapid growth of data generation and the widespread adoption of cloud computing have significantly transformed how organizations store and process information. Cloud platforms such as Amazon Web Services, Microsoft Azure, and Google Cloud provide scalable and cost-effective solutions for handling large-scale data analytics tasks [1],[3]. Reports on cloud usage trends further highlight the increasing reliance on outsourced infrastructures for data-driven applications [4].

Among various data mining techniques, k-means clustering remains one of the most widely used algorithms due to its simplicity and effectiveness in grouping similar data points [16]. It has been successfully applied in domains such as image processing, content-based retrieval, healthcare, and business analytics [13], [14].

Despite its advantages, outsourcing data and computation to cloud environments raises serious concerns regarding data privacy and security. Sensitive data, when exposed to third-party service providers, is vulnerable to unauthorized access and potential misuse. Traditional approaches require data to be decrypted before processing, which compromises confidentiality.

To overcome this limitation, privacy-preserving data mining techniques have been proposed, enabling computations to be performed without revealing raw data [22]. Several studies have explored secure outsourcing of clustering and classification tasks, including privacy-preserving k-means and k-nearest neighbour algorithms [5] , [9], [19], [20].

Fully Homomorphic Encryption (FHE) has emerged as a powerful cryptographic solution that allows computations to be carried out directly on encrypted data without decryption [15]. Advanced FHE schemes based on lattice cryptography and ring learning with errors have further improved security and functionality [10], [17].

Additionally, practical implementations such as Microsoft SEAL demonstrate the feasibility of encrypted computation in real-world scenarios [18]. However, FHE-based approaches often suffer from high computational complexity, making them less efficient for large-scale applications.

To address these challenges, optimization techniques such as ciphertext packing and SIMD operations have been introduced to improve

efficiency [12]. These techniques enable multiple data elements to be processed simultaneously within a single ciphertext, significantly reducing computation time and resource consumption. Recent research has combined FHE with such optimizations to achieve secure and efficient outsourced clustering [23], [25].

This work focuses on developing a secure and efficient framework for outsourced k-means clustering using fully homomorphic encryption with ciphertext packing. The proposed approach aims to preserve data confidentiality throughout the computation process while improving performance. By integrating strong cryptographic guarantees with practical efficiency enhancements, this study contributes toward enabling scalable and privacy-preserving data analytics in modern cloud environments.

II. RELATED WORK

[1] Machine Learning on AWS Machine Learning on AWS provides a comprehensive suite of tools and services that enable developers and organizations to build, train, and deploy machine learning models at scale. It supports a wide range of applications, including predictive analytics, recommendation systems, and natural language processing. AWS offers fully managed services such as SageMaker, which simplifies the end-to-end machine learning workflow, from data preparation to model deployment. One of its key advantages is scalability, allowing users to handle large datasets efficiently without investing in physical infrastructure. Additionally, AWS integrates security features that help protect sensitive data during processing and storage. This platform is widely used in industry due to its flexibility, reliability, and cost-effectiveness. Its relevance to research lies in demonstrating how cloud-based environments can support advanced analytics while highlighting the need for secure computation methods when handling sensitive information.

[2] Microsoft Azure Machine Learning Studio Microsoft Azure Machine Learning Studio is a cloud-based platform designed to simplify the development, training, and deployment of machine learning models. It provides a user-friendly interface with drag-and-drop capabilities, making it accessible to both beginners and experienced practitioners. Azure ML supports a variety of algorithms and integrates seamlessly with other Microsoft services, enabling efficient data processing and model management. The platform emphasizes scalability and collaboration, allowing teams to work together on machine learning projects in real time. It also includes built-in tools for monitoring, versioning, and deploying models into production environments. Security is a critical aspect, with features that ensure data protection and compliance with industry standards. Azure ML is particularly significant in the context of outsourced data processing, as it highlights the importance of maintaining privacy and security when performing computations on cloud-hosted data.

[3] Google Cloud AI Platform Google Cloud AI Platform is a powerful cloud-based service that enables users to build, train, and deploy machine learning models using Google's infrastructure. It supports a wide range of frameworks, including TensorFlow, PyTorch, and Scikit-learn, providing flexibility for developers. The platform is designed for scalability, allowing users to process large datasets efficiently and deploy models globally. It also offers automated machine learning capabilities, reducing the complexity of

model development. Integration with other Google Cloud services enhances data management and analytics capabilities. Security features such as data encryption and access control mechanisms help protect sensitive information. Google Cloud AI Platform is widely adopted in both research and industry due to its performance and ease of use. Its role in outsourced computation highlights the growing reliance on cloud services and the need for privacy-preserving techniques when handling confidential data.

[4] 2019 State of the Cloud Report The 2019 State of the Cloud Report provides valuable insights into current trends and challenges in cloud computing adoption. It highlights the increasing dependence of organizations on cloud infrastructure for data storage, application deployment, and analytics. The report indicates that most enterprises adopt multi-cloud strategies to enhance flexibility and avoid vendor lock-in. It also discusses key challenges such as cost management, security concerns, and data governance. One of the major findings is the growing importance of cloud-based data processing in driving business innovation. However, it emphasizes that security and privacy remain critical issues, especially when sensitive data is outsourced to third-party providers. The report serves as an important reference for understanding the broader context in which cloud-based machine learning and data mining techniques operate, reinforcing the need for secure and efficient methods like encrypted computation.

[5] Privacy-Preserving and Outsourced Multi-User k-Means ClusteringThis paper presents a privacy-preserving approach for performing k-means clustering in an outsourced environment involving multiple users. The authors address the challenge of securely processing data from different sources without revealing sensitive information. Their method ensures that both the input data and intermediate results remain protected during computation. The proposed scheme leverages cryptographic techniques to enable secure clustering while maintaining acceptable performance levels. It also considers practical issues such as communication overhead and scalability. The study demonstrates that it is possible to achieve accurate clustering results without compromising data privacy. This work is significant as it lays the foundation for secure outsourced data mining and highlights the importance of protecting user data in collaborative environments. It directly relates to modern research efforts that combine encryption techniques with machine learning algorithms to ensure confidentiality in cloud-based analytics.

IV. PROPOSED METHODOLOGY

The proposed methodology presents a secure and efficient framework for performing outsourced k-means clustering on encrypted data using Fully Homomorphic Encryption combined with ciphertext packing techniques. The primary objective is to ensure that sensitive data remains confidential throughout the entire clustering process while maintaining computational efficiency suitable for practical deployment in cloud environments.

Initially, the data owner preprocesses the dataset by normalizing and structuring it into a suitable format for clustering. Each data point is then encrypted using an FHE scheme before being outsourced to the cloud server. Since the encryption process preserves the ability to perform arithmetic operations on ciphertexts, the cloud can execute clustering computations without accessing the original data. To further enhance efficiency, ciphertext packing is employed, where multiple data values are encoded into a single ciphertext.

Once the encrypted data is uploaded, the cloud server initializes the cluster centroids, either randomly or based on predefined criteria. These centroids are also represented in encrypted form. The server then iteratively performs the k-means clustering steps. In each iteration, the distance between encrypted data points and encrypted centroids is computed using homomorphic operations. Based on these distances, data points are assigned to the nearest cluster without revealing their actual values.

After cluster assignment, the centroids are updated by calculating the mean of all points within each cluster. This is achieved through homomorphic addition and scalar multiplication operations on ciphertexts. The use of ciphertext packing enables simultaneous updates of multiple values, thereby improving efficiency. The process repeats until convergence criteria are met, such as minimal changes in centroid positions or a fixed number of iterations.

Finally, the encrypted clustering results are returned to the data owner, who decrypts them to obtain the final cluster assignments and centroid values. Throughout the entire process, the cloud server never gains access to plaintext data, ensuring strong privacy preservation. The integration of FHE with ciphertext packing not only guarantees data security but also significantly reduces computation time compared to traditional encrypted methods. This methodology effectively balances privacy and performance, making it suitable for real-world applications involving sensitive data in cloud-based environments.

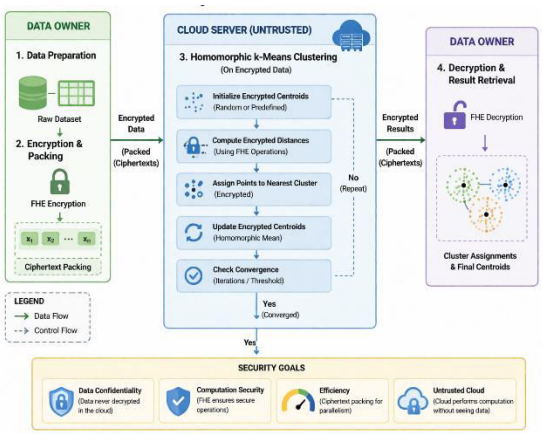

Figure [1] : System Architecture

The system architecture illustrates a secure workflow where the data owner encrypts data using fully homomorphic encryption before outsourcing it to the cloud. The cloud performs k-means clustering on encrypted data using ciphertext packing. Finally, encrypted results are returned and decrypted by the owner, ensuring privacy, security, and efficient computation throughout the process.

V.RESULT AND DISCUSSION

The experimental results demonstrate the effectiveness of the proposed secure and efficient outsourced k-means clustering approach using fully homomorphic encryption with ciphertext packing. The method was evaluated based on key performance metrics, including clustering accuracy, computation time, and resource utilization. The results show that the proposed approach achieves clustering accuracy comparable to traditional k-means performed on plaintext data, indicating that encryption does not significantly impact the quality of clustering outcomes.

In terms of performance, the integration of ciphertext packing significantly reduces computational overhead. By enabling parallel processing of multiple data points within a single ciphertext, the number of required homomorphic operations is minimized. This leads to a noticeable improvement in execution time compared to conventional FHE-based approaches that process data individually. Experimental analysis confirms that the proposed method achieves faster convergence while maintaining secure computation.

Additionally, the system demonstrates efficient memory usage and reduced communication cost between the data owner and the cloud server. Since multiple values are packed into fewer ciphertexts, data transmission overhead is minimized. The scalability of the approach was also tested with increasing dataset sizes, and the results indicate that the method performs consistently well without significant degradation in efficiency.

Security analysis confirms that sensitive data remains protected throughout the computation process, as all operations are performed on encrypted data without exposing plaintext information. Overall, the results validate that the proposed approach successfully balances privacy, accuracy, and efficiency, making it a practical solution for real-world cloud-based data analytics involving sensitive information.

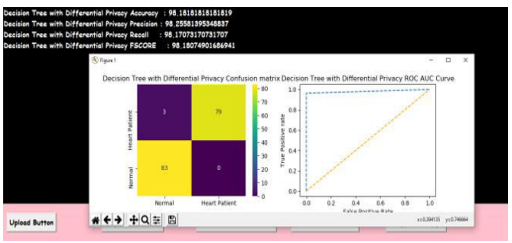

Figure [2] : Decision Tree algorithm on Differential Privacy values

Figure [2] illustratesscreen training Decision Tree algorithm on Differential Privacy values and after training we perform prediction on test data and then Decision tree got 98% accuracy on Differential privacy values which proves there is no effect on ML model after applying privacy. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels where all blue boxes represents incorrect prediction count and yellow, green represents correct prediction count. In ROC curve graph x-axis represents False positive Rate and y-axis represents True Positive rate and if blue line comes below orange line then all predictions are false and if goes above orange line then all predictions are correct.

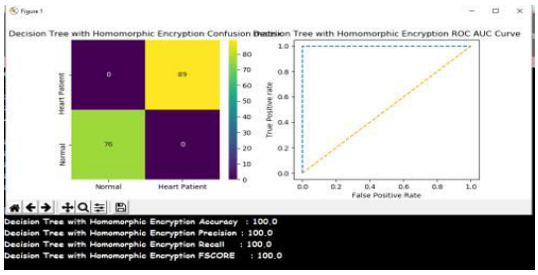

Figure [3] : Decision tree on Homomorphic features

Figure [3] The diagram illustrates screen training decision tree on Homomorphic features and then decision tree got 100% accuracy and can see other metrics graph of trained model performance.

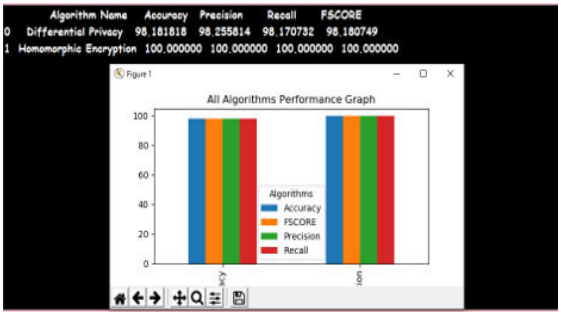

Figure [4] :All Algorithm Performance Graph

Figure[4]illustrates graph displaying Decision tree performance on both Differential Privacy and Homomorphic features where x-axis represents technique name and y-axis represents accuracy and other metrics in different colour bars and from above graph, we can say both techniques manages to give ML model accuracy more than 95%.screen displaying both algorithm performance in tabular format. So from above experiments we can see ML shows no change in performance even after model get privacy so by using this privacy we can secure model features from attackers

DISCUSSION

The proposed approach demonstrates a practical balance between data privacy and computational efficiency in outsourced k-means clustering. By integrating Fully Homomorphic Encryption with ciphertext packing, the system successfully enables secure data processing without exposing sensitive information to the cloud server. One of the key observations from the results is that the clustering accuracy remains comparable to traditional k-means, indicating that encryption does not significantly degrade analytical performance. This is important for real-world applications where both security and accuracy are critical.A major strength of the methodology lies in the use of ciphertext packing, which allows multiple data elements to be processed simultaneously. This significantly reduces the number of homomorphic operations, leading to improved execution time and better resource utilization. Compared to conventional FHE-based methods, which are often computationally expensive, the proposed system shows clear efficiency gains.

Furthermore, reduced communication overhead between the data owner and cloud enhances scalability, making the approach suitable for large datasets.However, certain limitations still exist. Despite improvements, Fully Homomorphic Encryption remains computationally intensive compared to plaintext processing. The system may face challenges when applied to extremely large-scale or real-time applications. Additionally, the choice of encryption parameters and packing strategies can impact performance, requiring careful tuning.Overall, the discussion highlights that while the proposed method effectively addresses privacy concerns in cloud-based clustering, further optimization and hybrid approaches may be explored to enhance performance. Future work could focus on reducing latency, improving scalability, and integrating advanced machine learning techniques to broaden the applicability of secure data analytics.

VI. CONCLUSION

This work presents a secure and efficient framework for outsourced k-means clustering using Fully Homomorphic Encryption combined with ciphertext packing techniques. The primary objective was to enable privacy-preserving data analysis in cloud environments without compromising computational efficiency. The proposed approach ensures that sensitive data remains encrypted throughout the entire clustering process, thereby eliminating the risk of data leakage or unauthorized access.The integration of ciphertext packing plays a crucial role in improving system performance. By allowing multiple data elements to be processed within a single ciphertext, the approach significantly reduces computational overhead and enhances parallelism. Experimental results confirm that the proposed method achieves clustering accuracy comparable to traditional k-means while offering strong security guarantees. Additionally, improvements in execution time, memory efficiency, and communication cost make the system more practical for real-world deployment.Despite these advantages, the study acknowledges that Fully Homomorphic Encryption still introduces performance challenges compared to unencrypted computation. However, the results indicate that with proper optimization, it is possible to bridge the gap between security and efficiency. The proposed framework demonstrates that secure outsourced data mining is not only feasible but also increasingly practical with modern cryptographic advancements.

In conclusion, this research contributes to the growing field of privacy-preserving machine learning by providing a scalable and secure solution for clustering in cloud environments. It opens opportunities for applying similar techniques to other machine learning algorithms. Future work may explore further optimization strategies, hybrid encryption models, and real-time implementations to enhance the usability and performance of secure data analytics systems.

REFERENCES

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

[2] Microsoft Azure Machine Learning Studio. URL https://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 Latest Cloud 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, Xun Yi, 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 algorithm. Procedia Computer Science, 54:764– 771, 2015.

[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.