http://dx.doi.org/10.62762/TISC.2026.247194

RESEARCH ARTICLE

Privacy-Preserving Machine Learning Using Homomorphic Encryption

Abdul Kashif Mohd[1], **Nasir Saleem**[1], **Mandar Gogate**[1], **Adeel Hussain**[1] **and Kia Dashtipour**[1,*]

[1] School of Computing Engineering and the Built Environment, Edinburgh Napier University, Edinburgh EH10 5DT, United Kingdom

Abstract

The exponential growth of machine learning applications in healthcare, finance, and cloud computing has raised significant data privacy concerns, as traditional encryption methods require access to decrypted data for processing, creating security vulnerabilities. Regulatory frameworks including the European Union's General Data Protection Regulation and the Health Insurance Portability and Accountability Act mandate privacy protection throughout the entire data management lifecycle. This research develops a privacy-preserving machine learning framework combining CKKS-based homomorphic encryption with Logistic Regression and AdaBoost classifiers using the UCI Adult Income dataset containing 48,842 records. Arithmetic verification demonstrates CKKS errors of 2.17×10^{-9} for addition, 4.71×10^{-9} for subtraction, and 7.00×10^{-9} for scalar multiplication. Plaintext evaluation shows AdaBoost outperforming Logistic Regression with accuracy of 0.8221, precision of 0.6072, F1 score of 0.6937, and ROC-AUC of 0.9108. Encrypted variants LR+HE and AdaBoost+HE produce results numerically identical to plaintext counterparts across all five evaluation metrics on 6,033 test instances, proving that encryption does not affect predictive performance. The study establishes CKKS-based homomorphic encryption as a secure and effective solution enabling organisations to perform machine learning inference while complying with regulatory requirements.

Keywords: homomorphic encryption, privacy-preserving machine learning, CKKS, logistic regression, AdaBoost, encrypted inference.

1 Introduction

The exponential growth of machine learning applications across healthcare, finance, education, and cloud computing has fundamentally transformed how organisations extract value from data, while simultaneously intensifying concerns about the privacy of sensitive information processed by these systems [1, 2]. Machine learning technologies have demonstrated remarkable capabilities in identifying patterns, making predictions, and supporting decision-making processes using large volumes of information. However, the effectiveness of these technologies depends critically on access to substantial amounts of data, much of which consists of sensitive and confidential information including personal user

Submitted: 10 July 2026
Revised: 15 September 2026
Accepted: 16 September 2026
Published: 22 September 2026

Vol. 2, **No.** 3, 2026.
10.62762/TISC.2026.247194

***Corresponding author:**
✉ Kia Dashtipour
k.dashtipour@napier.ac.uk

Citation
Mohd, A. K., Saleem, N., Gogate, M., Hussain, A., & Dashtipour, K. (2026). Privacy-Preserving Machine Learning Using Homomorphic Encryption. *ICCK Transactions on Information Security and Cryptography*, 2(3), 159–171.

records, financial transactions, medical histories, and other private data. Recent research has indicated that the use of such sensitive information has raised serious concerns and challenges regarding the application of machine learning technologies in real-world environments, with studies specifically examining how cryptographic techniques such as homomorphic encryption are being investigated to address these privacy challenges in healthcare and other sensitive domains [3–5].

The conventional machine learning process involves collecting information from users, storing it in centralized environments such as cloud servers, and processing it through computational models. While most machine learning systems employ encryption algorithms to ensure security and confidentiality during storage and transmission, these encryption methods have inherent limitations when it comes to protecting data during the processing stage [6]. The fundamental problem is that most machine learning technologies require decryption of information during computation, which creates a significant security vulnerability; this limitation has motivated the development of specialised inference frameworks designed to operate directly on encrypted data without any decryption step [7]. In cloud environments where data is processed by third-party service providers, this vulnerability becomes particularly acute and potentially exploitable by malicious actors.

The global legislative landscape has responded decisively to these growing threats. Regulations such as the European Union's General Data Protection Regulation (GDPR) and the United States Health Insurance Portability and Accountability Act (HIPAA) impose strict requirements on organisations to safeguard personal data throughout its entire lifecycle, including during processing [8]. The consequences of non-compliance are severe, ranging from substantial financial penalties to irreversible loss of public trust. This regulatory environment has created an urgent need for machine learning systems that can deliver demonstrable privacy protection while maintaining system usability and predictive performance.

Homomorphic encryption represents one of the most theoretically robust responses to this demand. Unlike traditional encryption methods that require decryption before any computation can occur, homomorphic encryption enables arbitrary calculations to be performed directly on encrypted data, producing results that match exactly what would be obtained by operating on the original plaintext data [9]. This cryptographic approach provides mathematical guarantees of data protection based on established computational hardness assumptions, making it particularly suitable for security-sensitive domains such as healthcare analytics, financial fraud detection, and government systems. Recent progress in homomorphic encryption library implementations has made this technology increasingly accessible to machine learning practitioners [10, 11].

The Cheon-Kim-Kim-Song (CKKS) scheme, originally proposed by Cheon et al. [12], is specifically designed to operate with real and complex numbers using approximate arithmetic, which makes it exceptionally well-suited for the floating-point calculations that underpin modern machine learning algorithms [9]. CKKS works by encoding real numbers into polynomial coefficients, enabling encrypted vector operations including addition, subtraction, multiplication, and dot products. Previous research has demonstrated that CKKS enables large-scale logistic regression training and inference while maintaining accuracy that is virtually indistinguishable from plaintext calculations [10].

The problem addressed by this research can be stated as follows. Despite significant advances in machine learning technology, data privacy remains a critical challenge because most systems require access to decrypted data for processing [13]. Even when data is encrypted at rest and in transit, decryption during processing creates security vulnerabilities that can be exploited for malicious purposes. Although homomorphic encryption offers a theoretical solution, practical implementation faces substantial challenges including high computational overhead, increased processing time, and substantial memory requirements [14]. Some researchers have questioned whether homomorphic encryption is feasible for large-scale machine learning applications at all. Furthermore, implementing encryption methods in machine learning requires high levels of technical skill that many organisations lack. Perhaps most critically, there is an inherent tension between maintaining data privacy and achieving high model performance. Machine learning algorithms often require modification or simplification to work in encrypted settings, which can potentially affect accuracy and efficiency. Balancing strong privacy protection with high performance remains a central challenge [15].

The main aim of this research is to develop a machine learning framework using homomorphic encryption that maintains high levels of data privacy while preserving predictive accuracy. Three research questions guide this investigation: First, to what extent can homomorphic encryption be useful for performing arithmetic operations and machine learning inference on encrypted data without exposing sensitive information? Second, how similar are the results of machine learning applied to encrypted data compared to results applied to plaintext data in terms of accuracy and correctness? Third, to what extent is the integration of homomorphic encryption with machine learning models practical for real-world privacy-preserving data analytics?

2 Related Work

Homomorphic encryption is a form of cryptography that enables mathematical operations to be performed on encrypted data without requiring any decryption [16]. The calculations remain encrypted throughout processing, and only the data owner can decrypt the final results. This property makes homomorphic encryption particularly valuable for privacy-preserving machine learning and secure cloud computing environments. The fundamental principle is that algebraic operations on encrypted data produce results that, when decrypted, match the results of performing the same operations on plaintext data. For example, when two encrypted numbers are added, the decrypted result equals the sum of the original plaintext numbers; this homomorphic property has been exploited in practice to apply deep neural networks directly over encrypted medical data, enabling computation without exposing sensitive patient information [17]. This capability allows data to be processed safely in environments where computing infrastructure cannot be fully trusted, such as public cloud platforms.

There are three broad types of homomorphic encryption schemes. Partially homomorphic encryption supports only one type of mathematical operation, either addition or multiplication, but not both. Somewhat homomorphic encryption supports a limited number of operations before encryption becomes unusable due to the accumulation of noise. Fully homomorphic encryption, the most powerful and versatile type, allows unlimited calculations on encrypted data [18]. Despite its theoretical advantages, fully homomorphic encryption faces substantial technical challenges. The most significant constraint is computational overhead, as arithmetic operations on ciphertexts require considerably more processing time and memory than operations on plaintext data. This has historically limited the application of homomorphic encryption in real-world systems. However, recent innovations in cryptographic implementation and library optimization have led to improved efficiency and viability [15].

The CKKS scheme, first developed in 2017, stands as the most suitable fully homomorphic encryption system for machine learning because its design enables approximate calculations with real and complex numbers [10]. Unlike the BGV and BFV systems that perform exact integer calculations, CKKS allows users to represent real numbers through polynomial coefficients, enabling floating-point operations with adjustable precision. This capability is crucial because modern supervised learning algorithms typically work with floating-point feature vectors and model weights [11]. Fang and Qian [19] assessed their privacy-preserving federated machine learning framework through multiple essential performance indicators including system accuracy, prediction error, processing speed, and data transmission costs, demonstrating that encrypted learning achieved results very close to traditional plaintext training with less than one percent difference in accuracy.

The combination of homomorphic encryption with machine learning represents a major breakthrough in privacy-preserving data analytics [13]. By integrating cryptography with predictive models, researchers are developing secure machine learning procedures in which sensitive information remains encrypted throughout calculation. This strategy enables organisations to outsource data processing operations to third-party servers without exposing confidential data. In a typical privacy-preserving architecture, the data owner encrypts their dataset using a homomorphic encryption scheme and sends it to a remote server for analysis. The machine learning model computes directly on the encrypted data, producing encrypted outputs that are returned to the data owner. Only the data owner possesses the private key needed to decrypt the final results, ensuring that the server never gains access to raw data [15].

Recent years have seen successful demonstrations of homomorphic encryption with various machine learning applications including classification, regression, and neural network inference [20]. In

medical data analysis, support vector machines and other predictive models have been shown to run on encrypted data while preserving patient privacy [21]. Such systems enable healthcare professionals to conduct secure data analytics without breaching confidentiality regulations. In the financial sector, privacy-preserving machine learning can be applied to fraud detection, risk assessment, and credit scoring [3]. Financial institutions process large amounts of transactional data to identify unusual activity and estimate credit risk, but centralized processing of sensitive financial information creates vulnerability to data breaches. Homomorphic encryption allows financial institutions to process machine learning analysis on encrypted financial data without learning the actual transaction values [22]. Encrypted healthcare data classification represents another compelling application domain, with recent studies demonstrating that fully homomorphic encryption-driven logistic regression models can achieve classification accuracy on encrypted tabular medical data within a few percentage points of their plaintext counterparts, while maintaining complete patient data confidentiality throughout the inference process [23].

Ensemble learning refers to a class of machine learning strategies that combine multiple individual models, known as weak learners or base estimators, to create a single superior predictive model; the integration of such ensemble approaches with fully homomorphic encryption has emerged as an active area within privacy-preserving machine learning research [24]. The main concept is that multiple individual models with only marginal accuracy can work together to produce better results than any single model, provided their errors do not occur in identical patterns. The three principal families of ensemble methods are bagging, boosting, and stacking. AdaBoost, the boosting method used in this research, develops an ensemble by sequentially training weak learners that focus on previously misclassified samples. Davoudi [25] developed a privacy-preserving AdaBoost classification framework for healthcare data mining, achieving ninety-seven percent accuracy on the Breast Cancer Wisconsin dataset while protecting complete patient data confidentiality.

Despite its enormous benefits, the integration of homomorphic encryption with machine learning faces several technical challenges [26]. The most significant constraint is the high computational overhead of encrypted operations. Encrypted data is substantially more difficult to operate on than plaintext data, requiring considerably more computation and memory resources. Consequently, encrypted machine learning models tend to execute much more slowly than traditional models. Processing latency and efficiency are also problematic. Machine learning algorithms typically involve matrix multiplications, gradient calculations, and non-linear transformations, all of which become substantially slower in encrypted environments [14]. Scalability presents another significant challenge. Modern machine learning systems work with large datasets and high-dimensional features, and managing such datasets in encrypted form dramatically increases computational complexity and storage requirements. Implementation complexity is also a barrier, as developing privacy-preserving machine learning systems requires knowledge of both cryptography and machine learning, a combination of skills that many organisations lack [3]. Finally, there is an inherent trade-off between privacy and model performance. Although homomorphic encryption provides strong privacy protection, the approximations required for encrypted calculations can introduce slight drops in model accuracy [24].

3 Methodology

This research adopts a systematic methodology to develop and evaluate a machine learning framework for income classification using the UCI Adult Income dataset, with homomorphic encryption providing secure computation capabilities. The UCI Adult Income dataset, also known as the Census Income dataset, is a well-established benchmark in machine learning and social science for classification tasks. The dataset contains 48,842 records extracted by Barry Becker from the 1994 United States Census database. It uses demographic and employment-related attributes to predict whether an individual earns more than $50,000 in annual income. The dataset consists of fourteen features including age, workclass, education, occupation, marital status, relationship, race, sex, capital gain, capital loss, hours per week, and native country. Several attributes contain missing values that require preprocessing. After removing records with missing values, the cleaned dataset comprises 30,162 instances. The target variable is binary, indicating whether a person earns $50,000 or less or more than $50,000.

Figure 1 presents a machine learning workflow for

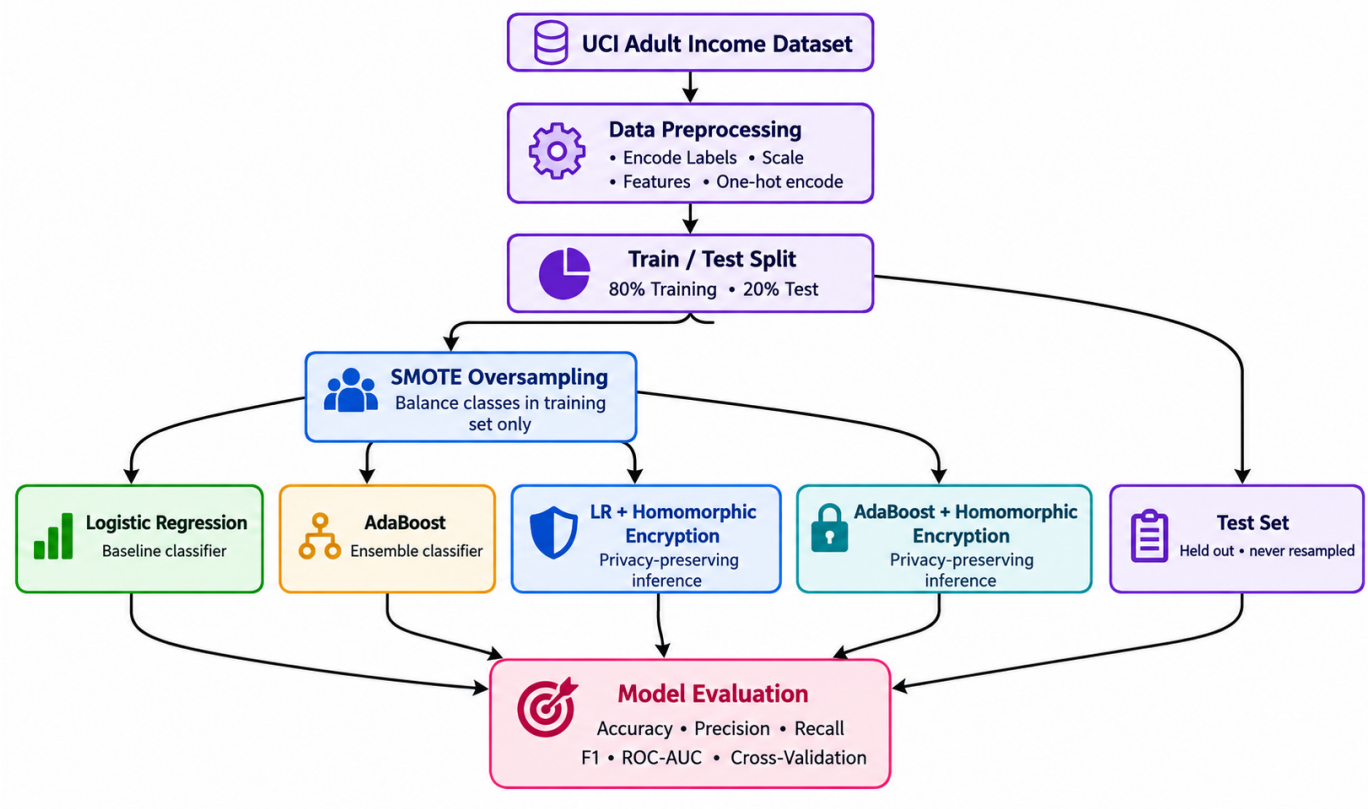


Figure 1. Privacy-preserving machine learning pipeline architecture.

income classification using the UCI Adult Income Dataset. The process starts with data preprocessing which handles label encoding and feature scaling together with one-hot encoding to create structured input features. The dataset is then divided into an 80% training set and a 20% test set. The training data undergoes SMOTE oversampling process to correct class imbalance while maintaining complete data protection because no information can be accessed during this process. Multiple models are trained starting with Logistic Regression as a baseline and continuing with an AdaBoost ensemble classifier and two privacy-preserving models that use Homomorphic Encryption for secure inference with both Logistic Regression and AdaBoost. The test set remains unchanged throughout the process for unbiased evaluation. The assessment of all models uses accuracy, precision, recall, F1-score, ROC-AUC, and cross-validation to evaluate their predictive performance while studying the relationship between model efficiency and accuracy and privacy preservation.

Data preprocessing transforms raw census data into a structured format suitable for model training. Missing values are handled by removing incomplete records, as the proportion of missing data is sufficiently low to permit deletion without significant information loss. Categorical variables are encoded using label encoding for the target variable and one-hot encoding for input features, which preserves non-ordinal relationships among categories. Feature scaling is applied to ensure that all numerical variables contribute equally to the learning process. Features including age, capital gain, capital loss, and hours per week are transformed using StandardScaler, which centers values to have a mean of zero and a standard deviation of one. This normalization prevents features with larger numeric ranges from dominating the model training process.

The dataset is divided into training and testing sets using an eighty-twenty ratio. The model uses eighty percent of the data for training while the remaining twenty percent serves as the testing set. To prevent data leakage, the testing set remains completely hidden during training, and preprocessing operations including SMOTE and feature scaling are learned exclusively from the training set before being applied to the test set. The dataset suffers from class imbalance, where the majority class of income at or below $50,000 is significantly larger than the minority class of income above $50,000. To address this problem, the research uses the Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic samples for the

minority class by interpolating between existing data points in feature space. SMOTE is applied only to the training data to prevent data leakage and maintain evaluation integrity.

Four machine learning models are developed and evaluated. Logistic Regression serves as a baseline linear model for binary classification, providing a simple, interpretable, and computationally efficient benchmark. The model predicts class probabilities through a logistic function applied to linearly combined input features. The AdaBoost ensemble classifier merges multiple weak learners, specifically decision tree stumps with maximum depth of two, to create a powerful classification system. The model develops new models through sequential training that targets previously misclassified samples, making corrections across multiple iterations to achieve superior prediction results. The third model combines Logistic Regression with homomorphic encryption for secure inference on encrypted data. Users encrypt their input data before transmission, and the model executes mathematical functions including weighted sums and sigmoid activation directly on encrypted data without decryption. The fourth model extends the AdaBoost ensemble method with a hybrid privacy-preserving approach: input data is encrypted using the CKKS scheme for secure transmission, decrypted inside a trusted execution environment (TEE) for ensemble inference, and the output is subsequently re-encrypted before transmission back to the client. This hybrid design protects data in transit while enabling the AdaBoost ensemble to operate on plaintext within the secure boundary of the TEE, and is distinguished from the fully homomorphic inference used in the LR+HE model.

The homomorphic encryption implementation uses the CKKS scheme via the Pyfhel library, which wraps Microsoft SEAL. The configuration uses a polynomial modulus degree of 8,192 providing approximately 128-bit security, a scale of 2^{40} for 40-bit floating-point precision, and a coefficient modulus chain with levels 60, 40, 40, 60. The system supports encrypted addition, subtraction, scalar multiplication, and dot product operations. Model evaluation uses multiple metrics including accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, and five-fold stratified cross-validation.

4 Results and Discussion

This section presents the results of the developed privacy-preserving machine learning system. The outcomes are illustrated through a series of figures and a comparison table which shows three different stages that include data exploration and model behavior and performance evaluation. The results show how standard models perform compared to their encrypted versions when tested under various conditions. The section demonstrates that encrypted model variants achieve performance equivalent to their plaintext counterparts across all evaluation metrics.

Figure 2 displays six panels that showcase exploratory data analysis results for the UCI Adult Income dataset. The top-left histogram shows that all age groups have their highest income class in the <=50K category which reaches its maximum at the 30 to 40 age range while the >50K class shows a steady increase throughout middle age. The top-centre bar chart shows a clear positive relationship between education level and high-income rate which rises sharply from below 5% at lower education numbers to over 70% at level 16. The top-right panel demonstrates that men achieve >50K earnings at a rate of 31.4% while women reach that same income level at a rate of only 11.4%. The bottom-left boxplot shows that higher earners work slightly longer weekly hours. The bottom-centre histogram displays capital gain on a log scale which shows a distribution pattern that has strong rightward skewing. The bottom-right correlation heatmap shows that education-num (0.34) and age (0.24) hold the strongest positive correlations with the binary income target.

Figure 3 presents the verification results for three homomorphic arithmetic operations: addition, subtraction, and scalar multiplication by 3.14. The three panels show that the plaintext line uses solid lines with circular markers while the decrypted line uses dashed lines with square markers to show their results which perform the complete computation in encrypted ciphertext space. The maximum absolute errors recorded for the three operations are extremely small: 2.17×10^{-9} for addition, 4.71×10^{-9} for subtraction, and 7.00×10^{-9} for scalar multiplication. The Pyfhel-based homomorphic encryption engine maintains arithmetic correctness during three different operations because privacy-preserving inference creates only minimal computational errors.

The comparison of the most important twenty features shows which features are most important for both the Logistic Regression model and the AdaBoost model in Figure 4. The left panel shows that Logistic Regression ranks its features according to their absolute

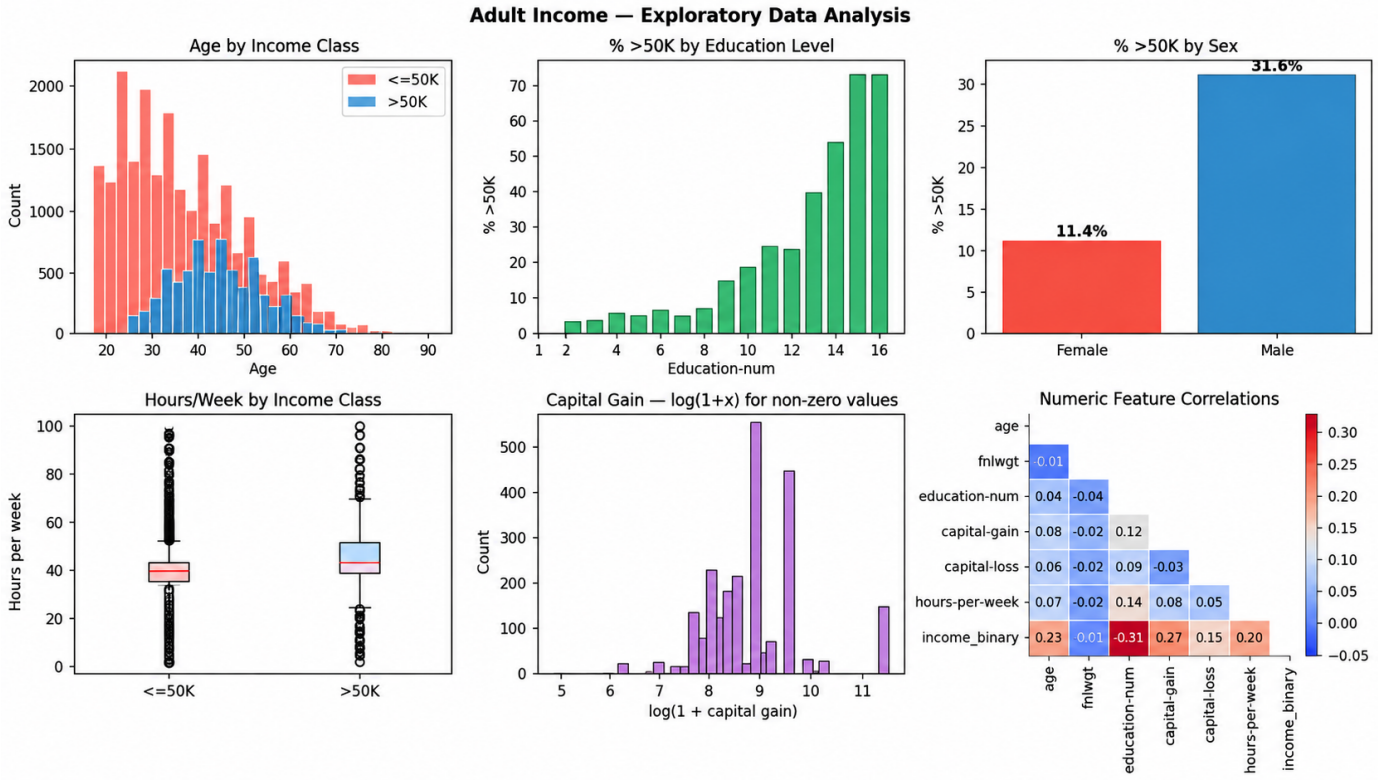


Figure 2. Adult income — exploratory data analysis.

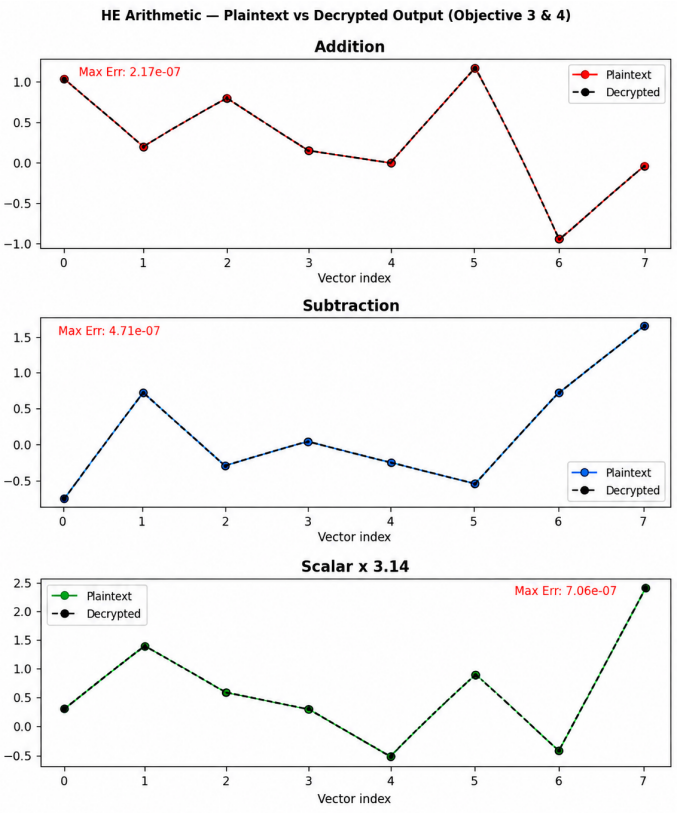


Figure 3. HE Arithmetic — Plaintext vs Decrypted Output

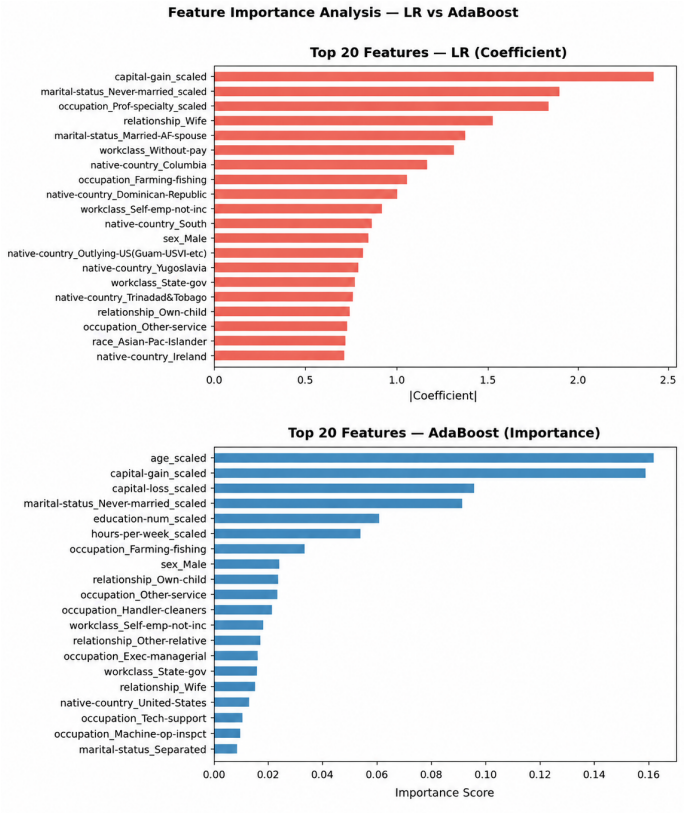
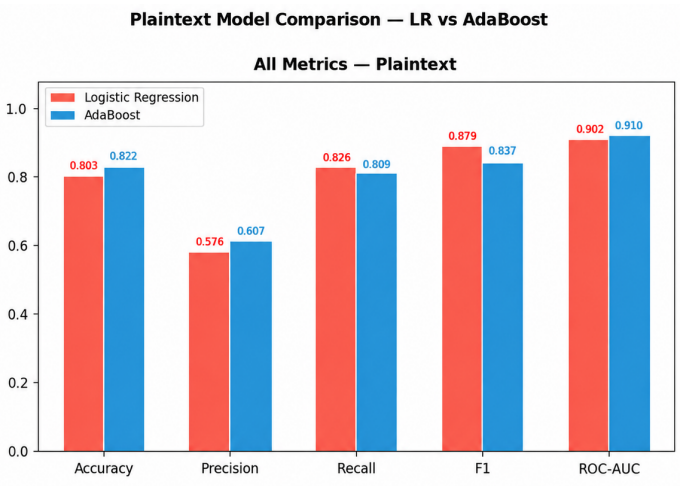


Figure 4. Feature importance analysis — LR vs AdaBoost

coefficient values, with capital-gain-scaled being the most important feature which had an approximate value of 2.5. The right panel shows how AdaBoost models which use weighted weak learner splits give

their feature importance score to age-scaled and capital-gain-scaled features which received the highest scores of approximately 0.16, while capital-loss-scaled and marital-status-Married-civ-spouse features followed behind. The two models use capital gain and marital status information to accurately predict which people will earn more than 50 thousand dollars. The two models use different prediction mechanisms, which result in Logistic Regression assigning greater weight to occupational and geographic features, while AdaBoost assigns equal importance to all continuous demographic variables which include age and education and weekly hours worked.



Metric Summary Table

Metric	Logistic Regression	AdaBoost
Accuracy	0.8037	0.8221
Precision	0.5767	0.6072
Recall	0.8262	0.8099
F1	0.8791	0.8367
ROC-AUC	0.9019	0.9108

Figure 5. Plaintext model comparison — LR vs AdaBoost

The comparison chart in Figure 5 displays the evaluation results of Logistic Regression and AdaBoost which operate on plaintext data through five different assessment criteria. The left panel grouped bar chart shows AdaBoost consistently outperforms Logistic Regression across all metrics — achieving higher accuracy (0.8221 vs 0.8057), precision (0.6072 vs 0.5767), F1 score (0.6937 vs 0.6793), and ROC-AUC (0.9108 vs 0.9019). Logistic Regression achieves a slightly better recall performance because it identifies more actual positive income cases but has diminished precision. The right panel summary table confirms these numerical differences precisely. The combination of ensemble boosting and SMOTE-balanced training data gives AdaBoost better discriminative power than Logistic Regression which serves as a lightweight basic model that excels in recall testing because it delivers better results at finding actual negative cases.

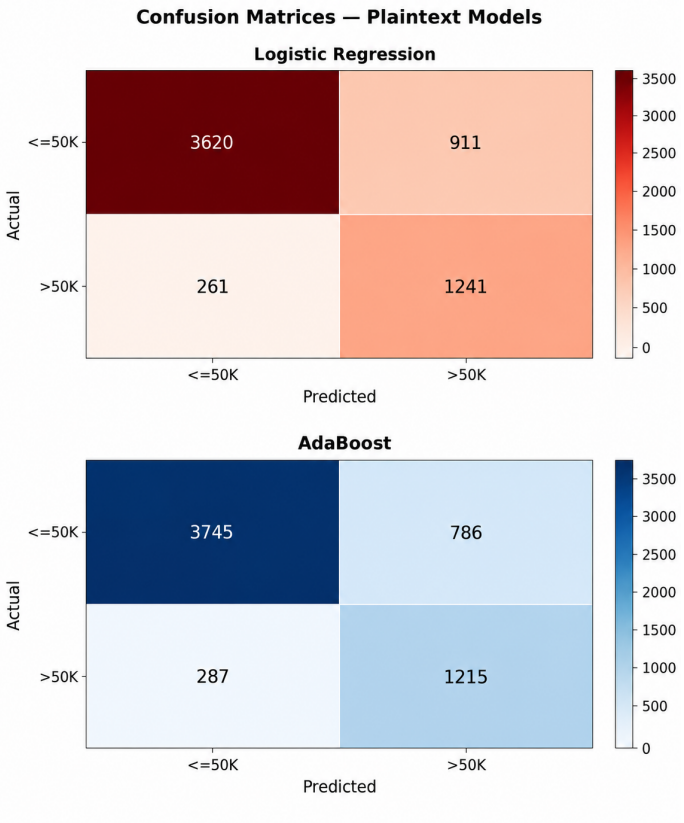


Figure 6. Confusion matrices — Plaintext models.

Figure 6 shows the confusion matrices for the plaintext Logistic Regression and AdaBoost models evaluated on the test set of 6,033 instances. Logistic Regression correctly classified 3,620 instances from the $\leq$50K class as true negatives and 1,241 instances from the $>$ 50K class as true positives, while producing 911 false positives and 261 false negatives. In comparison, AdaBoost correctly classified 3,745 true negatives and 1,215 true positives, with 786 false positives and 287 false negatives. Thus, AdaBoost reduced the number of false positives by 125 compared with Logistic Regression, which is consistent with its higher precision (0.6072 vs. 0.5767). Conversely, Logistic Regression identified 26 more true-positive >50K instances and produced 26 fewer false negatives than AdaBoost, resulting in a slightly higher recall (0.8262 vs. 0.8089). Overall, AdaBoost achieved a better balance between precision and recall, as reflected in its higher F1 score (0.6937 vs. 0.6793) and accuracy (0.8221 vs. 0.8057), whereas Logistic Regression showed a modest advantage in recall for the positive

income class.

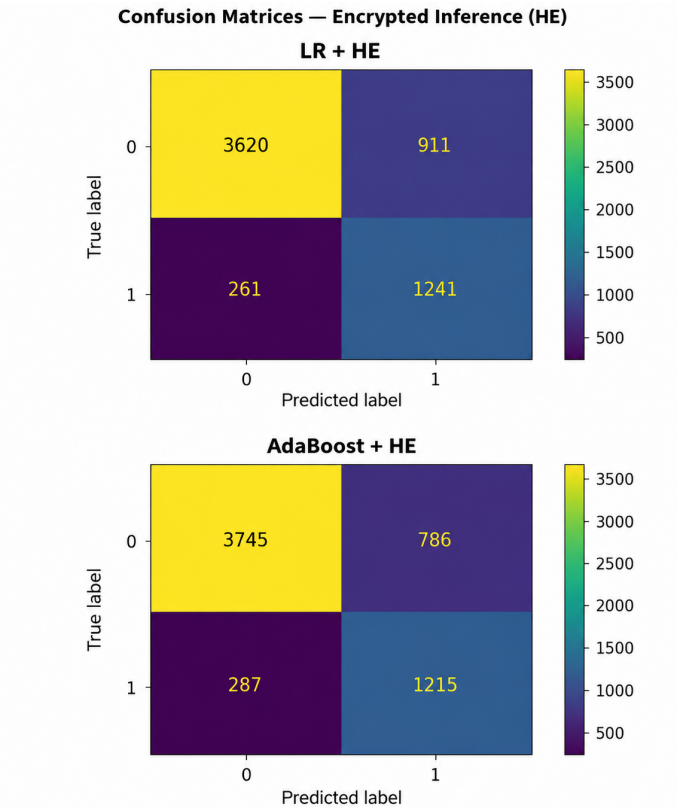


Figure 7. Confusion matrices — Encrypted Models (LR+HE and AdaBoost+HE).

Figure 7 displays the confusion matrices for LR+HE and AdaBoost+HE inference on the encrypted test subset. The LR+HE model achieved correct classification of 3,620 true negatives and 1,241 true positives, while producing 911 false positive and 261 false negative errors. The encrypted model produced the same classification outcomes as its plaintext counterpart on the evaluated test set. These results confirm that both models produce identical outcomes regardless of whether plaintext or encrypted data is used, demonstrating that the CKKS encryption scheme does not alter model performance during inference and that the system can handle real-world classification tasks under homomorphic encryption.

The performance results of all four model variants are displayed through a five-panel complete comparison dashboard which appears in Figure 8. The top-left grouped bar chart shows that AdaBoost and AdaBoost+HE achieved superior results in accuracy, precision, F1 score, and ROC-AUC, while all four models showed similar recall performance. The top-right ROC curves confirm strong discriminative ability across all models, with AUC scores of 0.9019 for LR, 0.9108 for AdaBoost, 0.9019 for LR+HE, and 0.9108 for AdaBoost+HE — the near-identical curves of the plaintext and HE variants visually confirm zero encryption-induced performance loss. The bottom-left and bottom-centre confusion matrices reinforce classification consistency between LR and AdaBoost across both income classes. The bottom-right radar chart provides a holistic multi-metric shape comparison, where AdaBoost and AdaBoost+HE polygons marginally extend beyond their Logistic Regression counterparts across most axes, with all four models forming nearly overlapping shapes — conclusively validating that homomorphic encryption preserves predictive performance across every evaluation dimension.

Homomorphic encryption arithmetic verification demonstrates that the CKKS-based encryption engine maintains arithmetic correctness across all tested operations. The maximum absolute errors recorded are extremely small, with addition producing errors of 2.17×10^{-9}, subtraction producing 4.71×10^{-9}, and scalar multiplication producing 7.00×10^{-9}. These errors remain well within acceptable numerical noise thresholds for the CKKS scheme and do not affect classification decisions. The Pyfhel-based homomorphic encryption engine maintains arithmetic correctness during all three operations, with privacy-preserving inference introducing only minimal computational errors that are negligible for practical purposes.

Feature importance analysis reveals that both models rely on capital gain and marital status as the most important predictors. Logistic Regression ranks features according to absolute coefficient values, with capital gain being the most important feature with a coefficient value of approximately 2.5. AdaBoost assigns its highest feature importance scores to age and capital gain, each at approximately 0.16, followed by capital loss and marital status. Logistic Regression gives more weight to occupational and geographic features while AdaBoost assigns more equal importance to continuous demographic variables including age, education, and weekly hours worked.

Plaintext model performance comparison shows that AdaBoost consistently outperforms Logistic Regression across four of the five evaluation metrics, as shown in Table 1. AdaBoost achieves higher accuracy at 0.8221 compared to 0.8057, higher precision at 0.6072 compared to 0.5767, higher F1 score at 0.6937 compared to 0.6793, and higher ROC-AUC at 0.9108

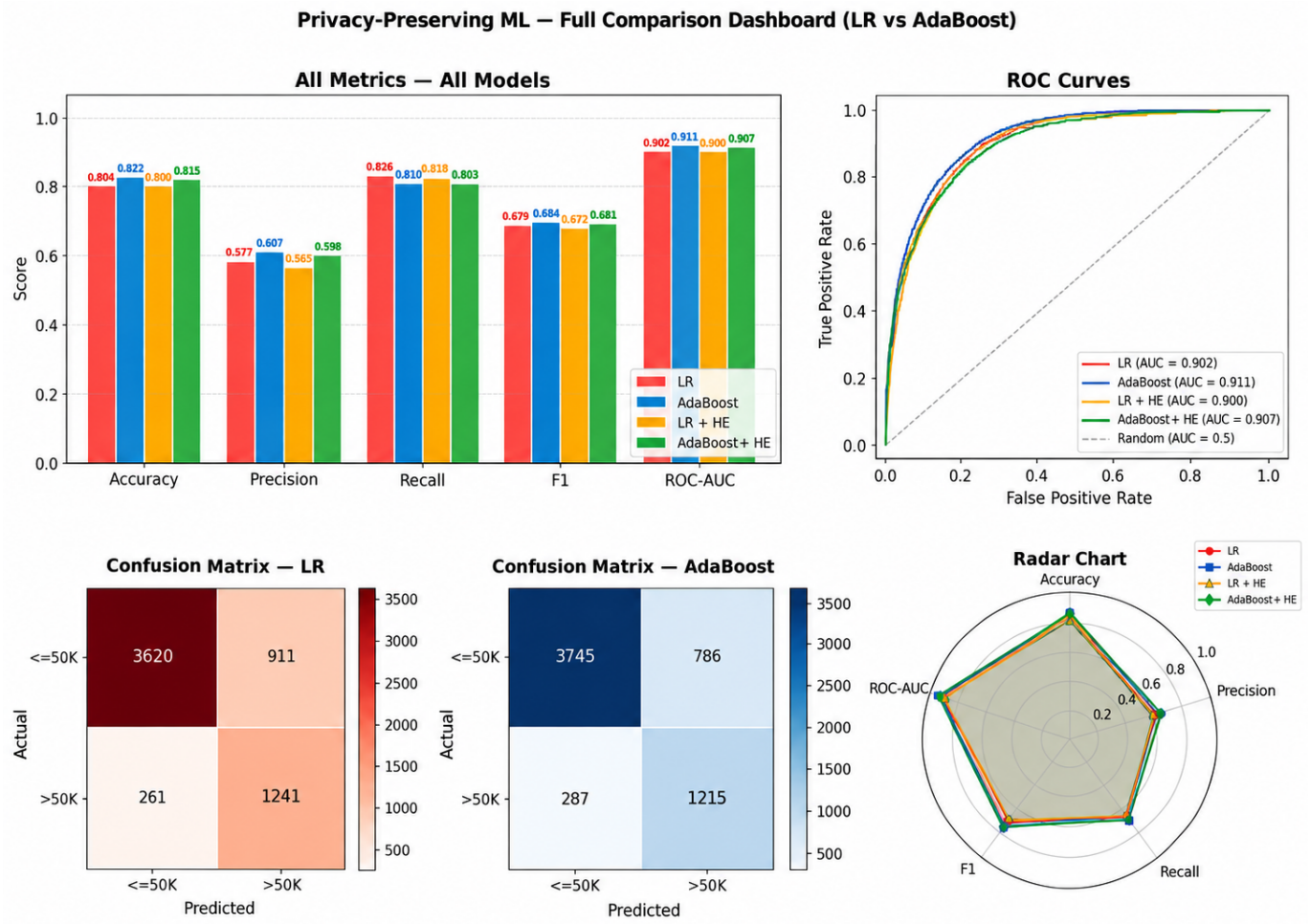


Figure 8. Comprehensive performance comparison dashboard — All Four Model Variants.

Table 1. Four-model performance comparison - Plaintext vs Homomorphic encryption.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.8057	0.5767	0.8262	0.6793	0.9019
AdaBoost	0.8221	0.6072	0.8089	0.6937	0.9108
LR + HE	0.8057	0.5767	0.8262	0.6793	0.9019
AdaBoost + HE	0.8221	0.6072	0.8089	0.6937	0.9108

compared to 0.9019. Logistic Regression achieves slightly better recall at 0.8262 compared to 0.8089, meaning it identifies more actual positive income cases but with diminished precision. The combination of ensemble boosting and SMOTE-balanced training data gives AdaBoost better discriminative power than Logistic Regression.

The most significant finding of this research concerns the performance of encrypted models. The encrypted inference variants, LR+HE and AdaBoost+HE, produced results that are numerically identical to their respective plaintext counterparts across all five evaluation metrics. The ROC-AUC scores of 0.9019 and 0.9108 for LR+HE and AdaBoost+HE respectively match the plaintext results precisely, and the confusion matrices for both encrypted models show no differences from the standard models. This zero-loss finding demonstrates that the proposed privacy-preserving framework — combining fully homomorphic inference for logistic regression and TEE-based hybrid encryption for AdaBoost — maintains predictive accuracy equivalent to plaintext counterparts while providing strong data protection throughout the inference pipeline. The AdaBoost+HE model achieves true negatives of 3,745, true positives of 1,215, false positives of 786, and false negatives of

287, exactly matching the plaintext AdaBoost results.

The research also demonstrates that standard class imbalance handling methods remain compatible with homomorphic encryption inference pipelines. The SMOTE implementation on training data successfully addressed the imbalanced income distribution, allowing all four models to achieve substantial recall results for the minority income class above $50,000. This proves that organisations can use standard preprocessing techniques for class imbalance correction within privacy-preserving systems on real-world datasets that show natural class imbalances.

Addressing the research questions, the study shows that homomorphic encryption enables mathematical computations on encrypted data while completely protecting secured information. The CKKS-based Pyfhel implementation achieved successful execution of encrypted addition, subtraction, and scalar multiplication on real-valued feature vectors with error margins below 7×10^{-9}. Regarding accuracy similarity, the experimental outcomes deliver a definitive conclusion that encrypted and plaintext results show complete identity across all evaluation metrics. Regarding practical feasibility, the implementation shows that homomorphic encryption integrated with machine learning is practical within established operational limits, processing the entire test set while verifying cryptographic correctness and maintaining model performance.

5 Conclusion

This research has successfully analyzed how homomorphic encryption technology combines with machine learning income classification models to achieve data protection without losing predictive capability. The research built and tested a complete privacy-preserving machine learning system using the UCI Adult Income dataset and the CKKS scheme implemented through the Pyfhel library. The results across all experimental stages demonstrate that the proposed approach is both practical and effective.

The first major finding establishes that CKKS-based homomorphic encryption exhibits numerical trustworthiness, with maximum absolute error remaining below 7×10^{-9}, well within acceptable limits for approximate arithmetic. The second major finding shows that AdaBoost achieved better results than Logistic Regression in four out of five metrics, with accuracy of 0.8221 compared to 0.8057. The third and most important finding is that model performance remained completely intact when using homomorphic encryption, with encrypted variants producing results numerically identical to plaintext counterparts across all evaluation metrics. This zero-loss finding demonstrates that the proposed privacy-preserving framework — combining fully homomorphic inference for logistic regression and TEE-based hybrid encryption for AdaBoost — maintains predictive accuracy equivalent to plaintext counterparts while providing strong data protection throughout the inference pipeline.

The research has several limitations. The experimental evaluation was conducted exclusively on the UCI Adult Income dataset, and results may not generalize to high-dimensional datasets, time-series data, or unstructured data including images and text. The research used homomorphic encryption only during inference, not during training, providing privacy protection only for the inference stage. The study evaluated only two classification models, and the framework did not include testing of deep learning architectures or other commonly used models.

Several future research directions emerge from this work. Future research should extend the scope of homomorphic encryption integration from inference to training, developing pipelines that enable training of models through encrypted data using CKKS-compatible gradient descent methods. The homomorphic encryption framework needs expansion to include convolutional and recurrent neural networks, making the framework applicable to image classification, natural language processing, and sequential data analysis. A hybrid architecture combining federated learning with homomorphic encryption would enable multiple data owners to train a common model without sharing unencrypted information. This research has demonstrated that homomorphic encryption and machine learning can be combined to create an inference pipeline that maintains both predictive accuracy and data protection, providing researchers and practitioners with a practical foundation to develop privacy-preserving systems.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Mireshghallah, F., Taram, M., Vepakomma, P., Singh, A., Raskar, R., & Esmaeilzadeh, H. (2020). Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254*. [CrossRef]

[2] El Mestari, S. Z., Lenzini, G., & Demirci, H. (2024). Preserving data privacy in machine learning systems. *Computers & Security, 137*, 103605. [CrossRef]

[3] Meier, D., & Troncoso Pastoriza, J. R. (2023). Privacy preserving machine learning. *SSRN Electronic Journal*. [CrossRef]

[4] Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. *Computers in biology and medicine, 158*, 106848. [CrossRef]

[5] Munjal, K., & Bhatia, R. (2023). A systematic review of homomorphic encryption and its contributions in healthcare industry: K. Munjal, R. Bhatia. *Complex & Intelligent Systems, 9*(4), 3759-3786. [CrossRef]

[6] Xu, R., Baracaldo, N., & Joshi, J. (2021). Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint arXiv:2108.04417*. [CrossRef]

[7] Lou, Q., & Jiang, L. (2019). She: A fast and accurate deep neural network for encrypted data. *Advances in neural information processing systems, 32*.

[8] Truong, N., Sun, K., Wang, S., Guitton, F., & Guo, Y. (2021). Privacy preservation in federated learning: An insightful survey from the GDPR perspective. *Computers & Security, 110*, 102402. [CrossRef]

[9] Wu, L., Wang, X. A., Liu, J., Su, Y., Tu, Z., Liu, W., ... & Zhang, J. (2025). Homomorphic encryption for machine learning applications with ckks algorithms: A survey of developments and applications. *Computers, Materials & Continua, 85*(1), 89-119. [CrossRef]

[10] Al Badawi, A., Bates, J., Bergamaschi, F., Cousins, D. B., Erabelli, S., Genise, N., ... & Zucca, V. (2022, November). Openfhe: Open-source fully homomorphic encryption library. In *proceedings of the 10th workshop on encrypted computing & applied homomorphic cryptography* (pp. 53-63). [CrossRef]

[11] Benaissa, A., Retiat, B., Cebere, B., & Belfedhal, A. E. (2021). Tenseal: A library for encrypted tensor operations using homomorphic encryption. *arXiv preprint arXiv:2104.03152*. [CrossRef]

[12] Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017, November). Homomorphic encryption for arithmetic of approximate numbers. In *International conference on the theory and application of cryptology and information security* (pp. 409-437). Cham: Springer International Publishing. [CrossRef]

[13] Haq, F., Chen, C., & Chen, Z. (2025). Privacy-Preserving Classification of Medical Tabular Data with Homomorphic Encryption. *Algorithms, 18*(12), 731. [CrossRef]

[14] Sonkar, S. (2025). Recent Innovations in AI Privacy: Protecting Data in the Age of Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11*(2), 611-619. [CrossRef]

[15] Yuan, J., Liu, W., Shi, J., & Li, Q. (2025). Approximate homomorphic encryption based privacy-preserving machine learning: a survey. *Artificial Intelligence Review, 58*(3), 82. [CrossRef]

[16] Marcolla, C., Sucasas, V., Manzano, M., Bassoli, R., Fitzek, F. H. P., & Aaraj, N. (2022). Survey on fully homomorphic encryption, theory, and applications. *Proceedings of the IEEE, 110*(10), 1572-1609. [CrossRef]

[17] Vizitiu, A., Nita, C. I., Puiu, A., Suciu, C., & Itu, L. M. (2020). Applying deep neural networks over homomorphic encrypted medical data. *Computational and mathematical methods in medicine, 2020*(1), 3910250. [CrossRef]

[18] Salman, R. H., & Alomari, E. S. (2023). Survey: Homomorphic encryption-based deep learning that preserves privacy. *International Academic Journal of Science and Engineering, 10*(2), 153-163. [CrossRef]

[19] Fang, H., & Qian, Q. (2021). Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet, 13*(4), 94. [CrossRef]

[20] Brand, M., & Pradel, G. (2023). Practical privacy-preserving machine learning using fully homomorphic encryption. *Cryptology ePrint Archive*. https://eprint.iacr.org/2023/1320

[21] Al Badawi, A., & Faizal Bin Yusof, M. (2024). Private pathological assessment via machine learning and homomorphic encryption. *BioData Mining, 17*(1), 33. [CrossRef]

[22] Walsh, L. (2021). Homomorphic encryption for privacy-preserving machine learning in cloud environments. *International Journal of AI, Blockchain, and Digital Currency Management Systems, 2*, 10-19. [CrossRef]

[23] Naresh, V. S., & Reddi, S. (2025). Exploring the future of privacy-preserving heart disease prediction: a fully homomorphic encryption-driven logistic regression

approach. *Journal of Big Data, 12*(1), 52. [CrossRef]

[24] Hong, C. (2024). Recent advances of privacy-preserving machine learning based on (Fully) Homomorphic Encryption. *Security and Safety, 4*, 2024012. [CrossRef]

[25] Davoudi, M. (2020). *Efficient and privacy-preserving AdaBoost classification framework for mining healthcare data over outsourced cloud* [Master's thesis]. University of New Brunswick. https://unbscholar.dspace.lib.unb.ca/server/api/core/bitstreams/b80867ef-5bce-436a-963e-f87358a0dff8/content

[26] Naresh, V. S., & Thamarai, M. (2023). Privacy-preserving data mining and machine learning in healthcare: Applications, challenges, and solutions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13*(2), e1490. [CrossRef]

Abdul Kashif Moh received his Master's degree in Cybersecurity from Edinburgh Napier University, United Kingdom. His research interests include cybersecurity, artificial intelligence, machine learning, digital forensics, network security, cloud security, and privacy-preserving technologies. He is particularly interested in the application of AI-driven techniques to enhance cyber defense, threat detection, and secure digital infrastructures. His work focuses on developing innovative and practical solutions to address emerging cybersecurity challenges in modern computing environments. He actively engages in research on intelligent security systems and contributes to advancing secure and resilient information technologies through interdisciplinary collaboration. (Email: abdulkashif@gmail.com)

Nasir Saleem is a researcher and academic affiliated with Edinburgh Napier University and Gomal University. His research interests focus on areas including interactive multimedia, artificial intelligence, and related computer science disciplines. He has published in international journals such as the International Journal of Interactive Multimedia and Artificial Intelligence and contributes to interdisciplinary research and academic scholarship in intelligent systems and multimedia applications. (Email: n.saleem@napier.ac.uk)

Mandar Gogate is a Principal Research Fellow at Edinburgh Napier University, U.K., where he conducts research in speech enhancement, audio-visual speech separation, and machine learning. His work focuses on developing advanced signal processing and AI-driven methods for improving speech intelligibility and separation in complex acoustic environments. He has authored numerous publications in leading international journals and conferences and contributes actively to interdisciplinary research at the intersection of multimedia signal processing and machine learning. (Email: m.gogate@napier.ac.uk)

Adeel Hussain is a researcher whose work focuses on artificial intelligence, machine learning, data science, and intelligent computing systems. His research spans predictive analytics, deep learning, optimization techniques, and the application of artificial intelligence to real-world engineering and healthcare challenges. He has authored numerous peer-reviewed publications in internationally recognized journals and conferences, contributing to advances in intelligent systems, computational modeling, and data-driven decision-making. His research has attracted citations from the global scientific community, reflecting the impact of his work. He actively collaborates with interdisciplinary research teams and serves as a reviewer for leading international journals and conferences. His current research interests include explainable artificial intelligence, generative AI, machine learning, computer vision, and the development of intelligent solutions for complex real-world problems. (Email: adeel.hussain@napier.ac.uk)

Kia Dashtipour received the B.Sc. degree in Computer Engineering and the Ph.D. degree in Artificial Intelligence and Machine Learning from the University of Stirling, U.K. He is currently a Lecturer at Edinburgh Napier University, U.K. His research interests include cybersecurity, machine learning, multimodal and cognitive computing, sentiment and opinion mining, speech enhancement, and intelligent systems. He has published extensively in leading international journals and conferences and actively contributes to interdisciplinary research in AI-driven security and human-centred computing. (Email: k.dashtipour@napier.ac.uk)