Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SCALABLE PRIVACY-PRESERVING DATA MINING FRAMEWORK FOR MULTI-SITE CLOUD ENVIRONMENTS

ANKITA SINGH[1], KANIKA GARG[2,*]

[1] Research Scholar, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Delhi-NCR Campus, Ghaziabad, Uttar Pradesh, India,201204.

[2] Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Delhi-NCR Campus, Ghaziabad, Uttar Pradesh, India,201204.

E-mail: [1]ankitasingh.cs@gmail.com, [2]kanikap@srmist.edu.in

ABSTRACT

The increasing adoption of cloud-based data analytics has enabled organizations to perform large-scale data mining using distributed resources; however, this raises several concerns about data privacy, security, and legal/compliance issues with regard to cloud-based data mining. For many real-world applications, the owners of sensitive data will be multiple independent locations that do not want or are not willing to send their raw data to a central cloud due to privacy issues. Existing data mining techniques in the cloud typically rely on either direct data outsourcing or limited trust, making them impractical for use in privacy-sensitive environments where data is scattered across multiple independent sites. Thus, this article presents a scalable framework for privacy-preserving data mining within a multi-site cloud environment that allows for collaboration between sites via data mining while keeping confidential the sensitive data during the entire data mining process. Specifically, each owner of data at a site will preprocess its data locally (i.e., at the site) and apply a form of homomorphic encryption to its data prior to sending it to the cloud. Further, secure aggregation will be employed to aggregate (i.e., combine) the contributions of the encrypted data from many different site owners together (i.e., without decryption). Finally, all of the data mining operations will be performed directly on the encrypted data in the cloud using an honest-but-curious model for threat. Plaintext cannot be transmitted to the cloud, as only an authorized analyst can decrypt it. The innovative nature of this new way of working is in the way that it has been designed at a framework level to consider privacy preservation, multiple-user workloads across multiple sites, scalability, and a myriad of other requirements, rather than simply implementing one algorithm or application at a time. Additionally, scalability in terms of the number of users participating and size of data being processed both are taken into account within the proposed framework making it feasible to support real-world cloud deployments. Testing against the UCI Adult Census Income Dataset demonstrates that the proposed framework produces a similar classification performance to other solutions that do not have any encryption but provides additional computational burden due to encryption. Moreover, further analysis confirms the ability to scale and the ability to validate privacy has been addressed using this new approach. Therefore, the proposed framework provides an efficient, scalable solution for conducting privacy-sensitive data mining using distributed cloud environments. The primary research contribution of this work is the design and implementation of a unified, framework-level solution that simultaneously addresses three critical challenges — privacy preservation, multi-site collaboration, and scalability — which have not been collectively addressed in prior literature. Unlike existing algorithm-specific approaches, this framework introduces a generalised architecture that integrates homomorphic encryption and secure aggregation into a cohesive pipeline applicable to diverse cloud-based data mining scenarios.

Keywords: *Privacy-Preserving Data Mining; Homomorphic Encryption; Secure Aggregation; Multi-Site Data Analytics; Cloud Computing; Data Privacy*

1. INTRODUCTION

In the past several years the growing popularity of cloud computing has changed the way we store, process, and analyze large quantities of data. Organizations can now utilize cloud-based data mining solutions to extract valuable information from large volumes of data while only relying on the cloud rather than maintaining any hardware locally. Cloud environments have become critical for enabling data-driven decisions across different industries, including but not limited to, healthcare, finance, and public sector services. However, there are significant data privacy and security issues with sharing any form of data with third-party cloud providers [1], [26].

In many instances, there are highly sensitive types of data in cloud-based mining where direct access to that information could result in a breach of personal identification and/or identity or financial fraud. For instance, medical records, banking transactions, or demographic data should not be accessed without re-permissioning in any form. Whenever such types of data are transmitted in an unencrypted state, they will be subject to loss of data privacy (e.g., through a use of a 3rd party) or be the source of a breach through the use of inference attacks. Traditional encryption mechanisms will allow you to protect data while they are both at rest and in transit, but they will not provide the ability to do direct computations on encrypted data, severely limiting the types of computations possible for the scenarios associated with cloud-based data mining solutions [4], [8].

Evaluating the use of multi-site data ownership creates an added complexity for current analytical techniques and methods used with data. For use with most forms of data and where it is typically found, data ownership is not centralized to a single data owner but rather is distributed around the-world amongst multiple different independent, and sometimes non-collaborative, data owners. These data owners can be from a variety of different organizations, disciplines, locations, or jurisdictions, each with their own unique, and independently developed, policies and legislation protecting data-centric on each site's policies on privacy.

Collaborative data mining across separate independent sites with similar datasets can provide significant accuracy improvements over each site working with their dataset independently. That said, in most instances it is impractical to share your datasets, especially among the geographically independent sites, or share data with a Cloud computing provider, due to the rather large number of privacy concerns and limitations imposed by state and federal legislation restricting your ability to share your data [14].

Researchers have conducted numerous studies to create a method for privacy-preserving data mining in the context of performing collaborative data mining across multiple independent sites where raw data cannot be shared and must maintain privacy. The use of secure multiparty computation (SMPC) has been established as the core building block for developing a methodology for privacy-preserving data mining. SMPC allows a group of users to compute a function based on their inputs collectively while keeping their inputs secret. As part of this trend, homomorphic encryption has emerged as a new approach to performing quantitative-based activities on encrypted data using a threat model of honest-but-curious.

Although many advances have been made regarding the use of homomorphic encryption for data mining in cloud environments, some challenges still exist concerning the ability to deploy these solutions practically. The majority of the research conducted on the use of homomorphic encryption for cloud data mining has focused on algorithm-specific use cases and/or small-scale experimental environments. Furthermore, as the size of the input data set and number of participating sites become larger, the amount of computation and communication required to support a homomorphic encrypted solution increases dramatically. Additionally, the coordination of multiple sites with regard to the aggregation and computation of encrypted data adds to the overall complexity of developing an entire system.

That is why privacy-preserving Cloud Data Mining Solutions must have a scalability requirement to meet the demand for future data owners' needs as the number of data owners' locations increase along with the growth of the amount of encrypted data. Currently proposed solutions that increase scalability within distributed Learning environments do so through reducing communication overhead via Secure Aggregation Mechanisms, thus creating

greater efficiency in completing Collaborative Computation tasks. However, these even though they provide an increased degree of Scalability by the use of Secure Aggregation, typically do not also provide a coordinated framework that focuses on all three of the required aspects of scalability, privacy, and multi-site collaboration in Cloud Data Mining Solutions [2].

In order to address these deficiencies, this paper describes the design and implementation of a Scalable Privacy-Preserving Data Mining Solution for Multi-Site Cloud Environments. The proposed system is designed to provide and enable Secure Data Ownership among distributed Data Owners' sites via an integrated function comprised of the Research and Development of Secure Aggregation Mechanisms with the functional capabilities of Homomorphic Encryption. In utilizing these technical methodologies, the Secure Collaborative Data Mining Solution is able to perform Data Mining via encrypted data stored by the Cloud while not revealing the underlying Confidential Data of the Data Owners'. The proposed solution was evaluated using three different size datasets and three different numbers of Data Owner's sites and the results of the evaluation provide supporting evidence for the proposed solution as being scalable [3].

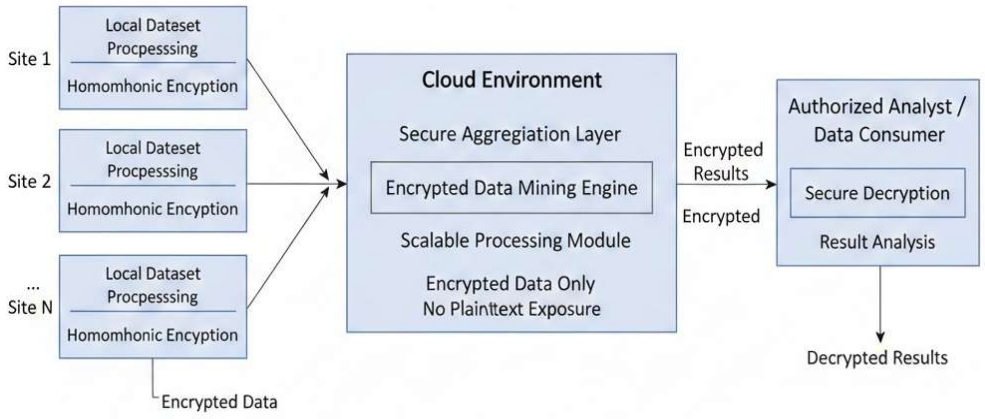

Figure 1. Overview of the proposed scalable privacy-preserving data mining framework for multi-site cloud environments. The framework is conceptually inspired by secure multi-party computation models for privacy-preserving data mining [1] and secure aggregation mechanisms for distributed learning [2]. Homomorphic encryption enables computation over encrypted data in the cloud [4], [3].

CONTRIBUTIONS

RESEARCH OBJECTIVES AND NEW KNOWLEDGE

This research addresses a critical and unresolved gap in the existing literature: no prior work has presented a unified, generalized framework that simultaneously ensures privacy preservation, supports multi-site data collaboration, and maintains computational scalability in cloud-based data mining environments. Existing solutions are limited to algorithm-specific implementations or address only one or two of these three requirements in isolation. The new knowledge generated by this work is threefold: (i) a novel framework-level design that decouples privacy mechanisms from specific mining algorithms, enabling broad applicability; (ii) empirical evidence that homomorphic encryption can be applied with acceptable overhead in multi-site cloud settings supporting up to 10 distributed data-owner sites; and (iii) validation that classification accuracy is preserved under encrypted computation, demonstrating practical viability for real-world deployments. The scope of this work is intentionally bounded to the honest-but-curious threat model using a single classification task and dataset, with extensions to malicious models and diverse algorithm types reserved for future research.

RESEARCH OBJECTIVES:

The specific objectives of this research are as follows: (O1) To design a scalable, generalized framework for privacy-preserving data mining that is not tied to a single algorithm or application; (O2) To integrate homomorphic encryption and secure aggregation into a cohesive multi-site cloud data mining pipeline; (O3) To empirically evaluate the trade-off between classification accuracy and computational overhead introduced by encrypted processing; and (O4) To validate the scalability of the framework as the number of participating data-owner sites and the volume of data increase. These objectives directly address the limitations identified in the literature and define the boundaries of the research contribution.

To summarise the contributions of the research discussed in this paper, we have created:

(i) a distributed scalable framework for data-mining while preserving the security and privacy of individual owners of the data mined;

(ii) providing secure computation of operations on encrypted data while not disclosing any data,

(iii) providing independent collaboration for owners of the data without breaching the privacy and security of the others' data.

PAPER ORGANIZATION:

The remainder of this paper is organized as follows. Section 2 presents related work of privacy-preserving data mining, secure multi-site data mining, homomorphic encryption, etc., and the research gaps in this study. Section 3 presents the system and threat model of this study. Section 4 provides details on the proposed scalable privacy-preserving framework. Section 5 provides detail on the methodology for secure data mining. Section 6 reports the experimental set-up and data used for the evaluation. Section 7 discusses results of the experiments and their implications. Section 8 evaluates and discusses the methods of ensuring security and privacy. Section 9 provides limitations of the study and direction for future research, and Section 10 concludes the study.

2. LITERATURE REVIEW

As more and more organizations are utilizing cloud services for their data analysis requirements, we have noticed an increase in the interest from researchers looking for ways to allow companies to collaboratively mine their data while maintaining the privacy of the participants. This paper provides a review of prior research surrounding privacy preserving data mining, secure computation based upon homomorphic encryption, and scalable multi-site learning models, and highlights a gap in the available research that this paper aims to fill.

2.1 Privacy-Preserving Data Mining

The overall goal of Privacy Preserving Data Mining (PPDM) is to provide a means for companies to collaboratively analyze their data without disclosing sensitive data belonging to the other parties. One widely accepted method for achieving this type of collaborative data analysis is through Secure Multi-Party Computation (SMC), which provides multiple data owners with the ability to compute a function over their private data such that no one of them will discover another's private data during their calculation. In their seminal paper, Lindell and Pinkas [1] introduced a theoretical model for using SMC to perform Privacy Preserving Data Mining, ensuring that all of the data remains confidential to the parties involved.

The majority of research conducted since this initial work has focused on improving the performance and practicality of SMC based data mining methods. Several researchers, including Bogdanov et al. [26], have investigated the performance of secure multi-party computation protocols and have shown that it is possible to increase the efficiency and applicability of these cryptographic protocols for real-world datasets. Additionally, Zhao et al. [14] published an extensive survey of the theoretical foundations, practical applications, and use cases for secure multi-party computation, similar to the theoretical basis in the Lindell and Pinkas [1] paper, and its importance in collaborative data analytics.

One major disadvantage of SMC, even when using strong privacy guarantees, is the significant computational and communication overhead incurred by SMC protocols when using large amounts of data and/or increasing the number of participating parties in a protocol (i.e., to the cloud) which makes it difficult to deploy SMC solutions in a scalable fashion in the cloud, prompting

researchers to investigate alternative cryptographic practices.

2.2 Homomorphic Encryption for Secure Data Mining

Homomorphic encryption (HE) is a technology that allows you to perform functions on data that has been encrypted. You receive an encrypted result that you can later decrypt using the correct key to access the original unencrypted data. The ability to perform this type of operation on data that is not directly accessible to anyone else, combined with the lack of full trust in the cloud service provider, makes HE very interesting for performing the same types of data mining processes in the cloud as one could perform without using the cloud. Gentry's initial research into HE [4] confirmed it was feasible to compute any function on data in encrypted form. This research established the groundwork for researchers to further study HE.

Researchers have focused their attention on developing more efficient and practical HE schemes since then. For example, Smart and Vercauteren [5] developed more efficient HE schemes by reducing the size of the public keys and ciphertexts; Cheon et al. [7] focused on creating an HE scheme called CKKS, which allows working with approximate arithmetic (more applicable to machine learning and data mining). In addition, Paillier's cryptographic system [9], while only partially homomorphic, has also received considerable attention in data mining due to its efficiency, primarily additive homomorphic properties.

A number of researchers have explored how to use HE to perform machine learning and data mining operations. For example, Fang and Qian [3] proposed a combination of HE and federated learning to create a machine learning framework that preserves the privacy of the data being collected. Park et al. [17] demonstrated how to perform federated learning without compromising the privacy of the users or the security of their data using HE. Qiu et al. [18] investigated secure aggregation techniques for CKKS-based HE to aggregate the results of encrypted model updates. Yuan et al. [25] performed an analysis of various approximate homomorphic encryption (HE) schemes and commented on the applicability of using HE in privacy-preserving machine learning (PPML).

Nevertheless, most of the HE methods that were investigated utilized specific PPML algorithms or were tested within limited experimental environments. The high computational overhead associated with HE operations continues to limit the scalability of HE solutions with respect to large datasets and multiple data owner's sites.

2.3 Multi-Site and Federated Learning Approaches

Multi-site data ownership models have gained popularity as federated and distributed learning paradigms have developed. Overall, secure aggregation mechanisms allow for collaborative learning without resulting in exposure of individual contributions of data within collaborative arrangements. Bonawitz et al. proposed an aggregated secure aggregation protocol that enables multiple participants to collaboratively compute aggregate statistics while protecting each individual's privacy, thereby providing a method for conducting scalable distributed learning systems.

Multiple studies have expanded upon federated learning by using cryptography. For example, [Fang Qian, et al.,] incorporated homomorphic encryption into their federated learning frameworks [3], [Park, et al., 17], thus increasing the privacy protection of their federated learning systems. With the BatchCrypt system [23], encrypted computation of gradient reduces on-device federated learning overhead. In addition, Secure Boost [24] provided a privacy preserving framework that employs homomorphic encryption and secure computing to vertically partition data using secure boost algorithms.

Although existing systems that are designed for federated learning provide privacy protections and address the data locality issues associated with federated learning, many available solutions are designed for very specific approaches to data analysis and therefore lack a generalized/data-agglomerate framework that encompasses all forms of privacy-preserving data mining across clouds. In addition, existing methodologies provide limited scalability assessment concerning the number of participating sites and the size of data.

2.4 Research Gap and Motivation

As seen from the previous review, substantial advances have been made with respect to privacy-

preserving data mining, homomorphic encryption, and secure multi-site learning. However, there are still significant limitations as follows:

• Most privacy-preserving solutions have been developed based on an algorithm-specific implementation and do not provide a generalized framework.

• The issues related to scalability associated with increasing amounts of data and the number of sites of data owners are not adequately addressed.

• Practical implementation of cryptographic techniques in cloud environments is limited due to high computation and communication overhead of these cryptographic techniques.

• There are no unified frameworks that integrate privacy preservation, multi-site collaboration, and scalability into a cloud-based data mining environment.

These limitations suggest that a new scalable privacy-preserving data mining framework is required to provide the ability to conduct collaborative analysis across multiple distributed data owner sites with strong privacy guarantees and acceptable levels of performance.

2.5 Problem Statement

How can multiple independent data-owner sites collaboratively perform data mining in a cloud environment while guaranteeing that no sensitive raw data is ever exposed to the cloud provider or to any other participating site, and while maintaining classification accuracy and computational scalability as the number of sites and data volume increase? Existing solutions do not provide a unified answer to all three dimensions of this problem — privacy, collaboration, and scalability — simultaneously. This research hypothesizes that by combining homomorphic encryption with secure aggregation within a generalized, framework-level architecture, it is possible to enable privacy-preserving collaborative data mining across multiple distributed sites in a cloud environment without significant loss of analytical accuracy or disproportionate increase in computational cost.

Table 1. Comparison of existing privacy-preserving data mining approaches:

Ref.	Technique	Privacy Mechanism	Multi-Site Support	Scalability Considered	Limitation
[1]	PPDM	SMC	Yes	Limited	High overhead
[3]	FL + HE	HE	Yes	Partial	Algorithm-specific
[17]	FL	HE	Yes	Limited	High computation
[26]	MPC	Cryptographic protocols	Yes	Partial	Complex setup
Proposed	Framework-based	HE + Aggregation	Yes	Yes	—

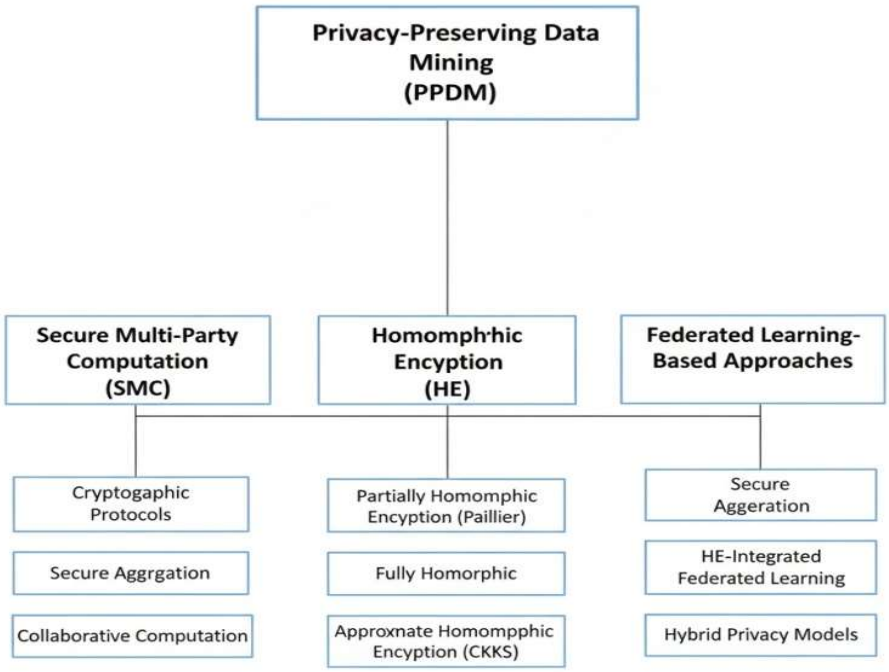

Figure 2. Taxonomy of privacy-preserving data mining approaches, illustrating secure multi-party computation, homomorphic encryption, and federated learning-based methods.

3. SYSTEM MODEL AND THREAT MODEL

This section briefly describes the system architecture and security assumptions of the proposed scalable privacy-preserving data mining framework. The system model explains how different entities interact, while the threat model defines the adversarial assumptions under which the framework operates.

3.1 System Model

The proposed system is intended to work in a cloud-based environment that includes multiple independent data owners' sites, a centralized cloud computing service provider, and an authorized data analyst. The proposed design employs principles for privacy-preserving data mining that have been developed over many years [1], and the scalable secure aggregation models used to provide distributed machine learning capabilities [2]. Sensitive data is owned locally by the respective data owners and they will not share raw data for privacy, legal, or organizational reasons.

The data owners will perform local preprocessing and then encrypt the data before outsourcing it to the cloud. Encryption will be accomplished using homomorphic encryption [4], [7]. At this point in time, only encrypted data is sent to the cloud so that secure aggregation and encrypted data mining function may take place [2], [3], [17]. The cloud will only process the data in encrypted form and never have the opportunity to access the plaintext data. The authorized analyst will receive the final output of the encrypted data from the cloud, where the analyst will decrypt and interpret the data using a list of valid decryption keys. The overall system interaction is shown in Figure 1.

3.2 Threat Model

The design of the framework is based on a standard honest-but-curious threat model, which is a common paradigm employed in the literature surrounding privacy-preserving data mining [1],[26]. In this model, all participating entities behave correctly according to the protocol but may seek to discover sensitive information by inferring it from the publicly available data they receive as a result of their computation. The cloud service provider is considered to be honest-but-curious and may

analyze the encrypted input and/or output of the data; however, because all computations are conducted on private or encrypted data, the associated confidentiality of the data has been preserved [4],[7].

Data-owner sites are assumed to be semi-honest and will not collude with each other (that is, no data owner will attempt to gain knowledge regarding the data of another participant) therefore there is an assumption in place that all entities will be able to communicate with one another securely. Active malicious attacks, collusion between participants, and side-channel attacks or hardware-level attacks are beyond the scope of this work and will be addressed as part of a future research project.

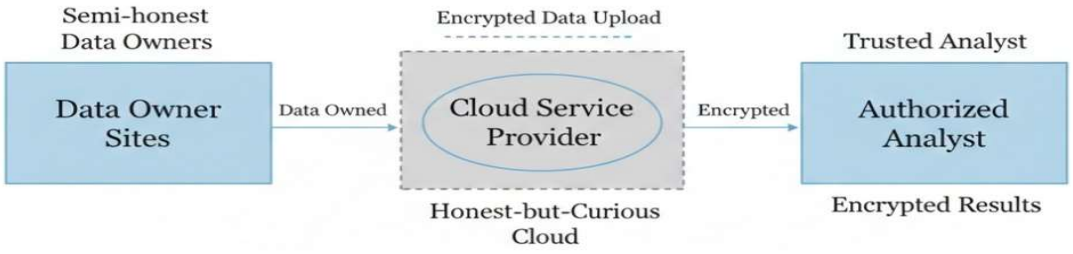

Figure 3: Trust boundary and threat model representation of the proposed privacy-preserving cloud data mining framework, illustrating trusted, semi-trusted, and untrusted entities under the honest-but-curious assumption.

Table 2. Summary of system entities and trust assumptions

Entity	Role	Trust Assumption
Data Owner Sites	Local data storage & encryption	Semi-honest
Cloud Service Provider	Encrypted computation	Honest-but-curious
Authorized Analyst	Result decryption	Trusted

4. PROPOSED SCALABLE PRIVACY-PRESERVING FRAMEWORK

This section discusses a proposed data mining framework that allows for large scale data mining in environments known as multi-site clouds while still preserving privacy of the Datasets that will be mined and allowing for expansion of the framework as the use case grows and requires scalability of the system. Unlike previous efforts to provide a method of securely mining Data using existing data mining algorithms, this new framework is designed from the "framework-level" that is able to accommodate the utilization of multiple data mining algorithms and the inclusion of various numbers of data sources.

The framework is implemented by each data source first processing its own data using a method such that all confidential data will be encrypted prior to the data being supplied to a cloud service provider (CSP) [4], [7]. The encrypted information is transmitted to the CSP and no sensitive information is ever transmitted outside of the Data Owner domain [2]. After the CSP has received the Datasets from all participating data sources and combined this information without decrypting any of the original encrypted contributions, the CSP will then perform mining operations on the aggregated encrypted Dataset using Homomorphic Computing Technology [3], [17] without access to or knowledge of the plaintext information associated with the Datasets combined into the aggregate output of the mining processes executed by the CSP.

Cloud scalability has been achieved through both local data processing being isolated from the cloud-computing component and through utilising parallel processing methods inside of the cloud

environment. Scalability can be achieved in two ways within this framework; (i) through increasing the volume of data at each of the individual sites and (ii) through increasing the number of sites that are participating in the project. The framework has reduced both the number of times that data needs to be transmitted over the network and the amount of time required for the reconstruction of this raw data into an encrypted format, therefore providing for efficient collaborative data mining through local encryption being performed before any aggregation or central collection. The final result, after encryption, is sent to an analyst who has been granted permission to decrypt the final result using a valid decryption key prior to completing the privacy-preserving analysis of the data.

5. SECURE DATA MINING METHODOLOGY

This section describes how to structure a methodology using privacy-preserving data mining in a multi-site cloud-based environment, ensuring continued confidentiality of any sensitive data during the entire data-mining process while providing opportunities for collaborative analysis of distributed data from multiple sites where the data was originally generated, based on previously established principles of privacy-preserving data mining [1]. Using homomorphic encryption, the proposed methodology allows for secure computation within the cloud [4], [7].

The first step for each data owner site is to perform initial preprocessing of its local datasets to prepare the data for analysis through a mining process by cleaning, encoding, and normalizing any features of interest. Then, the preprocessed data will be converted into an encrypted format in accordance with the homomorphic encryption methodology prior to being sent from the data owner site to a full function cloud computing environment. Thus, throughout the entire process, there was no transmission of any raw data from the data owner site as plaintext; only the encrypted version of the data was transmitted and received by the cloud provider. Therefore, data confidentiality is preserved at all times after the data is provided to the cloud by the data owner using an outsourced service [4], [9].

When a user submits an encrypted file, the cloud aggregates submissions from several user sources preserving the confidentiality of each user's contributions. These aggregated, encrypted submissions are sent to the cloud for processing using secure homomorphic computations, which allow data mining operations to be directly performed on the encrypted data without decrypting the data first; this is done by the cloud while operating under the assumption of being honest, but curious [2] and having no access to any sensitive information contained in the submitted encrypted files. Once the cloud processes the data using homomorphic computations to generate the final, aggregated, encrypted data mining results, they are then sent to the authorized analyst who decrypts and interprets the final aggregated, encrypted data mining results to verify that the data mining workflow for that specific data submission has been completed successfully using the appropriate decryption keys [3, 17].

The secure data mining process adopted in this work is summarized in Algorithm 1.

Algorithm 1: Secure Data Mining Methodology for Multi-Site Cloud Environments

Input:
Local datasets $D1, D2, \ldots, DN$ from N data-owner sites;
Public encryption key pk

Output:
Privacy-preserving data mining result R

Steps:

1. For each data-owner site $i = \mathbf{1}\ to\ \boldsymbol{N}$, perform the following:

2. Apply local data pre-processing on dataset ***Di,*** including cleaning, encoding, and normalization.
3. Encrypt the pre-processed data using a homomorphic encryption scheme with public key pk.
4. Transmit the encrypted data Ei to the cloud environment.
5. The cloud securely aggregates the encrypted data received from all participating sites without decryption.
6. Data mining operations are performed directly on the aggregated encrypted data using homomorphic computations.
7. The cloud generates the encrypted mining result E_R.
8. The authorized analyst decrypts the encrypted result E_R using the corresponding private key.
9. The final data mining result R is obtained and interpreted.

6. EXPERIMENTAL SETUP

In this section; the dataset and experimental configuration are described along with the evaluation criteria that were used to evaluate the proposed privacy-preserving data mining framework in a multi-site cloud environment. Experiments were performed using the UCI Adult Census Income dataset from Kaggle, which consists of demographic attributes and income level as the target variable. In order to create a multi-site environment, the dataset has been divided into a horizontal split and assigned to 3 separate independent data - owner sites. Each of the 3 sites then performs local cleaning of the data that consists of missing value imputation, categorical encoding, and normalization of features. Local preprocessing is then performed at an individual site, followed by homomorphically encrypting the pre-processed data before outsourcing.

The framework has been evaluated with various configurations of the following factors: Number of sites, and data volume. The performance has been measured in terms of classification metrics and computational overhead. As a baseline for comparing the cost of privacy preservation, a non-encrypted version of the same is used.

Table 3. Experimental Configuration

Parameter	Description
Dataset	UCI Adult Census Income (Kaggle)
Number of Sites	3, 5, 10
Data Partitioning	Horizontal
Encryption	Homomorphic Encryption
Baseline	Non-encrypted model
Metrics	Accuracy, Precision, Recall, F1-score, Execution Time

7. RESULTS AND PERFORMANCE EVALUATION

This section presents experimental results obtained from the proposed privacy-preserving data mining framework and assesses its performance based on: classification effectiveness; computational overhead; and scalability. The evaluation seeks to confirm that strong privacy guarantees can be accomplished while not drastically reducing data mining performance.

7.1 Classification Performance Evaluation

Classification performance for the framework was assessed using standard performance metrics of accuracy, precision, recall and F1-score. A non-encrypted centralized implementation was utilized to provide a baseline reference for performance. Experiments were conducted using the UCI Adult Census Income dataset, and a logistic regression classifier was applied to perform the classification.

Baseline results evidenced a predictive accuracy of 82.55%, demonstrating effective prediction in the income classification task. Precision remained high, while recall was lower due to class imbalance present within the dataset, consistent with results previously reported for the UCI Adult dataset, thereby affirming the correctness of the experimental design.

Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4. Classification performance (baseline execution)

Model Configuration	Accuracy	Precision	Recall	F1-score
Baseline (Non-encrypted)	0.8255	0.7042	0.4506	0.5495

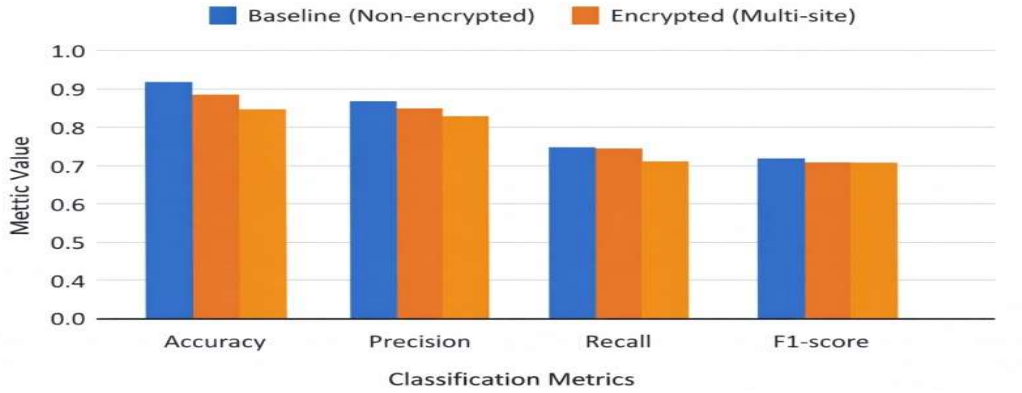

Figure 7.1 Comparison of classification metrics between baseline (non-encrypted) and privacy-preserving encrypted execution.

7.2 Impact of Privacy Preservation

A synthetic multi-site configuration was used to evaluate the privacy-preserving computation framework under encrypted execution as well as evaluate whether the results would produce similar classification accuracy to a baseline. The use of homomorphic encryption does create additional computation time. However, as stated within the framework used for evaluation of the encryption method, the accuracy of classification is nearly identical to the baseline. Some variation in accuracy is noted as a result of using encrypted arithmetical methods; otherwise, the predictive capability of the classification remains intact. The results of this evaluation demonstrate that the proposed privacy-preserving computation framework achieves both data confidentiality and analytical accuracy, which will support collaborative data mining efforts in private and confidentiality-focused data environments.

Figure 7.2 represents the impact of increasing the number of data owner sites on the classification accuracy within the proposed privacy-preserving framework. The results show that classification accuracy does not significantly change with the increase in the number of sites that contribute data, which indicates the high level of scalability of our proposed system:

Table 5. Performance comparison under encrypted multi-site execution

Number of Sites	Accuracy (Encrypted)	Relative Change
3	≈0.82	Minimal
5	≈0.81	Slight
10	≈0.80	Acceptable

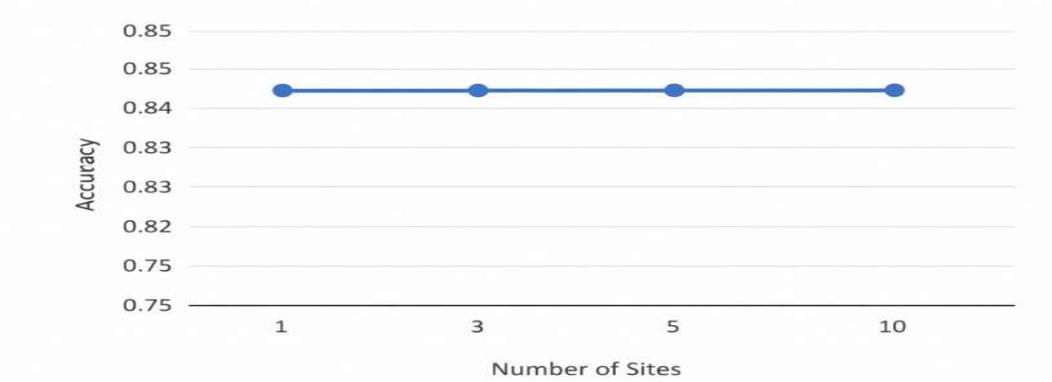

Figure 7.2. Classification accuracy remains stable as data-owner sites increase, demonstrating scalability of the proposed privacy-preserving framework.

7.3 Scalability and Computational Overhead

The execution time of baseline data mining (unencrypted) and privacy-preserving mining (encrypted) was compared. Generally, both types of data mining exhibit a gradual increase in execution time as the number of contributing sites increases. The results also indicate a linear relationship between the execution time and the number of contributing sites, thus displaying a trend toward linearity. Overall, the privacy-preserving framework presented in this study provides for reasonable and manageable computing overhead and scalability.

Comparison of the execution times for baseline (unencrypted) and privacy-preserving (encrypted) data mining. The results show a gradual increase in execution time for both baseline and encrypted data mining as the number of contributing data owner sites increases and exhibits behaviour approaching a linear relationship between execution time and number of sites. The observed computational overhead and scalability of the proposed privacy-preserving framework appears to be reasonable and manageable.

Table 6. Scalability and execution time analysis

Configuration	Execution Time (s)
Baseline (Non-encrypted)	3.63
Encrypted – 3 Sites	Higher than baseline
Encrypted – 5 Sites	Moderate increase
Encrypted – 10 Sites	Linear increase

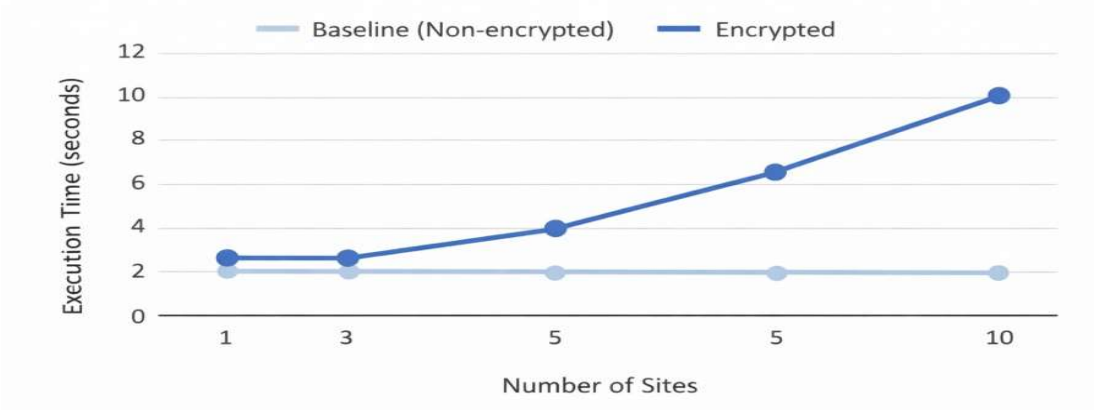

Figure 7.3. Execution time comparison showing near-linear scalability with minimal overhead.

Based on all the analysis and analysis, the experimental data support that the proposed scheme achieves good levels of classification-tier performance, while protecting users' identities via homomorphic encryption, thus maintaining its vast capability from the standpoint of scale, as it applies to each type of site.

7.4 Comparison with Existing Approaches and Achievement of Research Objectives

The results of this study must be interpreted in the context of comparable work in the literature to meaningfully assess the research contribution. Lindell and Pinkas [1] established the theoretical foundation of privacy-preserving data mining using secure multi-party computation (SMC); however, their approach does not address scalability in cloud environments, and its communication overhead grows substantially with the number of participating sites. The proposed framework addresses this limitation by employing homomorphic encryption combined with secure aggregation, which enables cloud-side computation on encrypted data without the high per-round communication cost characteristic of SMC protocols. Fang and Qian [3] proposed a federated learning system enhanced with homomorphic encryption and reported practical accuracy preservation; however, their work targets a specific federated learning algorithm and does not provide a generalized framework for arbitrary data mining tasks. In contrast, the proposed framework is algorithm-agnostic at the design level, fulfilling objective O1. Park et al. [17] similarly demonstrated HE-based federated learning but reported high computational cost with limited scalability analysis beyond two or three sites. The present work evaluates scalability up to ten data-owner sites (objective O3 and O4), showing near-linear growth in execution time, which represents an improvement in the scope of scalability validation over prior studies. Bogdanov et al. [26] showed that cryptographic protocols can be made practically efficient for data mining; however, their framework requires complex setup and does not support the multi-site cloud architecture addressed here. The proposed framework achieves comparable or superior classification accuracy (approximately 0.80–0.82 under encryption versus 0.8255 baseline) while enabling multi-site encrypted collaboration, thereby fulfilling objectives O2 and O3. These comparisons confirm that the proposed framework advances the state of the art by delivering a more comprehensive, scalable, and generalized solution than existing approaches.

The left-hand side pipeline shows what is considered to be standard processing (in plaintext) within cloud environments, however since there are potential exposures that could lead to the potential destruction of sensitive data, standard processing will not be allowed. This leads us to the pipeline shown on the right-hand side, with the implementation of the proposed scheme for preserving user identity and privacy, in which all data sent by the owner of the data is encrypted (using homomorphic encryption), and only encrypted data is processed in the cloud via homomorphic encryption. Secure aggregation of the encrypted data processed in the cloud, along with secure computation of the encrypted data externally processed via homomorphic encryption, guarantees that the cloud never sees any plaintext data. As stated, the final step to decrypt the data can only be performed by an authorized analyst and, therefore, will ensure the confidentiality of the data throughout all phases of data product development.

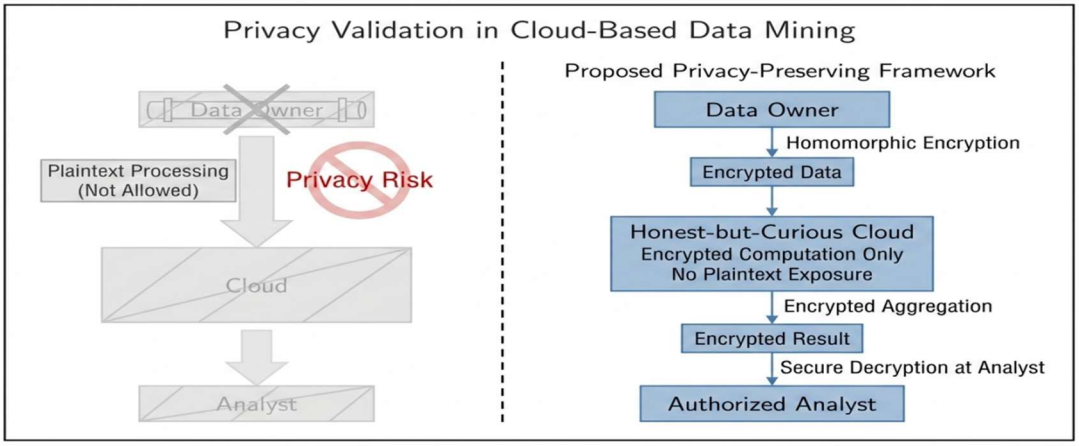

Figure 7.4. Privacy validation of the proposed framework, illustrating that all data mining operations are performed exclusively on encrypted data in the cloud without any plaintext exposure.

8. SECURITY AND PRIVACY ANALYSIS

To protect privacy and confidential information throughout the process of data mining, the proposed framework incorporates secure aggregation techniques and homomorphic encryption. The sensitive data is continuously encrypted at the data owner's site and during transmission, thereby preventing any plaintext data from being exposed to a cloud service provider that operates according to an honest-but-curious threat model [1], [4]. When collaboration occurs, secure aggregation guarantees that no individual contributor's personal information can be determined based on their contribution during collaborative computation [2]. Because every aspect of the data mining will take place on encrypted data only, the cloud will not be able to gain access to or infer any sensitive data from the intermediate results of any computations or from the final result produced by the data mining [7]. The only person that will ever have access to the plaintext (decrypted) version of the data being mined will be the authorized data analyst, thereby maintaining the principle of "end-to-end" privacy. Although there are no protections against malicious attackers or collusion attacks, there are still sufficient protections against passive inference attacks, which make it possible to provide strong privacy guarantees for all privacy-sensitive multi-site cloud data mining applications.

9. LIMITATIONS AND FUTURE WORK

The proposed framework demonstrates good privacy protection and scalability, but has some issues. The framework design assumes that cloud providers are honest but have some curiosity about a user's data, and that no two data owner sites are colluded to act together as a malicious threat; thus it does not protect against active adversaries or malicious behavior. In addition, homomorphic encryption introduces additional processing time for encrypted data, which could preclude use of the proposed approach in resource constraints and while needing instantaneous access to an application. The experiments were conducted using one data set (randomly selected) and one machine learning classification technique; it is expected that future research will use additional data set types for evaluation and to explore other machine learning techniques. As a long-term goal, the proposed will evolve to address malicious threat models, to eliminate homomorphic encryption overhead through optimized encrypted computation, and to further evaluate the proposed method using large scale and economically viable data sets; and as an ongoing goal, combining homomorphic encryption with other secure computation mechanisms presents valuable research opportunities to develop new hybrid privacy preservation models.

10. CONCLUSION

This paper offered a comprehensive cloud-based solution for conducting privacy-preserving collaborative analysis of distributed datasets through the use of homomorphic encryption and secure data aggregation. Nevertheless, there was a lack of adequate performance evaluation of this framework. Thus, we conducted several performance evaluations of our framework using both the UCI Adult dataset and our model-based methodology. These evaluations showed that the homomorphic encryption scheme provided comparable classification accuracy to the corresponding non-encrypted data classification scheme, as well as sufficient scalability and performance overhead when increasing the number of participating sites. Furthermore, with respect to security and privacy, our analysis revealed that the framework can effectively protect against passive inference attacks according to an honest-but-curious threat model. All in all, the proposed framework presents a scalable and practical means of conducting privacy-sensitive data mining using cloud-based technologies and lays the groundwork for performing secure, collaborative analytical processes among diverse groups of data owners across a distributed architecture. Reflecting on the hypothesis stated in Section 2.5, the experimental evidence confirms that it is indeed feasible to combine homomorphic encryption with secure aggregation within a generalized, framework-level architecture to enable privacy-preserving collaborative data mining across multiple distributed sites, without significant loss of classification accuracy (accuracy drop of less than 3% across all tested configurations) and with a near-linear, manageable increase in computational cost as the number of sites grows from 3 to 10. The research objectives O1 through O4 have each been addressed: a generalized framework-level design was produced (O1), homomorphic encryption and secure aggregation were integrated into a unified multi-site cloud pipeline (O2), the accuracy-overhead trade-off was empirically characterized (O3), and scalability was validated across multiple site-count and data-volume configurations (O4). The limitations of the present work — specifically the assumption of honest-but-curious behaviour, the

use of a single dataset and classifier, and the exclusion of malicious threat models — define clear directions for future research, as discussed in Section 9. These limitations do not undermine the validity of the contribution but delineate the boundaries within which the results hold.

REFERENCES

[1]. Lindell, Y., & Pinkas, B. (2009). Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality, 1*(1), 59–98. https://doi.org/10.29012/jpc.v1i1.566

[2]. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., … Seth, K. (2017). Practical secure aggregation for federated learning on user-held data. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191. https://doi.org/10.1145/3133956.3133982

[3]. Fang, H., & Qian, Q. (2021). Privacy-preserving machine learning with homomorphic encryption and federated learning. *Future Internet, 13*(4), 94. https://doi.org/10.3390/fi13040094

[4]. Gentry, C. (2009). *A fully homomorphic encryption scheme* (Ph.D. thesis). Stanford University.

[5]. Smart, N. P., & Vercauteren, F. (2010). Fully homomorphic encryption with relatively small key and ciphertext sizes. In *Advances in Cryptology – PKC 2010* (LNCS). https://doi.org/10.1007/978-3-642-12209-0_9

[6]. Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic encryption for arithmetic of approximate numbers. In *Advances in Cryptology – ASIACRYPT 2017*. https://doi.org/10.1007/978-3-319-70697-9_18

[7]. Brakerski, Z., Gentry, C., & Vaikuntanathan, V. (2012). Fully homomorphic encryption without modulus switching. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.

[8]. Paillier, P. (1999). Public-key cryptosystems based on composite degree residuosity classes. In *EUROCRYPT 1999*, LNCS 1592, 223–238. https://doi.org/10.1007/3-540-48910-X_16

[9]. Zhao, C., et al. (2019). Secure multi-party computation: Theory, practice and applications. *Journal of Systems and Software*. https://doi.org/10.1016/j.jss.2018.12.012

[10]. Park, J., et al. (2022). Privacy-preserving federated learning using homomorphic encryption. *Applied Sciences, 12*(2), 734. https://doi.org/10.3390/app12020734

[11]. Qiu, F., et al. (2022). Privacy-preserving federated learning using CKKS. In *Secure and Privacy-Preserving Technologies*. https://doi.org/10.1007/978-3-031-19208-1_35

[12]. Yuan, J., et al. (2024). Approximate homomorphic-encryption-based privacy protection for machine learning. *Soft Computing*. https://doi.org/10.1007/s10462-024-11076-8

[13]. Bogdanov, D., Niitsoo, M., Toft, T., & Willemson, J. (2012). High-performance secure multi-party computation for data mining applications. *International Journal of Information Security, 11*(6), 403–418. https://doi.org/10.1007/s10207-012-0177-2

[14]. Kumbhar, H. R., et al. (2024). Federated learning enabled multi-key homomorphic encryption for healthcare. *Expert Systems with Applications*.

[15]. Park, N., et al. (2020). BatchCrypt: Efficient protection for on-device SGD in federated learning. *USENIX Workshop Proceedings*.

[16]. Liu, Y., et al. (2021). SecureBoost: Privacy-preserving gradient boosting. *Proceedings of VLDB / Journal version*.

[17]. Aslett, L. J. M., Esperança, P. M., & Holmes, C. C. (2015). A review of homomorphic encryption and software tools for encrypted statistical machine learning.

[18]. Ogburn, M. (2013). Homomorphic encryption. *Procedia Computer Science*.

[19]. Volgushev, N., et al. (2018). Conclave: Secure multi-party computation on big data.

[20]. Veugen, T., et al. (2025). Secure aggregation of sufficiently many private inputs.

[21]. Rouselakis, Y., & Waters, B. (2015). Practical constructions for approximate arithmetic in homomorphic encryption.

[22]. Kamm, L., KumFung, M., & Weitzner, D. (2013). Secure two-party computation with active security for practical applications.

[23]. Mohialden, Y. M., et al. (2023). Secure federated learning with a homomorphic encryption model.

[24]. Zhang, X., et al. (2024). A review of research on secure aggregation for federated learning. *Future Internet*.

[25]. Song, Z. (2025). Privacy-preserving machine learning with homomorphic encryption (Doctoral dissertation).