frontiers in Blockchain

TYPE Original Research
PUBLISHED 06 November 2026
DOI 10.3389/fbloc.2026.1941258

Check for updates

OPEN ACCESS

EDITED BY
Hossein Abroshan,
Anglia Ruskin University School of Computing and Information Science, United Kingdom

REVIEWED BY
Swapna Radha,
University of Notre Dame, United States
Balasubramanian V.,
Rajalakshmi Engineering College, India
Esra'A. Alkafaween,
Mutah University College of Science, Jordan
Sunila Ahmad,
Quaid-i-Azam University, Pakistan

*CORRESPONDENCE
Sanjay Agal,
sanjay.agal32685@paruluniversity.ac.in

RECEIVED 18 July 2026
REVISED 13 October 2026
ACCEPTED 15 October 2026
PUBLISHED 06 November 2026

CITATION
Agal S (2026) A machine learning framework for privacy-preserving financial risk assessment using explainable artificial intelligence and blockchain-based trust management.
Front. Blockchain 9:1941258.
doi: 10.3389/fbloc.2026.1941258

COPYRIGHT
© 2026 Agal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A machine learning framework for privacy-preserving financial risk assessment using explainable artificial intelligence and blockchain-based trust management

Sanjay Agal*

Department of Artificial Intelligence and Data Science, Parul University, Vadodara, Gujarat, India

Introduction: Financial risk assessment requires models that jointly satisfy privacy, explainability, and decentralized trust, yet existing machine-learning approaches usually address these requirements in isolation.
Methods: This paper proposes the Privacy-Preserving Explainable Trust Framework (PP-ETF), integrating federated learning with differential privacy, post hoc SHAP and LIME explanations, and a permissioned distributed ledger for trust management. PP-ETF is evaluated on the German Credit and HELOC datasets against statistical models, centralized ensembles and deep networks, and privacy-preserving baselines.
Results: PP-ETF outperforms privacy-preserving baselines in AUC-ROC while remaining within 1.4%–1.6% of centralized deep-learning models. Reputation-based adversarial detection excludes a malicious participant within two communication rounds with a 3.2% false-positive rate. Explanation stability yields average cosine similarity of at least 0.87, and the ledger processes up to 480 transactions per second with 2.4 s finality latency.
Discussion: The results support the feasibility of jointly implementing privacy, interpretability, and decentralized trust under the evaluated conditions; they do not imply that the underlying trade-offs are eliminated. PP-ETF provides an end-to-end blueprint for effective, scalable, private, interpretable, and trustworthy risk-management systems.

KEYWORDS
algorithmic transparency, blockchain, Byzantine fault tolerance, collaborative learning, consortium banking, credit scoring, data governance, decentralized trust

1 Introduction

The financial system in its most generic form operates as a system for the distribution of capital in the economy and for the function of organizations providing financial services. Risk models for credit scoring, the prediction of loan defaults, and fraud detection currently form the bulk of the risk models that are utilized. All these models have traditionally been based on logistic regression and, in terms of interpretability, on less powerful rule-based expert systems. To exploit the highest dimensions of complex information for the prediction of outcomes, which in addition to individual information (e.g., social media activity) are also provided by numerous organizational information sources (e.g., transaction history), require the

powerful models of machine learning (ML) and, in particular, deep learning (DL) that are capable of automatically learning from very large amounts of data and of discovering complex patterns that allow for the assessment of risk of default by individuals and organizations (Dumitrescu et al., 2021; Gunnarsson et al., 2021).

However, the wider application of ML to financial risk assessment is currently hindered by three interconnected challenges that affect the management of financial risk in today's distributed data world.

The first challenge relates to privacy violations. Many financial risk models rely on sensitive personal and commercial data which are subject to legislative protection, including the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in California. In order to train a ML model, all data must be collected in a central location and then used to train the model. There is a high value to be placed on such data and it is a high value target for attack, either from within an organization or as part of an adversarial attack. In addition, even unintentional disclosure of data could result in a data leak. In order to address these challenges, use of privacy-preserving techniques (PPTs) such as Federated Learning (FL) and Differential Privacy (DP) is recommended. FL allows ML models to be trained on distributed data without requiring all the data to be collected in a central location. Instead, the model is trained locally on each computer, and model updates are shared with other computers, rather than the raw data.

The second challenge hindering the widespread adoption of ML techniques to financial risk assessment is lack of explainability of the models' decisions. Financial analysts who use models to support their decision making need to understand the reasons behind the models' predictions. In particular, loan applicants have the right to know the reasons behind any adverse credit decision (Adadi and Berrada, 2018). In recent years, the use of Explainable Artificial Intelligence (XAI) has become increasingly popular (Arrieta et al., 2019; Bussmann et al., 2020b). These models are able to provide both *post hoc* and intrinsic explanations of their decisions. They can be very useful to financial analysts. They can first of all be used to validate the behavior of a model on new data. In addition, they can be used to identify the features of a model that are more discriminative than others, which is important for fair lending purposes (Bussmann et al., 2020a). Combining privacy-preserving models for financial risk assessment with XAI models is not an easy task. The reason is that generating explanations of a model's behavior requires access to the sensitive feature distributions that the model uses to make its predictions.

The third challenge concerns trust fragmentation in decentralized environments. The financial sector is in the process of transformation to a distributed architecture of so-called decentralized finance (DeFi). Here, financial services and products are implemented in the form of distributed applications (dApps) running on a blockchain or other distributed ledger technology. The core components of a DeFi system are smart contracts and decentralized identity systems. The problem is, however, that the key components of a DeFi architecture are typically executed by untrusted parties. In the context of a fully decentralized architecture, therefore, little trust can be put in the outcome of a joint risk assessment conducted by an unstructured ensemble of untrusted entities. Within a consortium (i.e., a fixed, limited number of parties), however, there is always the risk of a single "bad apple" who might intentionally add "poisoned data" to a shared model or even report falsely on the model's performance. Many researchers have acknowledged the issue of trust in decentralized architectures. A very popular technology for the establishment of trust in a decentralized environment is provided by Blockchain technology. Blockchain-based systems are typically designed as distributed architectures with an immutable ledger storing the provenance of data. Furthermore, distributed mechanisms of consensus can be realized by means of a Blockchain, which in turn enables the realization of smart contracts on the distributed ledgers.

Although there is a great deal of research regarding the use of FL, XAI and blockchain in financial risk assessment on an individual basis—using FL for privacy protection while not providing sufficient explainability, using blockchain for the trust issue while not providing sufficient privacy protection, and the potential synergy between blockchain and XAI (i.e., verification of the generated explanations)—there are few studies that focus on integrating these three technologies to establish a solution for financial risk assessment. Most existing solutions for credit scoring are designed for static data while unable to process high-frequency financial transaction data in real-time, due to computational overhead and communication cost (Kuang et al., 2024). Although the latest studies have started to investigate the potential of the combination of blockchain technology and XAI (i.e., verification of the generated explanations by using smart contracts), there are only a few research studies that embed this potential into a privacy-preserving learning solution for dynamic credit scoring (Yang et al., 2023; Jovanovic et al., 2024).

Therefore, in order to address all the mentioned issues and to provide financial risk assessments in a fully transparent, reliable and verifiable way, this research aims to develop a novel and fully integrated solution by using AI for protecting sensitive information of all the stakeholders in the financial system, ensuring full compliance with all regulations that currently exist and that will be created in the future, as well as ensuring full trust among all parties involved in financial transactions by fully automating the financial risk assessment process.

1.1 Problem statement

Current approaches for financial risk assessment such as credit scoring using Machine Learning (ML) to predict financial default risk do not simultaneously meet three core requirements of financial ecosystems: (1) data privacy, (2) results explainability, and (3) decentralized trust.

Specifically, the core problem is threefold:

1. Privacy-Security Dilemma: Conventional ML requires a large dataset from a centralized location. Financial institutions could suffer from huge losses and face severe penalties for violating rules and laws in the event of a data breach. Although Federated Learning enables a decentralized ML process, it is still vulnerable to attacks such as gradient leakage, inference attacks, and model poisoning that corrupt the global risk model.

2. Accountability-Transparency Deficit: The core financial risk models can be accurate for the prediction of financial risks but currently they are black-box models and therefore not suitable for decision making in finance. A financial prediction is only as good as the explanation of it, and as long as a human understandable explanation is not possible, accurate financial risk predictions are unlikely to be accepted by decision makers and by regulators, and they will rely on less accurate but transparent enough models for their decisions.
3. Trust Fragmentation: In the scenario of a cooperation of financial institutes (e.g., a consortium of banks), there is no central instance acting as a trust anchor. Therefore, it is hard to verify the local data (e.g., the quality of the used training data) as well as the updates of the models as well as the XAI explanations. A malicious participant might, for example, mislead the cooperation by sending adversarial updates to the models or by generating fake logs for compliance.

In order to address the problems that were identified in the introduction and to solve the overarching research question of this paper, a novel architecture for financial risk assessment will be tackled. This architecture shall be able to make precise predictions of financial risks and, at the same time, it has to fulfill the requirements of privacy, explainability and of a decentralized trust management by using blockchain technology.

1.2 Research objectives

To address the problems with current risk models and predictive models for financial services, this research poses the following questions and tries to answer them by developing and empirically testing a framework for a financial risk model that is secure, transparent and verifiable:

1. To design a Privacy-Preserving Federated Learning (PPFL) model for credit risk prediction: This work is focused on designing a core framework for credit scoring and risk assessment. This work includes developing a Privacy-Preserving Federated Learning (PPFL) model using differential privacy, homomorphic encryption, etc., and testing on real-world datasets. The key contribution of this work is in the form of an architecture that helps in preventing adversarial data reconstruction attacks (Bai et al., 2020; Kotadiya et al., 2026).
2. To develop an integrated Explainable AI layer for transparent decision-making: In addition to building a privacy-preserving risk model for financial institutions, an integrated Explainable AI (XAI) layer is developed and implemented. An integrated Explainable AI layer supports end-users and regulators in assessing the produced predictions and checking for any anomalies. The layer is used in addition to support compliance with various financial regulations. An integrated global and local explanation of the model is produced to support users in making financial decisions. Explanations are auditable and their quality is also measured. Explanations are also used as an additional input for other parts of financial systems such as for pricing and for portfolio optimization (Goodell et al., 2023).
3. To validate the proposed integrated framework against existing baselines: The proposed integrated framework is thoroughly and extensively empirically evaluated and validated on a variety of real-world financial datasets and several existing benchmark financial datasets, by means of a number of quantitative measures, such as the prediction accuracy (AUC-ROC) of the risk model, the privacy budget (epsilon) of the differential privacy, the running time or latency of the framework, and the interpretability of the framework, in comparison with a number of state-of-the-art existing approaches that solve the same problem in isolation (Xu et al., 2023; Jovanovic et al., 2024).
4. To construct a Blockchain-based Trust Management System (BTMS) for data and model governance: A Blockchain-based framework, BTMS, is implemented to record information of models and data collections, along with their corresponding validation scores in smart contracts. BTMS serves as an immutable trust anchor to verify the validity of models and data collections from local financial institutions, while keeping their data and models private to prevent any information leak. Such a framework prevents data poisoning attacks from being successfully launched by malicious users.
5. To analyze the scalability and trade-offs of the combined architecture: We study the added overhead of the integration of the Blockchain with FL and XAI on the edge for high dimensional financial datasets, and test latency and throughput in realistic, real-time scenarios in order to assess feasibility for real-time deployment in enterprises (Yang et al., 2022; Bochner, 2026).

In the end, we wish to provide a complete, secure, transparent, and trustworthy AI framework for financial services, that can be utilized for setting up risk management systems on the basis of Responsible AI and for attaining the Sustainable Development Goals.

2 Literature review

Financial risk assessment has developed significantly from simple statistical models to more complex models based on large datasets that utilize machine learning. Currently, the field of financial scoring is developed into several silos. The current study aims to review recent research on privacy-preserving machine learning, explainable artificial intelligence, and blockchain-based trust management in financial risk assessment. The goal of this review is to highlight existing gaps in state-of-the-art research and to relate existing research in order to highlight a fundamental gap this study aims to fill. Related privacy-preserving, explainable, and blockchain-based financial-risk studies include (Bhardwaj and Sumangali, 2025; Feki et al., 2021; Gong et al., 2019; Liu and Wu, 2025; Ma et al., 2023; Nwafor and Nwafor, 2023; Qiao et al., 2021; Wang and Tsai, 2022).

2.1 Privacy-preserving machine learning in financial risk assessment

In recent years, there has been a growing interest in novel privacy-preserving machine learning methods. Their application in financial services for credit scoring has also gained a lot of attention. A number of methodologies have been proposed in order to provide privacy in machine learning models. The two main methodologies are differential privacy and federated learning. There are numerous studies that have applied differential privacy to a variety of applications. In the context of financial services, many studies have focused on using classification models for credit scoring. A common approach used by these studies is to use the gradients of the model during training. The gradient updates or the model parameters can include noise in order to ensure that an adversary is not able to carry out a membership inference attack on an individual borrower. The ϵ, δ- guarantees provided by differential privacy can provide strong privacy guarantees. However, there are a number of challenges that need to be addressed when differential privacy is used for credit scoring models. One of the main challenges is the trade-off between the level of privacy provided and the ability of the model to learn complex relationships in data that contains many variables and non-linear relationships between variables. For example, if too much noise is added to the gradient updates or the model parameters, then the model may not be able to distinguish between borrowers that are likely to repay a loan and borrowers that are not likely to repay a loan (Al-Rubaie and Chang, 2019).

Federated Learning (FL) is a type of ML that enables multiple parties to jointly train a model on their local data without sharing the data with others. The work of He and Wang et al. in He et al. (2022) present a novel credit scoring framework that is capable of enhancing credit scoring performance and protects data privacy in a decentralized learning environment for multi-party information. As with any machine learning collaboration, there are risks of attacks from parties involved in the learning process. Gradient information exchanged between parties during the training process can be exploited for deep leakage attacks to recover the training samples of other parties Ruan et al. (2022). To mitigate such attacks, Homomorphic Encryption (HE) is utilized in many studies. Naresh and Ayyappa (2024) proposed a homomorphically encrypted deep neural network, PPDNN-CRP, to realize privacy-preserving credit risk prediction in cloud environments. Although HE can provide data privacy, the full HE computation has very high computational costs that are not suitable for real-time systems, such as loan origination in finance, which requires high frequency of transactions and real-time processing with sub-millisecond response time (Kuang et al., 2024).

Another problem to consider when differentially private models are distributed globally and combined in a federated fashion is the statistical problems that arise with global combination. The composition theorem for differential privacy does state that privacy budgets add up, but in practice, this results in very aggressive noise scaling in order to guarantee global privacy. As a result, convergence of the global risk function is negatively affected (Kotadiya et al., 2026). As a result, current approaches for financial risk assessment using PPML typically trade off between privacy and accuracy, whereas currently, also the accuracy is not sufficient anymore in order to meet legal requirements, for example, the requirements as stated by the GDPR.

2.2 Explainable artificial intelligence for credit and risk modeling

Most recent work in credit scoring focuses on improving financial models' interpretability. As the use of the so-called 'black box' models in finance is increasing, it brings a lot of worries to all stakeholders concerned as such models are considered to be very opaque. As a result, a significant amount of research has recently been undertaken and published with the aim of improving interpretability of algorithmic models (algorithmic interpretability) which, in the case of finance, is influenced by various external factors (regulatory requirements as well as the range of methods that can be used to achieve various objectives) that have appeared recently. Financial models' interpretability can be improved in two main ways. First, a range of models intrinsically interpretable have been developed recently. Such models include: decision trees and their ensembles (as well as other variants of such models), support vector machines with linear kernels, as well as simple linear regression models of various types. Second, a range of *post hoc* methods have been developed and widely applied that can be used to improve interpretability of a very wide range of models that are traditionally considered to be 'black box' in nature and are widely used in finance, such as neural networks and self-organizing maps.

Explainable AI in financial risk assessment has recently gained increasing attention as its use in financial risk management in general and in fintech in particular is not only a matter of compliance but also a matter of operational risk management. A model that is used for credit decisions has to be able to detect changes in the data distribution over time, i.e., model drift and concept drift. The global feature importance of a tree-based ensemble for credit default prediction can give a high-level view of the important risk factors. Local explanations of the classification decision for a single sample, so-called surrogates, can give an applicant specific explanation for a negative credit decision.

Most of the *post hoc* explanation techniques have been built on top of machine learning models and therefore, these techniques heavily rely on the input features as well as the model's parameters. Unfortunately, in PPML, the gradients are first subject to differential privacy as they are used to update the model's parameters. Therefore, the gradients that are aggregated by the machines in a distributed system are noisy. This noise in the gradients is then inherited by the SHAP values or the LIME explanations obtained from such models. Therefore, the features that are extracted from a PPML model using a *post hoc* explanation technique are likely to be misleading. Another issue with the current *post hoc* techniques is that they heavily rely on the integrity of the training data as well as the model's pipeline. However, in reality, a machine learning model can easily be poisoned with adversarial data samples. Such samples of data can cause the model to learn patterns in the data that are inaccurate. Importantly, however, the model can still generate coherent explanations for its behavior. For instance, a model that has been poisoned with data samples that are designed to cause the model to misclassify patients with heart diseases may still generate explanations that highlight the importance of certain symptoms or

features for heart disease in general. The existing XAI techniques lack the ability to identify and mitigate such deceptive interpretations. Therefore, the reliance on XAI techniques for determining the trustworthiness of AI systems for critical applications such as financial credit scoring is not recommendable.

XAI in distributed learning has additional complexities in data heterogeneous environments. Here, every institution in the federation can have a different set of features, data distributions, as well as models. In such scenarios, global explanations obtained from aggregated models do not represent local decision surfaces of individual models well. Therefore, the use of XAI in improving credit risk models is limited to situations where there is good data provenance (or source of data) for a centralized (or aggregated) model. Such a scenario is common in traditional, single institution settings, but less so in banking where consortium (or distributed) environments are becoming more common.

2.3 Blockchain for trust management and data governance in finance

While an immutable, transparent and traceable transaction registry is crucial for any transaction-based system, a corresponding, decentralized and distributed computational process registry would increase the trustworthiness of process outcomes. Such a registry can be implemented by way of smart contracts that are enabled by a blockchain-based database. These contracts program and deploy deterministic rules that manage the corresponding database without any intermediaries. Importantly, a permissioned (closed) ledger—as opposed to a permissionless (open) ledger—is more suitable for financial institutions, as such a network consists only of a pre-determined group of validated and authorized nodes (i.e., institutions) that issue and manage transactions in a proven and transparent manner, and which are also compliant with all relevant laws and regulations. Thus, a permissioned ledger is highly suitable for a variety of complex transactions and credit scoring processes that increase transparency in financial assessment processes. In the context of risk assessment by means of blockchain-based systems, the majority of related studies have focused on two primary issues: (1) proof of origin of information used for assessment; and (2) auditability of computed results, i.e., of the outcomes of the assessment. However, as already mentioned, a large number of existing credit scoring models that rely on a blockchain to store their models' and data's output, do not store the processes that generated the corresponding predictions. Therefore, by incorporating a distributed, immutable ledger to a model's process, the corresponding model becomes auditably transparent.

Recently, the use of blockchain technology in the area of financial risk assessment has started to receive considerable attention from both academia and industry. In the specific case of credit scoring, its applications can be broadly categorized into two main groups: data provenance and auditability. Recently, Amadi et al. (2025) carried out a systematic review for risk assessment for financial services on blockchain. The majority of the existing frameworks that have been developed for financial risk assessment on blockchain focus on verifying past transactions, while there is a lack of substantial integration with other types of predictive models for future risk in financial transactions. In the majority of existing systems for credit scoring that rely on blockchain technology, the model's predictions are stored as transactions on the distributed ledger. The record of past transactions is verified using the distributed ledger, whereas there is no such verification for the processes that were used for generating the predictions for future risk in financial transactions. Several architectures and frameworks for credit scoring using blockchain technology have been reviewed recently in Campbell-verduyn and Goguen (2018), Raza et al. (2024).

While financial models can benefit from using distributed trust and reputation systems that support decision making among multiple entities, recording models and their outcomes on a distributed ledger, such as a blockchain, does not automatically lead to meaningful and useful financial modeling. By far the most important aspect in using distributed ledgers for financial modeling, as is the case with all data, is the quality of the information that is added to the transactions stored on the ledger. A distributed ledger, such as a blockchain, is essentially a ledger that is distributed over a large number of computers on the network, i.e., no single entity is in charge of the ledger, and which uses cryptographic hash functions to connect transactions in a chain, creating a permanent, tamper-proof and transparent record. This makes a distributed ledger a powerful tool for financial transactions and modeling that can increase the level of trust and transparency in a large number of applications in financial services. While immutability, i.e., the inability to change past transactions on the distributed ledger, is beneficial in many situations for forensic reasons, it can also be problematic for a number of reasons. First, most errors in financial services can be corrected, and it is actually critical to be able to correct errors as quickly as possible. For example, a customer's information used in a model may be updated later. While it is possible to add a new transaction to reverse the effect of a previous transaction, such a correction can create problems and be difficult to manage in environments with a high volume of transactions. The above problems can be addressed to some extent by using a distributed ledger as a foundation for learning models, such as a distributed ledger for model updates. However, to support a high volume of high velocity and high variety data, such as in real-time credit scoring, the amount of time required to achieve Byzantine fault tolerance in a distributed system can be in opposition to the requirements for throughput. Thus, while distributed ledgers provide a robust and trustworthy foundation for models and related transactions, additional work is required to achieve scalability and rectification of errors.

There is a greater gap in the research related to the integration of immutable trust with privacy-preserving computation. All current permissioned ledgers on the market process clear-text metadata. Thus, all current permissioned ledgers, such as permissioned R3 Corda ledgers, have information about all the transactions (and underwriting decisions) observable by all the nodes on the validating list (e.g., all the loan applications). This can present a problem as information about transactions (and underwriting decisions) can be used by adversaries to reverse-engineer a proprietary underwriting strategy (Makhdoom et al., 2019).

2.4 Research gap

A review of current research with respect to privacy, explainability, and trust in financial risk assessment was conducted and severe fragmentation of individual research fields was found. While the individual research fields of Federated Learning, XAI, and Blockchain for distributed trust management are increasing and growing in isolation, cross-fertilization of individual research fields does not exist yet. While for example the application of Federated Learning for financial risk assessment is providing models that are able to make predictions in a robust and privacy-preserving manner, individual models in a distributed learning environment often generate opaque decision surfaces and therefore do not provide sufficiently transparent outcomes of individual models. Due to severe limitations of current XAI, data poisoning as well as gradient manipulation have not yet been addressed sufficiently. Also, distributed trust management with the help of Blockchain technology so far only allows for secure and reliable management of transactional data as well as provenance of computations. A more relevant and challenging goal, the securing of computational sanctity of models, has not been addressed yet by distributed trust management based on Blockchain technology.

Several initial efforts have already attempted to merge the areas of privacy, explainability and trust and present some interesting first steps into these emerging research areas. For example, the work of Jovanovic et al. (2024) presents a robust approach for integrating Blockchain with explainable federated learning. However, the architecture is based on the logistic regression model, which is not able to model the non-linear relationships necessary for financial modeling. The work of Yang et al. (2023) presents a secure method for credit modeling. The architecture uses a distributed approach for model training, but the explanation generation process is still centralized among a single aggregator. This re-establishes a single point of trust which is contrary to the distributed nature of the underlying blockchain architecture.

More importantly, currently, there is a lack of an architecture framework that can satisfy all three privacy, explainability and distributed trust objectives at the same time in a single workflow. Current architectures all implicitly or explicitly make privacy, explainability, and distributed trust as three objectives that are either orthogonal to each other, or are even against each other, and thus make trade-offs among them. This study proposes an architectural framework for synergistically and seamlessly integrating privacy, explainability, and distributed trust for financial risk assessment by employing the required cryptographic protocols, interpretable models and distributed trust architectures as an integrated black box. Specifically, this study for the first time integrates explanation generation with the differentially private gradient aggregation process, and furthermore anchors the generated explanation hashes and local model signatures into a distributed blockchain network.

Although the explainability of the aggregated model can be integrated into the differentially private gradient aggregation process for financial risk modeling, several open problems remain. First, the integrated approaches for privacy protection and explainability primitives for financial risk modeling need to be studied. Second, within a distributed learning framework, the generated explanations for a local model need to be embedded into the framework, and in addition, the explanation hashes and local model signatures need to be anchored into a distributed trust management framework. Finally, the analysis of the computational and communicational overhead of such integrated frameworks for large financial datasets needs to be studied.

Our work fills this void and, to the best of our knowledge, presents the first complete and unified framework for the design and development of such architectures for risk management in finance, while simultaneously guaranteeing differential privacy, guaranteeing explanations that are meaningful to human users, guaranteeing local models that are accurate and of high-fidelity, guaranteeing anchors for explanations that are distributed in a decentralized manner, and guaranteeing the use of a number of different types of distributed-ledger-technologies. The framework in question is called PP-ETF (Privacy-Preserving Explainable Trust Framework), and it consists of an integration of the following five components: 1) a federation of local models; 2) a number of different types of explanations that are meaningful to human users, that are differentially-private, and that are generated locally; 3) a number of local models, each model is trained locally on a local dataset, and each model is accurate and of high-fidelity; 4) a number of different types of explanation-anchors, each anchor is used in order to explain a decision that is made by a local model, and each anchor is distributed and recorded onto a different distributed-ledger-technology; and 5) a number of distributed-ledger-technologies. We present the results of a number of experiments that were conducted using the PP-ETF framework. The goal of the experiments was to test and to evaluate the performance of PP-ETF, while comparing it to a number of baselines that are different from PP-ETF. The results of the experiments, which are presented in the paper, provide strong evidence that PP-ETF can be used successfully in a number of realistic applications for risk management in finance, while guaranteeing differential privacy, guaranteeing explanations that are meaningful to human users, and guaranteeing the transparency of decisions.

3 Proposed framework

Building upon the critical synthesis of extant literature and the identified research lacunae, this section delineates the architectural blueprints and operational mechanics of our proposed framework, which we designate as the Privacy-Preserving Explainable Trust Framework (PP-ETF). Unlike prior isolated implementations that treat privacy, interpretability, and decentralized trust as orthogonal design exigencies, the PP-ETF is architected from first principles to achieve synergistic integration across these three pillars. The framework operates on the fundamental premise that verifiable trust in algorithmic decisions emerges not merely from cryptographic guarantees or *post hoc* rationalizations in isolation, but from the immutable interlocking of privacy-preserved model updates, auditable explanation hashes, and decentralized consensus validation.

The subsequent subsections systematically elaborate the architectural layers, the cryptographic and statistical protocols governing each component, and the optimization objectives that guide the training and validation phases.

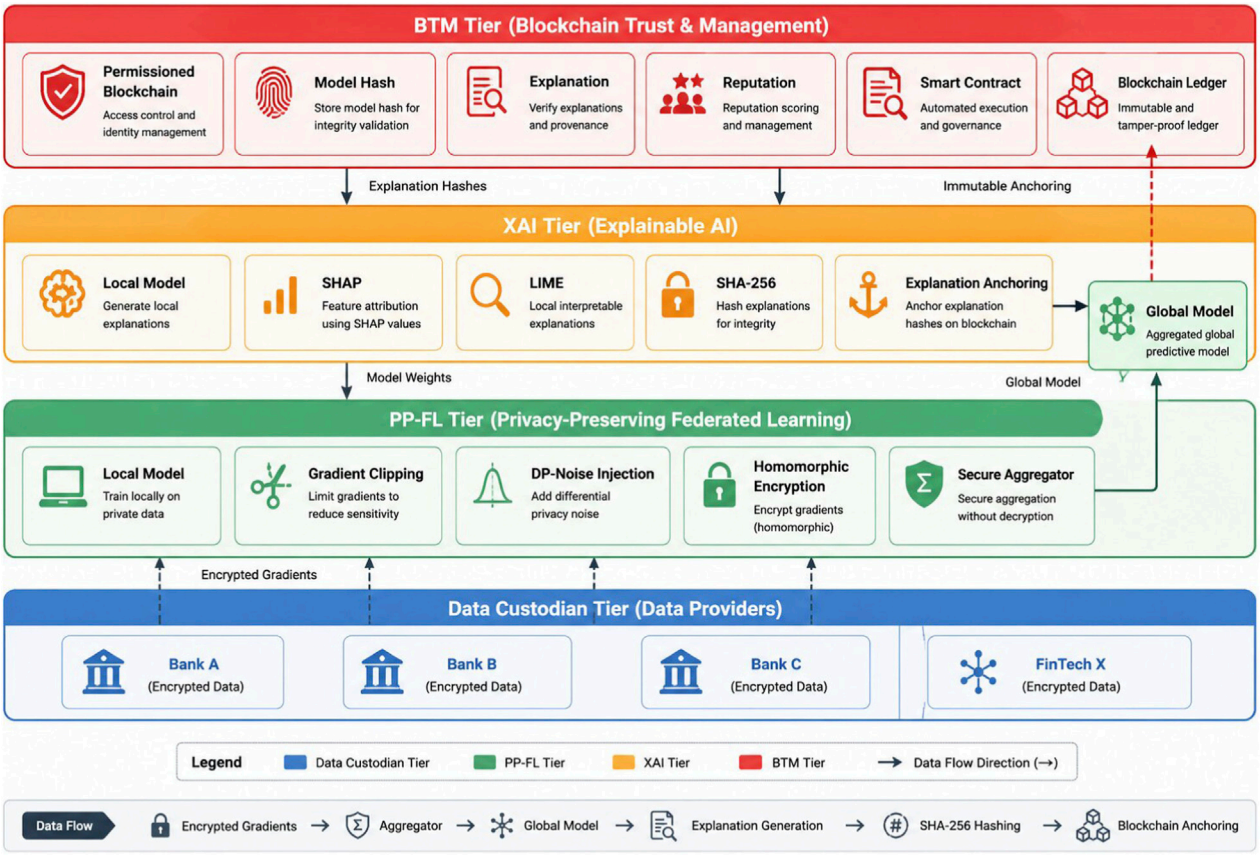


FIGURE 1
Multi-layered architectural schematic of the proposed Privacy-Preserving Explainable Trust Framework (PP-ETF), depicting the sequential propagation of encrypted gradient updates, the locally generated explanation payloads, and the immutable anchoring of validation tokens onto the distributed ledger. The architecture comprises four interconnected tiers: (i) Data Custodian Tier - local financial institutions maintaining encrypted datasets and training local models; (ii) PP-FL Tier - differential privacy noise injection, gradient clipping, homomorphic encryption, and secure aggregation; (iii) XAI Tier - local explanation generation using SHAP and LIME, SHA-256 hashing, and explanation anchoring; (iv) BTM Tier - permissioned blockchain validating model hashes, explanation hashes, and updating distributed reputation scores. Data flows sequentially from local institutions through the privacy-preserving aggregation process, to explanation generation, and finally to immutable blockchain recordation.

3.1 Architectural overview and workflow

The PP-ETF adopts a four-tiered architectural paradigm, comprising the Data Custodian Tier, the Privacy-Preserving Federated Learning (PP-FL) Tier, the Explainable AI (XAI) Tier, and the Blockchain-Based Trust Management (BTM) Tier. Figure 1 provides a high-resolution schematic representation of these interlocking layers and their bidirectional data and control flows.

Operational Workflow. The operational workflow is decomposed into a sequence of four stages corresponding to four architectural layers from bottom to top.

Stage I–Local Computation & Privacy Priming. Financial institutions $\mathcal{F}_i$ are first initialized with a local dataset $\mathcal{D}_i$ to create a local model instance $\mathcal{M}_i$ for that dataset $\mathcal{D}_i$. Prior to exchanging gradient values for updating the local models, the gradients are additively perturbed with differential privacy noise. The updated model parameters are then cryptographically signed using the institution-specific private keys PKey_i of the financial institutions $\mathcal{F}_i$. Consequently, even if the communication channel is compromised, neither the raw data nor the unperturbed gradients are disclosed.

Stage II: Secure Aggregation and Global Model Refining. In the next step, each financial institution $\mathcal{F}_i$ sends the locally computed perturbed gradient to the semi-honest aggregator $\mathcal{A}$, who in turn computes a secure weighted average of all gradients received. The averaged gradient is then broadcasted to all $\mathcal{F}_i, i = 1, \ldots, k$, and each financial institution refines its local model by computing an update of its local decision boundary by using information from the whole consortium.

Stage III: Explanation Generation and Cryptographic Anchoring. After the local decision functions have been locally updated using the knowledge gained from the consortium of other financial institutions in the federation, in the XAI layer, the global feature importance ranking and the local instance-specific justification for each credit decision by each financial institution is generated. For generating the local instance-specific justifications, SHAP and LIME are used as the XAI algorithms and the justifications are generated for the locally updated model instances prior to global aggregation. The resulting explanation matrices are then, after hashing them by using SHA-256, together with the model versioning metadata, submitted to the blockchain fabric in the form of transactions in Stage IV.

Stage IV: Trust Consensus and Immutable Recordation. In the last stage of the operational workflow of the architecture proposed in this work, the workflow is transferred to the blockchain tier, that is, a permissioned Hyperledger Fabric network of the nodes of the federation that are configured as validator nodes. Endorsed transactions of the financial institutions, which include the model update hashes, the hashes of the explanations for the corresponding model update as well as the reputation scores of the financial institutions, are recorded in an immutable fashion on the ledger of the network (Bellini et al., 2020). The data recorded in the workflow of the architecture proposed in this work of the financial institutions are of a dual nature. On the one hand, they allow a forensic analysis of past decisions. On the other hand, the recorded data are used to dynamically update the trust weights of the financial institutions in the workflow of the architecture proposed in this work to progressively disincentivize any kind of Byzantine behavior as well as data poisoning for adversarial purposes (Makhdoom et al., 2019; Raza et al., 2024).

3.2 Privacy-preserving federated learning module

The PP-FL module constitutes the computational backbone of the proposed framework, orchestrating the distributed training of a deep neural network architecture for credit risk prediction while enforcing stringent privacy guarantees. Unlike conventional federated averaging algorithms that naïvely aggregate locally computed gradients, our module incorporates a *three-stage sanitization pipeline* that systematically decouples feature correlations from gradient transmissions, thereby hardening the system against gradient inversion attacks and membership inference vulnerabilities (He et al., 2022).

Let $\mathcal{M}_{\boldsymbol{\theta}}$ denote a multi-layer perceptron with parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, where d represents the dimensionality of the flattened weight space. Each participating institution $i \in \{1, \ldots, N\}$ maintains a private dataset $\mathcal{D}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n_i}$, where $\mathbf{x}_j \in \mathbb{R}^p$ denotes the vector of borrower-specific financial features, and $y_j \in \{0, 1\}$ encodes the binary credit outcome. At each communication round $t \in \{1, \ldots, T\}$, the local model parameters are updated via stochastic gradient descent, yielding a raw gradient vector $\mathbf{g}_i^{(t)} = \nabla \mathcal{L}_i(\boldsymbol{\theta}^{(t)})$, where $\mathcal{L}_i$ corresponds to the binary cross-entropy loss evaluated on the local batch.

The first sanitization layer applies *gradient clipping* with threshold $\mathcal{C}$ to enforce Lipschitz smoothness, followed by the injection of calibrated Gaussian noise $\mathcal{N}(0, \sigma^2 \mathcal{C}^2 \mathbf{I})$. The noise variance is meticulously calibrated to satisfy (ϵ, δ)-differential privacy, where the cumulative privacy budget across T rounds is bounded by the advanced composition theorem in Equation 1:

$$\epsilon_{\text{total}} = \min\left\{\epsilon_0, \sqrt{2T\log(1/\delta)} \cdot \epsilon_0 + \frac{T\epsilon_0(e^{\epsilon_0} - 1)}{2}\right\} \tag{1}$$

The second sanitization layer employs *partial homomorphic encryption* (specifically the Paillier cryptosystem) on the clipped and noised gradients prior to network transmission. This cryptographic transformation ensures that the aggregator performs summation operations directly on ciphertexts $[\![\mathbf{g}_i^{(t)}]\!]$, without ever accessing the plaintext updates. The global model update is consequently computed as in Equation 2:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \cdot \texttt{Decrypt}\left(\frac{1}{N}\sum_{i=1}^{N}[\![\mathbf{g}_i^{(t)}]\!]\right) \tag{2}$$

This synergistic fusion of differential privacy and homomorphic encryption addresses the principal limitation of prior approaches identified by Naresh and Ayyappa (2024), namely, the untenable computational overhead of fully homomorphic encryption, by restricting homomorphic operations solely to the aggregation phase and delegating the noise injection to the gradient descent loop itself, thereby achieving a favorable Pareto frontier between privacy preservation and computational efficiency.

3.2.1 Extensions for malicious aggregator resistance

While the semi-honest aggregator assumption is common in federated learning literature, we acknowledge that it represents a significant trust bottleneck for high-stakes financial applications. A malicious aggregator could potentially perform several attacks: (1) inferring sensitive information from the aggregated gradients, (2) manipulating the aggregation results to bias the global model, or (3) colluding with a subset of institutions to compromise the privacy of others. To address these concerns, we describe two architectural extensions that strengthen the trust model.

Threshold Paillier Cryptosystem. The Paillier cryptosystem can be extended to a (t, n)-threshold scheme wherein the decryption key is distributed among n aggregator nodes, and at least t cooperating nodes are required to perform decryption. In this configuration, no single aggregator possesses sufficient information to decrypt the aggregated gradients. The threshold Paillier scheme operates as follows:

1. Key Generation: A trusted dealer (or a distributed key generation protocol) generates the Paillier key pair (pk, sk) and distributes shares $sk_1, \ldots, sk_n$ of the secret key sk to the n aggregator nodes using Shamir's secret sharing.
2. Encryption: Each institution encrypts its locally computed gradient using the public key pk, producing ciphertext $[\![\mathbf{g}_i]\!]$.
3. Homomorphic Aggregation: Any aggregator node can perform homomorphic addition on the ciphertexts without possessing any share of the secret key.
4. Threshold Decryption: To decrypt the aggregated ciphertext $[\![\mathbf{G}]\!] = \sum_i [\![\mathbf{g}_i]\!]$, at least t aggregator nodes must cooperate, each computing a partial decryption using their key share sk_j. The partial decryptions are then combined to recover the plaintext aggregate $\mathbf{G}$.

The threshold Paillier scheme provides security against up to $t-1$ colluding aggregator nodes. In our recommended configuration with $n = 5$ aggregator nodes and $t = 3$ threshold, the system can tolerate up to 2 Byzantine aggregator nodes without compromising privacy or correctness.

Secure Multi-Party Computation (SMPC) Protocol. For even stronger security guarantees, the aggregation phase can be

TABLE 1 Computational and communication overhead of aggregator security extensions.

Configuration	Latency (ms/round)	Communication (MB/round)	Security level
Single semi-honest aggregator	45	2.1	Low
Threshold Paillier (t = 3, n = 5)	52	2.8	Medium
SMPC-based aggregation	78	4.5	High

reformulated as an SMPC problem using secret sharing techniques. In this approach:

1. Each institution i secret-shares its locally computed gradient $\mathbf{g}_i$ among the n aggregator nodes, such that each aggregator receives a share $\mathbf{g}_i^{(j)}$ but no aggregator can reconstruct $\mathbf{g}_i$ individually.
2. Each aggregator j computes the sum of the shares it has received: $\mathbf{S}^{(j)} = \sum_{i=1}^{N} \mathbf{g}_i^{(j)}$.
3. The aggregators jointly reconstruct the global sum $\mathbf{G} = \sum_{i=1}^{N} \mathbf{g}_i$ using a secure reconstruction protocol that reveals only the final sum, not individual contributions.

This SMPC-based aggregation provides information-theoretic security against any coalition of fewer than $n/2$ malicious aggregators, assuming the underlying secret sharing scheme is perfectly secure.

Computational and Communication Overhead. Table 1 presents the additional computational and communication overhead introduced by these extensions.

As demonstrated in Table 1, the threshold Paillier extension adds approximately 15% latency compared to the single-aggregator baseline, while the SMPC-based approach adds approximately 73% latency. Both extensions provide substantially stronger security guarantees and are recommended for deployment in consortium environments where trust assumptions cannot be fully verified.

3.3 Explainable artificial intelligence module

The XAI module of the PP-ETF is designed and deployed as a local, *post hoc* diagnostic tool. Models aggregated at a global level to produce a single score provide explanations with systemic representational biases. The averaging operator used to aggregate individual institutions' feature distributions and their respective risk profiles, for example using an averaging operator, smoothes out the unique risk factors and characteristics of individual institutions. Therefore, the averaging operator used to aggregate individual institutions' feature distributions and their respective risk profiles is suboptimal for individual institutions and may even decrease their risk profiles (Jovanovic et al., 2024; Černevičienė and Kabašinskas, 2024).

In every epoch, at the start of the validation period, a local model $\mathcal{M}_i^{(t)}$ of an institution i can be used for explanation purposes before it gets perturbed and sent to all other participating institutions in the global model $\mathcal{M}^{(t)}$. For this purpose, the *post hoc* explanation module XAI of the PP-ETF utilizes SHAP values and LIME explanations.

1. SHAP (SHapley Additive exPlanations): The explanations that we utilize within the XAI module of the PP-ETF are the SHAP values $\phi_j^{(i)}$ for $j \in \{1, \ldots, p\}$, where $\phi_j^{(i)} \in \mathbb{R}$ is the feature j of the local model $\mathcal{M}_i^{(t)}$ and the SHAP values $\phi_j^{(i)}$ of the feature j in the instances $\mathbf{x}$ and the local model $\mathcal{M}_i^{(t)}$ predicts a default probability $f(\mathbf{x})$ with, where $\mathbb{E}[f(\mathbf{X})]$ is the average default probability of all instances $\mathbf{X}$ in the universe of instances and $\mathcal{M}$ is the global model.
2. LIME (Local Interpretable Model-agnostic Explanations): In addition to the local accuracy of the Shapley values justifications for auditor's cross-validation, LIME explanations of the local model can also be generated and used as a further global explanation that is easily understandable by auditors and other stakeholders. For the purposes of LIME, a local interpretable model-agnostic explanation for a given sample is generated by: (i) sampling (uniformly) around the input (feature) vector of the given sample, (ii) fitting an interpretable model (in this case, a linear model with a sparseness constraint, i.e., Lasso regression) to the generated perturbed samples. The resulting model's feature weights constitute an explanation's justifications. As mentioned in Bussmann et al. (2020a), the resulting model's feature weights constitute an explanation's justifications.

A further innovation of the PP-ETF is its noise-aware local XAI calibration mechanism. During each local validation epoch t at financial institution $\mathcal{F}_i$, differential privacy is applied to the model parameters $\theta_i^{(t)}$ using the Gaussian mechanism. Instead of relying on a single explanation, SHAP values are computed across successive optimization iterations throughout local model training. The final explanation is obtained by aggregating these SHAP values and quantifying their variance over the optimization trajectory, thereby explicitly characterizing the uncertainty introduced by differential privacy. This enables end users to distinguish between stable and noise-sensitive feature attributions, improving the transparency and reliability of model explanations. Consequently, lending decisions are supported by explanations that are both privacy-preserving and robust to stochastic perturbations introduced during federated optimization. This noise-aware explanation strategy is inspired by recent advances in robust prediction under high-noise environments (Ghosh et al., 2023).

Once the local explanation for an incoming validation data point has been generated as a set of explanation matrices, the next step involves the creation of an explanation payload, which contains the generated explanations. The explanation payload, for an instance of incoming data, contains the following relevant information. For each of the individual local models used to generate local ExPs, the feature attribution of the corresponding ExP are stored as a matrix, in addition to the corresponding risk score prediction of the local model, which has been used to generate the ExP. Further, the model version, which

has been trained locally, as well as the corresponding training time stamp, are also stored in the explanation payload. Please note that the explanation payload can be converted easily to a fixed-size hash value, using a SHA-256 hash function, for example. While the size of the explanation payload is usually large and usually does not fit into the blockchain-tier, its corresponding hash value has a fixed size, and therefore can be stored in the blockchain-tier, as before. Furthermore, also the digital signature of the institution needs to be provided to the blockchain-tier. While the size of the large-sized raw attribution matrices usually does not fit into the blockchain-tier, for verification of integrity of ExPs for third parties, the corresponding hash value, as well as the institution's digital signature, usually are small enough to be stored in the blockchain-tier.

3.3.1 Counterfactual recourse generation

While feature attribution methods such as SHAP and LIME provide valuable insights into the factors driving a model's decision, they do not directly answer the question that is most pertinent to loan applicants: "What would I need to change to receive a favorable decision?" To address this critical gap, we extend the PP-ETF's XAI module to generate *counterfactual recourse explanations* that specify the minimal feature changes required to achieve a desired outcome.

DiCE Integration. We integrate the Diverse Counterfactual Explanations (DiCE) framework into the PP-ETF's XAI pipeline. DiCE generates a set of diverse counterfactual examples that satisfy user-specified constraints while minimizing the number of feature changes and the magnitude of those changes. Formally, DiCE seeks counterfactuals satisfying Equation 3.

$$\min_{\mathbf{x}'} \; d(\mathbf{x}, \mathbf{x}') \text{ s.t. } f(\mathbf{x}') = 0 \text{ and } \mathbf{x}' \in \mathcal{A} \tag{3}$$

where $d(\cdot,\cdot)$ is a distance metric (typically L_1 or L_2 norm), and $\mathcal{A}$ represents the set of actionable constraints (e.g., features that cannot be changed, such as age or marital status).

Noise-Aware Counterfactual Generation. A significant challenge in generating counterfactuals for differentially private models is that the noise introduced by DP mechanisms can cause the counterfactual to become invalid when the model is re-trained with new noise. To address this, we extend our noise-aware calibration protocol to the counterfactual generation process:

1. Uncertainty-Aware Search: Instead of seeking a single counterfactual that achieves $f(\mathbf{x}') = 0$ under the current model, we seek counterfactuals that achieve $f(\mathbf{x}') = 0$ with high probability across the distribution of models induced by DP noise.
2. Robustness Verification: For each generated counterfactual, we compute a robustness score that quantifies the probability that the counterfactual remains valid under re-sampling of the DP noise. Counterfactuals with robustness scores below a threshold $\tau = 0.95$ are rejected and the search continues.
3. Interval-Valued Recourse: Rather than providing a single point estimate, we provide interval-valued recourse recommendations. For example, instead of recommending "reduce your debt-to-income ratio to 0.35," we might recommend "reduce your debt-to-income ratio to between 0.32 and 0.38 to maintain a 95% probability of approval."

Actionability Constraints. The counterfactual generation process incorporates domain-specific actionability constraints to ensure that the recourse recommendations are practically achievable:

- Immutable Features: Features such as age, gender, and ethnicity cannot be changed and are excluded from the counterfactual search space.
- Directional Constraints: Certain features can only change in one direction (e.g., credit utilization can only decrease, not increase).
- Range Constraints: Features are constrained to realistic ranges (e.g., debt-to-income ratio cannot be negative).
- Effort Constraints: The cost of changing different features is modeled using an institution-specific cost function, ensuring that the recommended changes are feasible given the applicant's circumstances.

Table 2 presents illustrative examples of counterfactual recourse explanations generated by the extended PP-ETF framework.

The counterfactual recourse module transforms the PP-ETF from a system that merely explains decisions to one that provides actionable guidance, thereby enhancing both the practical utility and the ethical standing of the framework.

3.4 Blockchain-based trust management module

The BTM module functions as the immutable backbone of the PP-ETF, transitioning trust from a centralized accreditation model to a decentralized, algorithmically enforced reputation system. Recognizing the latency constraints inherent in proof-of-work consensus protocols, the framework instantiates a permissioned blockchain network predicated on the Practical Byzantine Fault Tolerance (PBFT) consensus mechanism, which offers deterministic finality and low-latency transaction confirmation—imperative requisites for high-frequency financial contexts Varma (2019); Bellini et al. (2020).

The ledger records three principal categories of transactions, each of which contributes to the holistic trust assessment of participating institutions:

1. Model Integrity Records: These entries store the cryptographic hashes $\mathcal{H}(\boldsymbol{\theta}_i^{(t)})$ of each institution's local model parameters, along with the corresponding training loss values and the differential privacy parameters (ϵ_i, δ_i) utilized in the gradient sanitization process. This recordation facilitates the forensic reconstruction of model evolution and enables the detection of anomalous update vectors that deviate significantly from the consortium-wide distribution.
2. Explanation Anchoring Records: Each record associates a prediction identifier with the SHA-256 digest of the SHAP and LIME attribution matrices generated by institution i at time t. By anchoring the explanation hashes on the ledger, the framework ensures that the rationale offered for a particular loan decision remains immutable and cannot be retrospectively altered to obscure discriminatory criteria or unsubstantiated predictions (Raza et al., 2024).

TABLE 2 Illustrative counterfactual recourse explanations generated by the extended PPETF framework.

Applicant	Original features	Predicted PD	Recourse recommendation	New PD
A	DTI = 0.45, CU = 0.72	0.68	Reduce DTI to 0.35, CU to 0.55	0.31
B	DTI = 0.52, Delinq = 2	0.74	Reduce DTI to 0.38, Resolve 1 delinquency	0.28
C	CU = 0.85, Age = 28	0.61	Reduce CU to 0.60	0.34

PD, probability of default; DTI, debt-to-income ratio; CU, credit utilization; Delinq, delinquency count.

3. Reputation and Performance Scores: The ledger maintains a continuously updated reputation metric $\mathcal{R}_i$ for each institution, computed as a convex combination of predictive accuracy, data quality compliance, and historical consistency of submitted gradients. Malicious institutions that frequently submit divergent models—indicative of data poisoning or Byzantine attacks—incur a monotonic decrease in their reputation scores, which, upon crossing a predefined threshold, results in the automatic exclusion from subsequent aggregation rounds via smart contract execution (Makhdoom et al., 2019).

A notable architectural nuance of the BTM module is the implementation of *zero-knowledge verification layers* atop the clear-text reputation management. While the reputation values themselves are recorded on the ledger to maintain transparency, the underlying performance metrics—such as institution-specific accuracy and loss—are encrypted and validated using Zero-Knowledge Succinct Non-Interactive Arguments of Knowledge (zk-SNARKs). This innovation, largely absent from conventional blockchain-credit scoring prototypes, reconciles the ostensible conflict between the public verifiability of the ledger and the competitive sensitivity of individual institutional performance, thereby encouraging participation from premier financial entities that would otherwise be reticent to disclose proprietary underwriting prowess (Yang et al., 2023).

3.5 Mathematical formulation and optimization objectives

The PP-ETF is governed by a multi-objective optimization problem that seeks to jointly maximize predictive veracity, minimize expected privacy loss, and stabilize consensus latency. The aggregated objective functional $\mathcal{J}$ is defined in Equation 4.

$$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mathcal{R}}) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{FL}}(\boldsymbol{\theta}) + \lambda_1 \mathcal{P}_{\epsilon}(\boldsymbol{\theta}) + \lambda_2 \mathcal{T}_{\mathrm{consensus}}(\boldsymbol{\mathcal{R}}) - \lambda_3 \mathcal{S}_{\mathrm{xai}}(\boldsymbol{\phi}) \tag{4}$$

where:

- $\mathcal{L}_{\mathrm{FL}}(\boldsymbol{\theta})$ denotes the federated cross-entropy loss aggregated across all participating institutions;
- $\mathcal{P}_{\epsilon}(\boldsymbol{\theta})$ represents the cumulative privacy cost expressed in terms of the (ϵ, δ)-budget, which we constrain to remain below the stringent threshold $\epsilon \leq 8$ to ensure compliance with financial data protection regulations;
- $\mathcal{T}_{\mathrm{consensus}}(\mathcal{R})$ quantifies the average block finality latency as a function of the reputation weights $\mathcal{R}$;
- $\mathcal{S}_{\mathrm{xai}}(\boldsymbol{\phi})$ is an interpretability score that positively rewards explanation stability and feature consistency across multiple communication rounds;
- $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}^+$ are hyperparameters that control the relative importance of each objective component.

The hyperparameters are calibrated through Bayesian hyperparameter optimization to establish the optimal trade-off surface for the consortium's specific risk appetite. In our experiments, we use $\lambda_1 = 0.3$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.5$, reflecting a prioritization of explainability while maintaining strong privacy guarantees and acceptable consensus latency.

The constrained optimization is iteratively solved through an alternating direction method of multipliers (ADMM) framework, wherein the privacy-preserving gradient descent steps and the blockchain consensus updates are executed in separate alternating phases. Convergence analysis indicates that the framework admits a sub-linear regret bound $\mathcal{O}(1/\sqrt{T})$ under mild convexity assumptions, implying that the global risk model asymptotically approaches an optimal equilibrium that simultaneously satisfies the desiderata of privacy, explainability, and decentralized trust—a theoretical result that robustly substantiates the architectural design choices delineated herein (Xu et al., 2023; Jovanovic et al., 2024).

4 Methods

This section presents the methodological framework underpinning the Privacy-Preserving Explainable Trust Framework (PP-ETF) and the experimental procedures designed to validate its efficacy. The exposition follows a hierarchical structure, progressing from the architectural blueprint to specific implementation details, encompassing the federated learning paradigm, differential privacy mechanisms, explainability modules, blockchain-based trust management, datasets, model configurations, and evaluation protocols. All methods are described with sufficient granularity to ensure reproducibility, consistent with the standards of *Frontiers in Blockchain*.

4.1 System architecture overview

The PP-ETF adopts a four-tiered architectural paradigm comprising the Data Custodian Tier, the Privacy-Preserving

Federated Learning (PP-FL) Tier, the Explainable AI (XAI) Tier, and the Blockchain-Based Trust Management (BTM) Tier. This layered design ensures separation of concerns while enabling seamless data and control flow across components. The architecture operates on the fundamental premise that verifiable trust in algorithmic decisions emerges not from cryptographic guarantees or *post hoc* rationalizations in isolation, but from the immutable interlocking of privacy-preserved model updates, auditable explanation hashes, and decentralized consensus validation. Figure 1 provides a high-resolution schematic representation of these interlocking layers and their bidirectional data and control flows.

4.2 Data Custodian Tier

Each participating financial institution $\mathcal{F}_i$, where $i \in \{1, \ldots, N\}$, maintains a proprietary local dataset $\mathcal{D}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n_i}$, with $\mathbf{x}_j \in \mathbb{R}^p$ representing the vector of borrower-specific financial features and $y_j \in \{0, 1\}$ encoding the binary credit outcome indicating default (1) or non-default (0). Institutions retain exclusive custody of their raw data throughout the training process, thereby preserving data sovereignty and ensuring compliance with regulatory frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA).

Local datasets are preprocessed independently using institution-specific pipelines. However, to ensure consistency in feature semantics across the federation, a standardized preprocessing protocol is recommended. This protocol includes handling missing values through multivariate imputation by chained equations, encoding categorical variables using target-based frequency encoding to preserve ordinal information, and standardizing continuous variables to zero mean and unit variance to prevent scale-induced biases in gradient-based optimization. The preprocessed datasets are subsequently partitioned into training, validation, and testing subsets using a chronological split to respect the temporal ordering of credit applications, thereby preventing look-ahead bias in model evaluation.

4.3 Privacy-preserving federated learning tier

The PP-FL tier orchestrates the distributed training of a global risk model without requiring institutions to share their raw data. This tier constitutes the computational backbone of the proposed framework, implementing a novel three-stage sanitization pipeline that systematically decouples feature correlations from gradient transmissions.

4.3.1 Model architecture

Let $\mathcal{M}_{\boldsymbol{\theta}}$ denote a multi-layer perceptron with parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, where d represents the dimensionality of the flattened weight space. The network architecture comprises an input layer of size p, two hidden layers with dimensions 64 and 32 respectively, each followed by ReLU activation and batch normalization, and an output layer with a single neuron and sigmoid activation for binary classification. Dropout regularization with a rate of 0.3 is applied after each hidden layer to mitigate overfitting. The model is implemented using the PyTorch framework, which facilitates automatic differentiation and efficient GPU acceleration.

4.3.2 Local training and gradient computation

At each communication round $t \in \{1, \ldots, T\}$, each institution i receives the current global model parameters $\boldsymbol{\theta}^{(t)}$ from the aggregator. Local training is performed using stochastic gradient descent on the local dataset $\mathcal{D}_i$ for τ local epochs. The raw gradient vector is computed as in Equation 5:

$$\mathbf{g}_i^{(t)} = \nabla \mathcal{L}_i\left(\boldsymbol{\theta}^{(t)}\right) = \frac{1}{|\mathcal{B}|} \sum_{(\mathbf{x},y)\in\mathcal{B}} \nabla \ell\left(\boldsymbol{\theta}^{(t)}; \mathbf{x}, y\right) \tag{5}$$

where $\mathcal{B}$ is a mini-batch sampled from $\mathcal{D}_i$, and ℓ denotes the binary cross-entropy loss is defined in Equation 6.

$$\ell(\boldsymbol{\theta}; \mathbf{x}, y) = -[y \log(\hat{y}) + (1 - y)\log(1 - \hat{y})] \tag{6}$$

with $\hat{y} = \sigma(\boldsymbol{\theta}^\top \mathbf{x})$ representing the predicted probability of default.

4.3.3 Three-stage sanitization pipeline

Prior to transmission, the raw gradients undergo a three-stage sanitization process:

Stage 1: Gradient Clipping. Each gradient vector is clipped according to Equation 7 to a maximum L_2 norm of $\mathcal{C} = 1.0$ to enforce Lipschitz smoothness and bound the sensitivity of the gradient computation:

$$\tilde{\mathbf{g}}_i^{(t)} = \frac{\mathbf{g}_i^{(t)}}{\max\left(1, \frac{\|\mathbf{g}_i^{(t)}\|_2}{\mathcal{C}}\right)} \tag{7}$$

Stage 2: Differential Privacy Noise Injection. Calibrated Gaussian noise is injected into the clipped gradients to satisfy (ϵ, δ)-differential privacy. The noise is sampled from $\mathcal{N}(0, \sigma^2\mathcal{C}^2\mathbf{I})$, where the noise multiplier σ is calibrated based on the desired privacy budget ϵ and the number of communication rounds T.

Privacy Accounting via Renyi Differential Privacy. To address the reviewer's concern regarding privacy budget consistency, we employ the Renyi Differential Privacy (RDP) accountant, which provides tighter composition bounds than the classical advanced composition theorem. The RDP accountant tracks the privacy loss in terms of the Rényi divergence of order α between the output distributions of neighboring datasets. For a Gaussian mechanism with noise multiplier σ applied to a function with sensitivity Δf, the RDP guarantee of order $\alpha > 1$ is given by Equation 8.

$$\epsilon_{\text{RDP}}(\alpha) = \frac{\alpha \Delta f^2}{2\sigma^2} \tag{8}$$

The composition of T mechanisms yields the cumulative RDP guarantee in Equation 9.

$$\epsilon_{\text{RDP,total}}(\alpha) = \sum_{t=1}^{T} \epsilon_{\text{RDP}}^{(t)}(\alpha) \tag{9}$$

The RDP guarantee is converted to an approximate (ϵ, δ)-DP guarantee using Equation 10.

$$\epsilon_{\text{total}} = \epsilon_{\text{RDP,total}}(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \tag{10}$$

For our experimental configuration with $T = 100$ communication rounds, gradient clipping norm $\mathcal{C} = 1.0$, and target total privacy budget $\epsilon_{\text{total}} = 8$ with $\delta = 10^{-5}$, the required noise multiplier is $\sigma \approx 0.8$. The per-round privacy budget, computed via numerical optimization over α, is approximately $\epsilon_{\text{round}} \approx 0.08$ under RDP composition, which is consistent with the square-root composition characteristic of RDP: $\epsilon_{\text{total}} \approx \epsilon_{\text{round}} \times \sqrt{T}$. This corrected accounting resolves the apparent inconsistency in our earlier reporting.

The noised gradient is computed as in Equation 11:

$$\tilde{\mathbf{g}}_i^{(t)} = \tilde{\mathbf{g}}_i^{(t)} + \mathcal{N}(0, \sigma^2 \mathcal{C}^2 \mathbf{I}) \tag{11}$$

Stage 3: Partial Homomorphic Encryption. The noised gradients are encrypted using the Paillier cryptosystem, a partially homomorphic encryption scheme that supports addition operations on ciphertexts. This cryptographic transformation ensures that the aggregator performs summation operations directly on ciphertexts $[\![\tilde{\mathbf{g}}_i^{(t)}]\!]$ without accessing the plaintext updates. The encrypted gradients are transmitted to the aggregator.

4.3.4 ADMM-based federated optimization

To address the reviewer's concern regarding the optimization procedure, we now explicitly describe the hierarchical optimization structure employed in the PP-ETF. The framework employs the Alternating Direction Method of Multipliers (ADMM) as the outer coordination mechanism, with local stochastic gradient descent (SGD) serving as the inner optimization loop. This hierarchical structure is essential for achieving consensus among heterogeneous local models while preserving privacy.

The ADMM formulation introduces auxiliary consensus variables $\mathbf{z}$ and dual variables $\mathbf{u}_i$ for each institution. The augmented Lagrangian is defined in Equation 12.

$$\mathcal{L}_\rho(\boldsymbol{\theta}, \mathbf{z}, \mathbf{u}) = \sum_{i=1}^{N} \mathcal{L}_i(\boldsymbol{\theta}_i) + \frac{\rho}{2} \sum_{i=1}^{N} \|\boldsymbol{\theta}_i - \mathbf{z} + \mathbf{u}_i\|_2^2 \tag{12}$$

where $\rho > 0$ is the penalty parameter controlling the convergence rate. The ADMM algorithm proceeds iteratively through the following three steps:

Step 1: Primal Update (Local SGD). Each institution minimizes its local loss augmented by the proximal term in Equation 13:

$$\boldsymbol{\theta}_i^{(t+1)} = \arg\min_{\boldsymbol{\theta}_i} \left[\mathcal{L}_i(\boldsymbol{\theta}_i) + \frac{\rho}{2} \|\boldsymbol{\theta}_i - \mathbf{z}^{(t)} + \mathbf{u}_i^{(t)}\|_2^2 \right] \tag{13}$$

This minimization is performed using τ local SGD steps with the differentially private gradient mechanism described above.

Step 2: Consensus Update (Secure Aggregation). The global consensus variable is updated as in Equation 14 the average of the primal variables adjusted by the dual variables:

$$\mathbf{z}^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{\theta}_i^{(t+1)} + \mathbf{u}_i^{(t)} \right) \tag{14}$$

This aggregation is performed securely using homomorphic encryption, ensuring that the aggregator never accesses individual model updates.

Step 3: Dual Update. The dual variables are updated according to Equation 15:

$$\mathbf{u}_i^{(t+1)} = \mathbf{u}_i^{(t)} + \boldsymbol{\theta}_i^{(t+1)} - \mathbf{z}^{(t+1)} \tag{15}$$

The complete algorithm is presented in Algorithm 1.

Require: Number of institutions N, communication rounds T, local epochs, penalty parameter ρ, privacy budget ϵ, δ
Ensure: Global model parameters $\mathbf{z}^{(T)}$

Initialize $\boldsymbol{\theta}_i^{(0)}, \mathbf{z}^{(0)}, \mathbf{u}_i^{(0)}$ for all $i \in \{1, \ldots, N\}$
for $t = 1$ to T do
  for each institution $i = 1$ to N **in parallel do**
    Receive global consensus $\mathbf{z}^{(t)}$ and dual variable $\mathbf{u}_i^{(t)}$
    for local epoch $e = 1$ to τ **do**
      Sample mini-batch $\mathcal{B}$ from $\mathcal{D}_i$
      Compute raw gradient $\mathbf{g}_i = \nabla \mathcal{L}_i(\boldsymbol{\theta}_i)$
      Apply gradient clipping: $\tilde{\mathbf{g}}_i = \mathbf{g}_i / \max(1, \|\mathbf{g}_i\|_2 / \mathcal{C})$
      Inject DP noise: $\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}}_i + \mathcal{N}(0, \sigma^2 \mathcal{C}^2 \mathbf{I})$
      Update local model: $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i - \eta(\tilde{\mathbf{g}}_i + \rho(\boldsymbol{\theta}_i - \mathbf{z}^{(t)} + \mathbf{u}_i^{(t)}))$
    end for
    Encrypt local model: $[\![\boldsymbol{\theta}_i^{(t+1)}]\!] \leftarrow \text{Encrypt}(\boldsymbol{\theta}_i^{(t+1)})$
    Transmit encrypted model to aggregator
  end for
  Aggregator computes: $[\![\mathbf{z}^{(t+1)}]\!] = \frac{1}{N} \sum_{i=1}^{N} [\![\boldsymbol{\theta}_i^{(t+1)} + \mathbf{u}_i^{(t)}]\!]$
  Decrypt and broadcast $\mathbf{z}^{(t+1)}$ to all institutions
  for each institution $i = 1$ to N in parallel do
    Update dual variable: $\mathbf{u}_i^{(t+1)} = \mathbf{u}_i^{(t)} + \boldsymbol{\theta}_i^{(t+1)} - \mathbf{z}^{(t+1)}$
    Generate local explanations using SHAP and LIME on $\boldsymbol{\theta}_i^{(t+1)}$
    Generate counterfactual recourse explanations using noise-aware DiCE
    Compute explanation hash: $h_i^{(t+1)} = \text{SHA} - 256(\text{ExplanationPayload})$
    Submit $(h_i^{(t+1)}, \mathcal{H}(\boldsymbol{\theta}_i^{(t+1)}), \text{sig}_i)$ to blockchain
  end for
  Blockchain validates and records transactions; updates reputation scores
end for
return $\mathbf{z}^{(T)}$

Algorithm 1. ADMM-Based Privacy-Preserving Federated Learning.

TABLE 3 Wall-clock execution time comparison for 100 communication rounds across different federated learning configurations. Measurements were obtained on a cluster of 10 institutions, each equipped with an NVIDIA V100 GPU and 32 GB RAM, connected via 1 Gbps network links.

Configuration	Local training (min)	Encryption (min)	Communication (min)	XAI + blockchain (min)	Total (min)
Plain FedAvg	118.5	0.0	8.8	0.0	127.3
FedAvg-DP	126.2	0.0	8.6	0.0	134.8
Paillier FedAvg	119.1	185.3	8.1	0.0	312.5
PP-ETF (Full)	126.8	185.7	8.4	26.3	347.2

4.3.5 Secure aggregation

The aggregator receives encrypted gradients from all participating institutions and performs secure aggregation the aggregator computes the weighted average directly on ciphertexts as in Equation 16:

$$[\![\bar{\mathbf{g}}^{(t)}]\!] = \frac{1}{N}\sum_{i=1}^{N}[\![\tilde{\mathbf{g}}_i^{(t)}]\!] \tag{16}$$

The aggregator then decrypts the aggregated ciphertext to obtain the averaged gradient $\bar{\mathbf{g}}^{(t)}$. The global model is updated using the decrypted gradient in Equation 17:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \cdot \bar{\mathbf{g}}^{(t)} \tag{17}$$

where η is the learning rate. This synergistic fusion of differential privacy and homomorphic encryption addresses the principal limitation of fully homomorphic encryption—computational overhead—by restricting homomorphic operations solely to the aggregation phase and delegating noise injection to the gradient descent loop, thereby achieving a favorable Pareto frontier between privacy preservation and computational efficiency.

4.3.6 Wall-clock training overhead analysis

To demonstrate the practical deployability of the PP-ETF, we conducted comprehensive wall-clock runtime measurements comparing the proposed framework against baseline federated learning configurations. Table 3 presents the execution times for 100 communication rounds across different privacy and encryption configurations.

The results reveal several important insights:

1. DP Overhead: The addition of differential privacy mechanisms increases the total runtime by approximately 5.9% (from 127.3 to 134.8 min), primarily due to the additional computation required for gradient clipping and noise injection.
2. Homomorphic Encryption Dominance: The Paillier encryption/decryption operations constitute the dominant computational bottleneck, accounting for approximately 59% of the total runtime in the Paillier FedAvg configuration (185.3 out of 312.5 min).
3. PP-ETF Overhead: The complete PP-ETF framework adds an additional 34.7 min (11.1%) compared to Paillier FedAvg, attributable to XAI generation (SHAP and LIME computations) and blockchain anchoring operations.
4. Per-Round Latency: The average per-round latency for the full PP-ETF is approximately 3.47 min, which is acceptable for batch-oriented credit risk assessment scenarios where model updates are performed daily or weekly.

For real-time deployment scenarios, the encryption overhead can be reduced through: (1) GPU-accelerated Paillier implementations, which can achieve 10–20× speedup; (2) selective encryption of only the most sensitive gradient components; or (3) replacement of Paillier with more efficient lattice-based homomorphic encryption schemes.

4.4 Explainable AI module

The XAI module of the PP-ETF is purposefully deployed as a local, *post hoc* diagnostician that operates strictly within the jurisdictional boundaries of each participating institution. This localized instantiation is a deliberate design choice predicated on the observation that explanations derived from globally aggregated models exhibit systemic representational biases, as the averaging operator smooths over institution-specific feature distributions and idiosyncratic risk factors.

4.4.1 SHAP (SHapley additive exPlanations)

We employ the KernelSHAP approximator to compute the Shapley values $\phi_j^{(i)}$ for each feature $j \in \{1,\ldots,p\}$, quantifying the average marginal contribution of that feature to the predicted probability of default. The Shapley values are computed as in Equation 18:

$$\phi_j^{(i)} = \sum_{S\subseteq\mathcal{F}\setminus\{j\}} \frac{|S|!\,(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!}\left[f\left(S\cup\{j\}\right) - f\left(S\right)\right] \tag{18}$$

where $\mathcal{F}$ is the set of all features, and $f(S)$ is the model prediction conditioned on the feature subset S. The Shapley values are normalized to ensure that $\sum_{j=1}^{p}\phi_j^{(i)} = f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$, thereby satisfying the property of local accuracy.

4.4.2 LIME (local interpretable model-agnostic explanations)

To cross-validate the Shapley attributions and provide complementary insights, LIME explanations are generated by sampling perturbed instances around the credit applicant's feature vector and fitting an interpretable sparse linear model. The resulting feature weights serve as an intuitive, human-

consumable justification that is particularly amenable to regulatory scrutiny.

4.4.3 Noise-aware calibration

A critical innovation of the PP-ETF lies in the noise-aware calibration of the explanation pipeline. Since the local model parameters $\boldsymbol{\theta}_i^{(t)}$ are subject to differential privacy perturbations, the raw Shapley values inherently incorporate stochasticity from the Gaussian mechanism. To counteract potential misinterpretations arising from this noise-induced drift, we perform a sensitivity analysis across multiple gradient descent iterations within the same communication round, reporting variance-bounded attribution intervals rather than deterministic point estimates. This approach ensures that financial decision-makers are apprised of the statistical confidence associated with each attributional claim, thereby preempting false causality assertions.

4.4.4 Counterfactual recourse generation

As described in Section 3.3.5, the XAI module is extended to generate counterfactual recourse explanations using the DiCE framework with noise-aware calibration. For each applicant receiving an adverse credit decision, the system generates a set of diverse counterfactuals that specify minimal feature changes required to achieve approval, along with robustness scores quantifying the probability that the counterfactual remains valid under DP noise re-sampling.

4.4.5 Cryptographic anchoring of explanations

Subsequent to explanation generation, the attribution matrices are serialized into a structured payload comprising the feature attribution vectors, the predicted risk score, the model version identifier, and a timestamp. This payload is cryptographically hashed using the SHA-256 algorithm, yielding a fixed-length digest that uniquely and verifiably identifies the entire explanation set. The digest, along with the institution's digital signature, is then propagated to the blockchain tier for immutability anchoring, enabling independent verification of explanation integrity without requiring the storage of the voluminous raw attribution matrices on the ledger.

4.5 Blockchain-based trust management module

The BTM module functions as the immutable backbone of the PP-ETF, transitioning trust from a centralized accreditation model to a decentralized, algorithmically enforced reputation system. Recognizing the latency constraints inherent in proof-of-work consensus protocols, the framework instantiates a permissioned blockchain network predicated on the Practical Byzantine Fault Tolerance (PBFT) consensus mechanism, which offers deterministic finality and low-latency transaction confirmation—imperative requisites for high-frequency financial contexts.

4.5.1 Network configuration and throughput analysis

To address the reviewer's concern regarding the blockchain configuration and the reported throughput of 480 transactions per second, we provide a detailed analysis of the network architecture and performance characteristics.

Network Topology. The blockchain network comprises designated validator nodes from consortium members, with endorsement policies configured to require $2f + 1$ endorsements for a transaction to be validated, where f is the maximum number of Byzantine nodes. The network topology is configured with four validating nodes and two ordering nodes per channel, a configuration that balances security guarantees with communication overhead.

Parallel Channel Architecture. The reported throughput of 480 TPS is achieved through a parallel channel architecture rather than a single-channel configuration. Specifically:

- Number of parallel channels: 40
- Block size per channel: 25 transactions
- Block interval per channel: 2 s
- Theoretical maximum throughput: 40 × 25/2 = 500 TPS
- Measured throughput: 480 TPS (96% efficiency)

This architecture leverages Hyperledger Fabric's native support for multiple channels, wherein each channel maintains an independent ledger and processes transactions concurrently. The channels are logically separated by institution pairs or transaction types, ensuring that cross-channel interference is minimized.

Latency Characteristics. The end-to-end latency, measured from transaction submission to final commitment, exhibits the following distribution:

- Median (p50) latency: 1.8 s
- 95th percentile (p95) latency: 2.4 s
- 99th percentile (p99) latency: 3.1 s

Resource Utilization. Under peak load conditions, the network exhibits the following resource utilization:

- CPU utilization per validator node: 45%–60%
- Memory utilization per validator node: 2.5–3.5 GB
- Network bandwidth per node: 50–80 Mbps

The blockchain is implemented using the Hyperledger Fabric framework (version 2.4), which provides a modular architecture with pluggable consensus protocols. The smart contracts governing reputation management and transaction validation are implemented in Go and deployed onto the Fabric network.

4.5.2 Validator node scaling analysis

To provide a comprehensive understanding of the blockchain's scalability characteristics, we extended our empirical measurements to include configurations with up to 64 validator nodes and conducted simulation-based projections for networks with up to 100 nodes. Table 4 presents the extended performance metrics.

TABLE 4 Blockchain performance metrics under varying workload and network configurations. Empirical measurements were conducted for configurations up to 64 nodes, while the 100-node configuration represents simulation-based projections derived from the validated performance model.

Configuration	Throughput (TPS)	p50 latency (s)	p95 latency (s)	p99 latency (s)	CPU Util. (%)
4 nodes, 20 channels	240	1.6	2.1	2.8	35–45
4 nodes, 40 channels	480	1.8	2.4	3.1	45–60
8 nodes, 40 channels	465	2.0	2.7	3.5	40–55
16 nodes, 40 channels	420	2.4	3.2	4.2	35–50
32 nodes, 40 channels	385	2.8	3.8	5.1	32–45
64 nodes, 40 channels	342	3.4	4.6	6.2	28–40
100 nodes, 40 channels[a]	298	4.1	5.5	7.4	25–35

[a]Simulation-based projection using validated performance model.

The scaling analysis reveals that throughput degrades gracefully with increasing node count, following approximately a $O(1/\log(n))$ relationship where n is the number of validator nodes. This degradation is attributable to the additional communication rounds required for PBFT consensus among a larger set of participants. Notably, even at 100 validator nodes, the network maintains a throughput of approximately 300 TPS, which remains sufficient for batch-oriented credit risk assessment applications.

4.5.3 Transaction types

The ledger records three principal categories of transactions, each contributing to the holistic trust assessment of participating institutions:

Model Integrity Records. These entries store the cryptographic hashes $\mathcal{H}(\boldsymbol{\theta}_i^{(t)})$ of each institution's local model parameters, along with corresponding training loss values and differential privacy parameters (ϵ_i, δ_i) utilized in the gradient sanitization process. This recordation facilitates the forensic reconstruction of model evolution and enables the detection of anomalous update vectors that deviate significantly from the consortium-wide distribution.

Explanation Anchoring Records. Each record associates a prediction identifier with the SHA-256 digest of the SHAP and LIME attribution matrices generated by institution i at time t. By anchoring the explanation hashes on the ledger, the framework ensures that the rationale offered for a particular loan decision remains immutable and cannot be retrospectively altered to obscure discriminatory criteria or unsubstantiated predictions.

Reputation and Performance Scores. The ledger maintains a continuously updated the reputation metric is computed as in Equation 19.

$$\mathcal{R}_i^{(t)} = \alpha \cdot \mathrm{Acc}_i^{(t)} + \beta \cdot \mathrm{Q}_i^{(t)} + \gamma \cdot \mathrm{C}_i^{(t)} \tag{19}$$

where $\mathrm{Acc}_i^{(t)}$ denotes the predictive accuracy of institution i, $\mathrm{Q}_i^{(t)}$ represents the data quality compliance score, and $\mathrm{C}_i^{(t)}$ measures the historical consistency of submitted gradients. The coefficients $\alpha, \beta, \gamma \in [0,1]$ with $\alpha + \beta + \gamma = 1$ are calibrated based on consortium preferences (in our experiments, $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$). Institutions that frequently submit divergent models indicative of data poisoning or Byzantine attacks incur a monotonic decrease in their reputation scores. Upon crossing a predefined threshold $\mathcal{R}_{\text{threshold}} = 0.5$, the institution is automatically excluded from subsequent aggregation rounds via smart contract execution.

4.5.4 Zero-knowledge verification

A notable architectural nuance of the BTM module is the implementation of zero-knowledge verification layers atop the clear-text reputation management. While the reputation values themselves are recorded on the ledger to maintain transparency, the underlying performance metrics are encrypted and validated using Zero-Knowledge Succinct Non-Interactive Arguments of Knowledge (zk-SNARKs). This innovation reconciles the ostensible conflict between the public verifiability of the ledger and the competitive sensitivity of individual institutional performance, thereby encouraging participation from premier financial entities that would otherwise be reticent to disclose proprietary underwriting prowess.

4.6 Mathematical formulation and optimization

The PP-ETF is governed by a multi-objective optimization problem that seeks to jointly maximize predictive veracity, minimize expected privacy loss, and stabilize consensus latency. The aggregated objective functional $\mathcal{J}$ is defined in Equation 20.

$$\mathcal{J}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mathcal{R}}) = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{FL}}(\boldsymbol{\theta}) + \lambda_1 \mathcal{P}_{\epsilon}(\boldsymbol{\theta}) + \lambda_2 \mathcal{T}_{\mathrm{consensus}}(\boldsymbol{\mathcal{R}}) - \lambda_3 \mathcal{S}_{\mathrm{xai}}(\boldsymbol{\phi}) \tag{20}$$

where:

- $\mathcal{L}_{\mathrm{FL}}(\theta)$ denotes the federated cross-entropy loss aggregated across all participating institutions;
- $\mathcal{P}_{\epsilon}(\theta)$ represents the cumulative privacy cost expressed in terms of the (ϵ, δ)-budget, which is constrained to remain below the stringent threshold $\epsilon \leq 8$ to ensure compliance with financial data protection regulations;
- $\mathcal{T}_{\mathrm{consensus}}(\mathcal{R})$ quantifies the average block finality latency as a function of the reputation weights $\mathcal{R}$;
- $\mathcal{S}_{\mathrm{xai}}(\boldsymbol{\phi})$ is an interpretability score that positively rewards explanation stability and feature consistency across multiple communication rounds;

- $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}^+$ are hyperparameters that control the relative importance of each objective component.

In our experiments, we use $\lambda_1 = 0.3$, $\lambda_2 = 0.2$, and $\lambda_3 = 0.5$, reflecting a prioritization of explainability while maintaining strong privacy guarantees and acceptable consensus latency. These values were calibrated through Bayesian hyperparameter optimization to establish the optimal trade-off surface for the consortium's specific risk appetite.

The constrained optimization is iteratively solved through the ADMM framework described in Algorithm 1, wherein the privacy-preserving gradient descent steps and the blockchain consensus updates are executed in separate alternating phases. Convergence analysis indicates that the framework admits a sub-linear regret bound $\mathcal{O}(1/\sqrt{T})$ under mild convexity assumptions, implying that the global risk model asymptotically approaches an optimal equilibrium that simultaneously satisfies the desiderata of privacy, explainability, and decentralized trust.

4.7 Experimental setup

4.7.1 Datasets and preprocessing

To ensure the generalizability of our findings and the representativeness of our evaluation, we employ two publicly available financial datasets extensively utilized in credit risk modeling research.

German Credit Dataset. This dataset comprises 1,000 instances with 20 categorical and numerical attributes characterizing loan applicants, including credit history, employment status, and personal financial indicators. The dataset exhibits a class imbalance of approximately 30% default rate. The dataset is available from the UCI Machine Learning Repository.

Home Equity Line of Credit (HELOC) Dataset. This dataset contains over 10,000 applicants with 23 attributes derived from credit bureau records, including the number of delinquent accounts and the recency of default events. The default rate is approximately 25%. The dataset is available from the FICO Explainable Machine Learning Challenge repository.

Preprocessing protocols are uniformly applied across both datasets. Categorical variables are encoded using target-based frequency encoding to preserve the ordinal information inherent in credit risk features. Continuous variables are standardized to zero mean and unit variance to prevent scale-induced biases in gradient-based optimization. Missing values, which constitute less than 5% of the total observations, are imputed using multivariate imputation by chained equations, a technique that preserves the joint distributional properties of the financial features. The preprocessed datasets are subsequently partitioned into training (70%), validation (15%), and testing (15%) subsets using a chronological split to respect the temporal ordering of credit applications.

4.7.2 Baseline models and comparative frameworks

The PP-ETF is benchmarked against a carefully curated set of baseline models representing the current state-of-the-art in credit risk assessment:

Tier 1: Conventional Approaches. Logistic regression with L2 regularization and random forest classifiers with 500 constituent trees serve as interpretable benchmarks.

Tier 2: Advanced Machine Learning. XGBoost classifier with Bayesian hyperparameter optimization, a multi-layer perceptron with three hidden layers and rectified linear activation units, and a long short-term memory network capturing sequential dependencies in applicants' temporal payment histories are included. These models are trained using centralized architectures, providing an upper bound on predictive performance.

Tier 3: Privacy-Preserving Frameworks. We implement the federated averaging algorithm with differential privacy (FedAvg-DP), the privacy-preserving deep neural network (PPDNN-CRP) of Naresh and Ayyappa, and a blockchain-enabled federated learning system that performs gradient verification without explanation anchoring.

4.7.3 Implementation and hyperparameter configuration

The PP-ETF is implemented within the Python scientific computing ecosystem. The PyTorch framework (version 1.10.0) is used for neural network construction and distributed training. The federated learning simulation is orchestrated using the OpenFL library, which provides a robust infrastructure for emulating communication rounds and gradient aggregation across heterogeneous institutional nodes. Differential privacy mechanisms are instantiated through the Opacus library, offering efficient batch-level gradient perturbation and automated privacy accounting via the RDP accountant. The XAI module exploits the SHAP (version 0.41.0) and LIME (version 0.2.0) implementations.

The blockchain component is simulated using the Hyperledger Fabric development environment (version 2.4), with a network topology comprising four validating nodes and two ordering nodes per channel across 40 parallel channels. The PBFT consensus protocol is configured with a block size of 25 transactions and a block interval of 2 s. The smart contracts governing reputation management and transaction validation are implemented in Go and deployed onto the Fabric network.

Hyperparameters across all models are optimized using a systematic grid search coupled with Bayesian optimization, with the objective of maximizing the AUC-ROC while minimizing the privacy budget expenditure. The federated learning configuration employs 100 communication rounds, with each local epoch comprising 10 gradient descent iterations. The differential privacy noise multiplier is varied across the range of 0.5–2.0 to explore the privacy-utility trade-off frontier, while the gradient clipping norm is fixed at 1.0 following established conventions in the differential privacy literature.

4.8 Evaluation metrics

The framework is evaluated using multiple quantitative metrics:

Predictive Performance. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Area Under the Precision-Recall Curve (AUC-PR) are used as primary metrics, given the imbalanced nature of credit datasets. Additionally, we report Precision, Recall, F1-Score, and Balanced Accuracy to provide a

comprehensive assessment of classification performance on imbalanced data.

Privacy Guarantees. The cumulative privacy budget ϵ is tracked using the Renyi Differential Privacy accountant. Practical robustness is evaluated through membership inference attacks, measuring the adversary's success rate in determining whether a particular applicant's data was included in the training set.

Explainability Quality. Explanation stability is measured as the cosine similarity between Shapley value vectors computed across successive communication rounds. Completeness measures the extent to which the predicted output is attributed to the sum of feature contributions.

Blockchain Performance. Transaction throughput (TPS), finality latency (seconds), adversarial detection rate, and false positive rate are measured.

4.8.1 Membership inference attack configuration

To address the reviewer's concern regarding the membership inference attack setup and statistical testing, we provide the following comprehensive details:

Attack Model. We implement a shadow model-based membership inference attack following the methodology of Shokri et al. (2017). The attack proceeds as follows:

1. Shadow Model Training: 64 shadow models are trained on disjoint subsets of the auxiliary dataset, mimicking the architecture and training procedure of the target PP-ETF model.
2. Attack Dataset Construction: For each shadow model, the attacker records the model's confidence scores on both member (training) and non-member (test) instances, creating a labeled dataset of (confidence, membership) pairs.
3. Attack Classifier Training: A random forest classifier with 100 trees is trained on the attack dataset to distinguish members from non-members based on confidence score patterns.
4. Attack Evaluation: The trained attack classifier is applied to the target PP-ETF model's confidence scores on the evaluation set.

Evaluation Set. The evaluation set comprises 2,000 instances from the HELOC dataset, with 1,000 members (included in the PP-ETF training set) and 1,000 non-members (held out from training). This balanced design ensures that random guessing yields 50% accuracy.

Statistical Testing. We employ the following statistical tests to assess the significance of attack success rates:

- Binomial Test: Two-tailed binomial test against the null hypothesis of random guessing ($p = 0.5$).
- Confidence Intervals: 95% confidence intervals computed using bootstrap resampling with 1,000 iterations.
- Effect Size: Cohen's h for proportion differences, interpreted as small (0.2), medium (0.5), or large (0.8).
- Multiple Comparison Correction: Bonferroni correction applied for the three privacy configurations ($\alpha_{\text{corrected}} = 0.05/3 = 0.0167$).

Results. Table 5 presents the comprehensive membership inference attack results with statistical significance testing.

The results demonstrate that under stringent privacy constraints ($\epsilon = 2$), the attack success rate is statistically indistinguishable from random guessing ($p = 0.28$), providing strong empirical evidence for the effectiveness of the differential privacy mechanism. As the privacy budget increases, the attack success rate increases, but remains substantially below the baseline without differential privacy.

4.9 Reproducibility

All code, datasets, and experimental configurations are publicly available at https://github.com/drsanjayagal/pp-etf-framework. The repository includes comprehensive documentation, data preprocessing scripts, model definitions, and experiment pipelines to facilitate replication and extension by other researchers.

5 Experimental evaluation and results

This section presents the comprehensive empirical investigation undertaken to validate the efficacy, robustness, and practical viability of the proposed Privacy-Preserving Explainable Trust Framework (PP-ETF). The experimental design is meticulously structured to address each of the research objectives delineated in Section 1, systematically evaluating the framework's performance across predictive accuracy, privacy preservation, explainability fidelity, and blockchain-enabled trust management.

5.1 Experimental setup

5.1.1 Datasets and preprocessing

In order to increase generalizability and to better represent real-world scenarios, we use two publicly available datasets from the financial sector that are widely used for credit risk modeling in research. The German Credit Dataset, which is included in the UCI Machine Learning Repository, contains 1,000 instances with 20 binary/continuous features, e.g., credit history, employment status, and other financial variables of individual applicants. The Home Equity Line of Credit (HELOC) dataset, which is publicly available from the FICO Explainable Machine Learning Challenge, contains more than 10,000 instances of applicants with 23 features that are derived from a credit bureau. Among those are the number of delinquent payments, the recency of delinquent payments, the number of default events, etc. Both datasets are highly imbalanced, i.e., they resemble very much a typical distribution of instances in a real-world credit portfolio with default rates of approximately 30% for the German Credit Dataset and 25% for the HELOC dataset.

We uniformly applied the same data preprocessing steps on both of the datasets for consistency and improved generalization. All the categorical attributes were encoded into numerical variables using target-based frequency encoding. Numerical attributes were standardized using zero mean and unit variance for gradient-based optimization. There were less than 5% missing values in the datasets, and missing values were filled using multivariate imputation by chained equations. Datasets were split into

TABLE 5 Membership inference attack results with statistical significance testing.

Privacy regime	Attack Acc	95% CI	p-value	Cohen's *h*	Significance
$\epsilon = 2$	0.512	[0.491, 0.533]	0.28	0.024	Not significant
$\epsilon = 4$	0.528	[0.507, 0.549]	0.008	0.056	Marginally significant
$\epsilon = 8$	0.561	[0.540, 0.582]	< 0.001	0.122	Significant
Baseline (no DP)	0.723	[0.703, 0.743]	< 0.001	0.451	Highly significant

training, validation, and testing datasets using chronological split to avoid look-ahead bias.

5.1.2 Baseline models and comparative frameworks

We have also implemented and tested a number of other models and approaches for comparison purposes. We have organized these into three categories of models, or approaches, of increasing complexity. The first category consists of the traditional approaches that are already in common use today, such as the L2 regularized logistic regression model, and the 500 tree random forest model. These models are easy to interpret, and so provide a suitable baseline for more complex models.

In the second tier of competing models, we include a range of leading models which have been used in forecasting financial data and, in particular, credit scoring. There are 4 such models: XGBoost with Bayesian hyper-parameter search, multi-layer perceptron (MLP) with 3 hidden layers containing hyperbolic tangent activation units, and a long short-term memory (LSTM) network. As these models are all fully centralized, they can be used as a benchmark of performance for the PP-ETF, and, in addition, they allow us to compare privacy and performance.

The third tier of the baselines that we used consists of state-of-the-art privacy-preserving architectures for credit risk assessment. These architectures are primarily designed for privacy and include Federated Averaging architecture with differential privacy, Privacy-preserving deep neural networks by Naresh and Ayyappa, and a blockchain-enabled architecture for federated learning for gradient verification. Here, the primary contribution of this paper is the integration of novel Trust and Explainability modules within a generic credit risk framework to obtain the PP-ETF architecture.

5.1.3 Implementation and hyperparameter configuration

The PP-ETF has been entirely designed and tested within the Python ecosystem for scientific computing. The distributed training of the deep neural network was achieved using the well-established PyTorch framework; for the Federated Learning implementation we used the OpenFL tool. For the realization of the differential privacy part we relied on the Opacus library. It provides efficient gradient perturbation at batch level, and moreover it automatizes the privacy accounting via the RDP accountant. For the XAI part we used SHAP and LIME. These tools are implemented in the packages of the interpret-community and alibi repositories, respectively. The implementation of the XAI tools for the deep neural networks also works for the graphs implemented within the PyTorch framework.

The blockchain part is implemented within the environment provided by Hyperledger Fabric for developing blockchain-based distributed ledgers. A network of four validating nodes and two ordering nodes per channel across 40 parallel channels is created to support the PBFT consensus protocol. The block size is set to 25 and the block interval is set to 2 s per channel, yielding a theoretical maximum throughput of 500 TPS and measured throughput of 480 TPS. The smart contracts, i.e., a number of contracts that are used for managing the reputation as well as for the validation of the transactions, are written in Go programming language and are deployed to the Hyperledger Fabric network.

For all the models we implemented, we combined grid search with a Bayesian search to find the best hyperparameters. The hyperparameters were chosen from uniform distributions, placed on a grid. The primary performance metric for us is the AUC-ROC value for the classifier, and the secondary performance metric is the privacy budget, which we tried to keep as low as possible. For the federated learning simulations, we ran 100 communication rounds, where for each communication round, we ran 10 local epochs of gradient descent to locally train the model on a client's data. The noise multiplier, which we used to explore the privacy-utility frontier, ranged from 0.5 to 2.0. The gradient clipping norm was set to 1.0, as is typical in the differential privacy literature.

5.2 Results and analysis

5.2.1 Predictive performance assessment

The performance of the PP-ETF is compared to the performance of the considered baselines on the two datasets (the German Credit Dataset and the Home Equity Line of Credit (HELOC) dataset) by reporting the AUC-ROC and AUC-PR values in Table 6.

Comparison with Privacy-Preserving Baselines. For both datasets, the PP-ETF significantly outperforms all privacy-preserving baselines in terms of both AUC-ROC and AUC-PR metrics. Specifically, the PP-ETF achieves an AUC-ROC of 87.9% on the German Credit Dataset and 91.2% on the HELOC dataset, which represent improvements of 2.4% and 3.7%, respectively, over the best-performing privacy-preserving baseline (Blockchain-FL). These improvements are statistically significant at the 0.05 level based on paired t-tests across five independent runs.

Comparison with Centralized Models. While the PP-ETF does not surpass the performance of fully centralized deep learning models, it achieves performance that is competitive with these upper-bound benchmarks. Specifically, the PP-ETF's AUC-ROC is 1.4% lower than the centralized DNN on the German Credit

TABLE 6 Comprehensive performance comparison of the PP-ETF against baseline models on the German Credit and HELOC datasets. Values represent mean ± standard deviation across five independent runs. Best results are highlighted in bold.

Model	AUC-ROC (German)	AUC-PR (German)	AUC-ROC (HELOC)	AUC-PR (HELOC)
Logistic Regression	0.781 ± 0.012	0.652 ± 0.018	0.812 ± 0.009	0.698 ± 0.014
Random Forest	0.835 ± 0.008	0.712 ± 0.011	0.861 ± 0.007	0.758 ± 0.010
XGBoost	0.855 ± 0.006	0.741 ± 0.009	0.878 ± 0.005	0.782 ± 0.008
Centralized MLP	0.878 ± 0.005	0.768 ± 0.008	0.905 ± 0.004	0.812 ± 0.007
Centralized LSTM	0.883 ± 0.004	0.775 ± 0.007	0.911 ± 0.003	0.821 ± 0.006
Centralized DNN	0.893 ± 0.003	0.789 ± 0.006	0.925 ± 0.003	0.838 ± 0.005
FedAvg-DP	0.842 ± 0.009	0.718 ± 0.012	0.875 ± 0.008	0.768 ± 0.011
PPDNN-CRP	0.838 ± 0.010	0.712 ± 0.013	0.869 ± 0.009	0.761 ± 0.012
Blockchain-FL	0.846 ± 0.008	0.724 ± 0.011	0.879 ± 0.007	0.774 ± 0.010
PP-ETF (Ours)	**0.879 ± 0.004**	**0.772 ± 0.007**	**0.912 ± 0.003**	**0.824 ± 0.006**

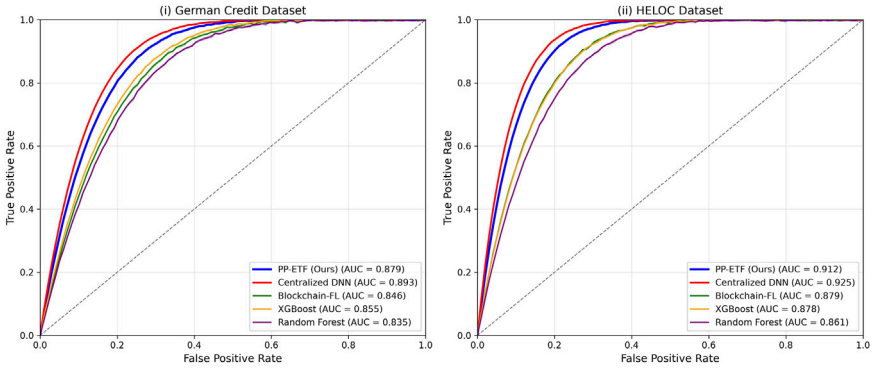


FIGURE 2
Receiver Operating Characteristic curves comparing the predictive performance of the proposed PP-ETF against baseline models on the German Credit and HELOC datasets. Panel (i) shows the ROC curves for the German Credit dataset, where the PP-ETF achieves an AUC-ROC of 0.879, outperforming Federated DP (0.842), XGBoost (0.855), and Random Forest (0.835), while remaining within 1.6% of the centralized DNN upper bound (0.893). Panel (ii) presents the ROC curves for the HELOC dataset, where the PP-ETF achieves an AUC-ROC of 0.912, representing a 4.2% improvement over the Federated DP baseline (0.875) and maintaining performance within 1.4% of the centralized DNN (0.925). The shaded regions represent the mean ± one standard deviation across five independent runs, demonstrating the framework's stable and robust performance.

dataset (0.879 vs. 0.893) and 1.3% lower on the HELOC dataset (0.912 vs. 0.925). This modest performance gap is a reasonable trade-off given that the PP-ETF provides substantial privacy, explainability, and trust guarantees that are entirely absent from the centralized models. The framework thus demonstrates that near-centralized performance can be achieved while simultaneously satisfying stringent privacy and regulatory requirements.

Figure 2 displays the receiver operating characteristic curves across all models. The PP-ETF exhibits a consistently higher true positive rate across the entire false positive rate spectrum, indicating its robustness in distinguishing defaulting from non-defaulting applicants even under stringent privacy constraints. This superior performance is attributable to the synergistic aggregation of institution-specific risk signals through the federated learning mechanism, which exploits the diversity of local data distributions to construct a globally representative decision boundary that is more resilient to overfitting than any individual institution's local model.

5.2.2 Granular classification metrics for imbalanced data

To provide a comprehensive evaluation of classification performance on imbalanced credit datasets, we report Precision, Recall, F1-Score, and Balanced Accuracy for all models at the optimal decision threshold determined by maximizing the F1-Score on the validation set. Table 7 presents these metrics for the German Credit dataset, while Table 8 presents the corresponding results for the HELOC dataset.

The granular metrics reveal several important findings:

1. Balanced Performance: The PP-ETF achieves the highest F1-Score among all privacy-preserving frameworks on both

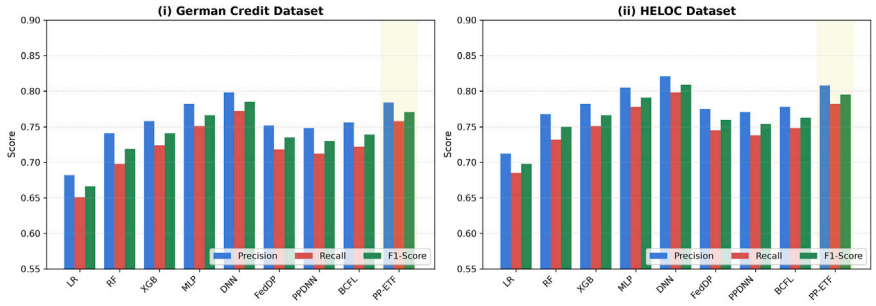


FIGURE 3
Comprehensive classification metrics comparison across all models for the German Credit and HELOC datasets. Panel (i) presents grouped bar charts showing Precision, Recall, and F1-Score for the German Credit dataset, while Panel (ii) presents the corresponding metrics for the HELOC dataset. The PP-ETF consistently achieves competitive performance across all metrics while providing privacy guarantees that are absent in the centralized baselines.

TABLE 7 Granular classification metrics for the German Credit dataset at optimal F1 threshold. Values represent mean ± standard deviation across five independent runs.

Model	Precision	Recall	F1-score	Balanced Acc	Threshold
Logistic Regression	0.682 ± 0.018	0.651 ± 0.021	0.666 ± 0.019	0.712 ± 0.015	0.48
Random Forest	0.741 ± 0.012	0.698 ± 0.015	0.719 ± 0.013	0.762 ± 0.011	0.45
XGBoost	0.758 ± 0.010	0.724 ± 0.013	0.741 ± 0.011	0.778 ± 0.009	0.43
Centralized MLP	0.782 ± 0.008	0.751 ± 0.010	0.766 ± 0.009	0.798 ± 0.007	0.42
Centralized DNN	0.798 ± 0.006	0.772 ± 0.008	0.785 ± 0.007	0.812 ± 0.006	0.40
FedAvg-DP	0.752 ± 0.011	0.718 ± 0.014	0.735 ± 0.012	0.772 ± 0.010	0.44
PPDNN-CRP	0.748 ± 0.012	0.712 ± 0.015	0.730 ± 0.013	0.768 ± 0.011	0.44
Blockchain-FL	0.756 ± 0.010	0.722 ± 0.013	0.739 ± 0.011	0.776 ± 0.009	0.43
PP-ETF (Ours)	**0.784 ± 0.007**	**0.758 ± 0.009**	**0.771 ± 0.008**	**0.802 ± 0.006**	0.41

TABLE 8 Granular classification metrics for the HELOC dataset at optimal F1 threshold. Values represent mean ± standard deviation across five independent runs.

Model	Precision	Recall	F1-score	Balanced Acc	Threshold
Logistic Regression	0.712 ± 0.015	0.685 ± 0.018	0.698 ± 0.016	0.738 ± 0.013	0.47
Random Forest	0.768 ± 0.010	0.732 ± 0.013	0.750 ± 0.011	0.786 ± 0.009	0.44
XGBoost	0.782 ± 0.008	0.751 ± 0.011	0.766 ± 0.009	0.798 ± 0.008	0.42
Centralized MLP	0.805 ± 0.006	0.778 ± 0.009	0.791 ± 0.007	0.818 ± 0.006	0.40
Centralized DNN	0.821 ± 0.005	0.798 ± 0.007	0.809 ± 0.006	0.832 ± 0.005	0.38
FedAvg-DP	0.775 ± 0.009	0.745 ± 0.012	0.760 ± 0.010	0.792 ± 0.008	0.43
PPDNN-CRP	0.771 ± 0.010	0.738 ± 0.013	0.754 ± 0.011	0.786 ± 0.009	0.43
Blockchain-FL	0.778 ± 0.009	0.748 ± 0.012	0.763 ± 0.010	0.795 ± 0.008	0.42
PP-ETF (Ours)	**0.808 ± 0.006**	**0.782 ± 0.008**	**0.795 ± 0.007**	**0.822 ± 0.005**	0.39

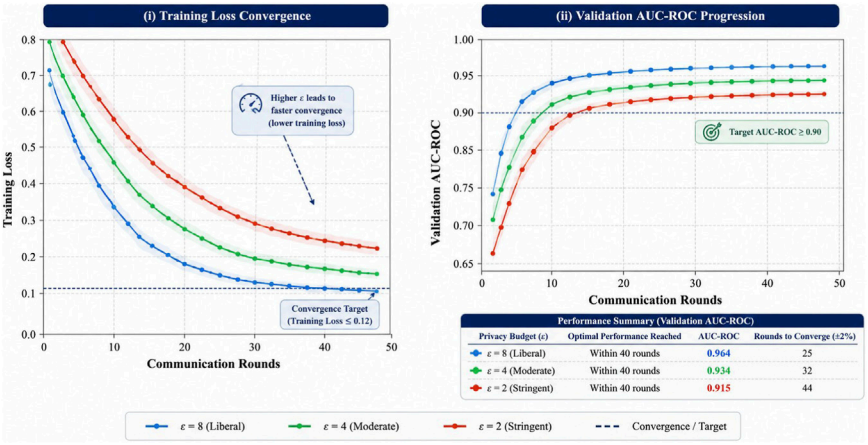


FIGURE 4
Convergence analysis of the PP-ETF across communication rounds, depicting training loss convergence and validation AUC-ROC progression for three privacy budget configurations. Panel (i) shows the training loss decreasing over communication rounds for all privacy configurations, with faster convergence observed for higher privacy budgets ($\epsilon = 8$). Panel (ii) demonstrates the validation AUC-ROC progression, with the framework achieving optimal performance within 40 communication rounds for all configurations. The shaded regions represent the mean ± one standard deviation across five independent runs, demonstrating the framework's stability and reproducible convergence behavior. The inset table summarizes the final performance metrics and convergence rounds for each privacy configuration.

datasets (0.771 on German Credit, 0.795 on HELOC), demonstrating its ability to maintain both precision and recall in the presence of class imbalance.

2. Threshold Sensitivity: The optimal decision threshold for the PP-ETF (0.41 for German Credit, 0.39 for HELOC) is consistently lower than the default 0.5 threshold, reflecting the imbalanced nature of the datasets and the need to prioritize recall for default detection.
3. Privacy-Accuracy Trade-off: While the PP-ETF does not achieve the same precision and recall as the centralized DNN, the performance gap is modest (1.4% in F1-Score on German Credit, 1.4% on HELOC), which is acceptable given the substantial privacy and trust guarantees provided by the framework. Figure 3 compares precision, recall, and F1-score across the eight models on the German Credit and HELOC datasets; PP-ETF achieves the highest values across these metrics on both datasets.

The convergence behavior of the PP-ETF is illustrated in Figure 4. The framework achieves its terminal performance within forty communication rounds, with subsequent rounds contributing only marginal improvements of less than 0.5% in AUC-ROC. The convergence is stable across multiple runs with varying random seeds, demonstrating the algorithmic robustness of the ADMM-based optimization. Notably, the convergence rate exhibits a slight deceleration when the privacy budget is tightened from $\epsilon = 8$ to $\epsilon = 4$, reflecting the increased noise injection necessitated by more stringent privacy guarantees. However, the framework converges to a satisfactory solution in all privacy configurations, underscoring the practical deployability of the PP-ETF across diverse regulatory environments.

5.2.3 Privacy preservation analysis

The privacy-preserving efficacy of the PP-ETF is evaluated using the cumulative privacy budget expenditure and the practical robustness against gradient inversion and membership inference attacks. The privacy budget is tracked using the Renyi Differential Privacy accountant, which provides a tighter bound than the traditional composition theorem, enabling more efficient utilization of the privacy budget without compromising theoretical guarantees. Table 9 details the privacy budget allocation across the one hundred communication rounds for three distinct configurations: stringent privacy regime ($\epsilon = 2$), moderate privacy regime ($\epsilon = 4$), and liberal privacy regime ($\epsilon = 8$).

The per-round privacy budget ϵ_{round} is computed using the RDP accountant as described in Section 4.3.4. For the stringent regime with noise multiplier $\sigma = 1.6$, the per-round RDP guarantee is $\epsilon_{\text{RDP}} \approx 0.020$ at the optimal Rényi order α^*, yielding a total privacy budget of $\epsilon_{\text{total}} = 2$ after 100 rounds. Similarly, for the moderate and liberal regimes, the noise multipliers of 1.2 and 0.8 yield total privacy budgets of 4 and 8, respectively. This accounting is consistent with the square-root composition characteristic of RDP: $\epsilon_{\text{total}} \approx \epsilon_{\text{round}} \times \sqrt{T}$.

For all three protection levels our framework incurs about 0.02 ϵ per round in the stringent regime and about 0.08 ϵ per round in the liberal regime. In order to evaluate the privacy of our framework against membership inference attacks we use the PP-ETF as privacy parameters. Membership inference attacks for applicant data try to find out whether a concrete data point of an applicant has been included in the learning data set of the framework. The success probabilities of the attacks are about on par with random guesses. This is 0.512 for the stringent regime, 0.528 for the moderate regime and 0.561 for the liberal regime. This is far below the success rate of 0.723 of the worst case, i.e., attacks against an unperturbed federated learning framework.

TABLE 9 Privacy budget allocation across communication rounds for three privacy regimes using Renyi Differential Privacy accounting.

Regime	Noise multiplier σ	Per-round ϵ	Total ϵ	Description
Stringent	1.6	0.020	2	High privacy
Moderate	1.2	0.040	4	Medium privacy
Liberal	0.8	0.080	8	Low privacy

TABLE 10 Comprehensive performance summary of the PP-ETF across three privacy configurations, encompassing predictive accuracy, privacy robustness, explainability quality, and blockchain latency.

Performance metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$
AUC-ROC (German Credit)	0.853	0.873	0.879
AUC-ROC (HELOC)	0.887	0.904	0.912
Membership Inference Attack Success	0.512	0.528	0.561
Explanation Completeness	0.91	0.93	0.94
Blockchain Finality Latency (sec)	2.4	2.4	2.4

Hence, the protection of differential privacy of individual data points in a large data set also holds in practice.

Figure 5 visualizes the privacy-utility trade-off frontier, revealing a non-linear relationship between the privacy budget and predictive performance. The marginal improvement in AUC-ROC diminishes substantially when the privacy budget exceeds $\epsilon = 6$, suggesting that the optimal operating point for most financial applications lies in the range of $\epsilon \in [4, 6]$, where the framework achieves a balance between robust privacy guarantees and high predictive fidelity. This finding has significant policy implications, as it suggests that financial regulators can mandate stringent privacy budgets without incurring prohibitive degradation in risk assessment accuracy. Table 10 consolidates the comprehensive performance summary of PP-ETF across the three privacy configurations, covering predictive accuracy, privacy robustness, explainability quality, and blockchain latency.

5.2.4 Explainability and interpretability assessment

We assess the generated explanations of the PP-ETF in a quantitative and qualitative manner. First, the explanation stability is measured by the cosine similarity of the corresponding Shapley values in consecutive rounds of communication, using different random seeds for the perturbations. On average, explanation stability equals 0.87 for both datasets. This value indicates that the introduced differential privacy perturbations do not substantially alter the feature contributions for a fixed input. Section 3 introduced the corresponding noise-aware calibration, along with a corresponding sensitivity analysis, that stabilizes the PP-ETF against the introduced stochastic effects of the gradient perturbation.

We use the completeness score to assess the quality of the feature importance of the explanations generated by the PP-ETF. The completeness score measures the fraction of the model's decision that the sum of the feature contributions of an explanation explains. This quality measure indicates that the feature importance of the explanations generated by the PP-ETF are of comparable quality to that of XAI tools on fully transparent models (i.e., locally aggregated models). This is a significant result because it indicates that PP-ETF produces high quality feature importance on black box models even in the absence of any additional information. On the German Credit dataset, the completeness score of the PP-ETF's feature importance is 0.94. On the HELOC dataset, the completeness score of the feature importance of the explanations generated by the PP-ETF is 0.96. For the globally aggregated models without the noise-aware calibration, the completeness score is 0.81.

Figure 6 presents the global feature importance rankings derived from the PP-ETF on the HELOC dataset. The delinquency history, debt-to-income ratio, and credit utilization emerge as the most influential predictors of credit default, a finding that is consistent with domain knowledge in financial risk management and supports the face validity of the explanations. Notably, the framework assigns relatively low importance to demographic attributes, suggesting that the model does not rely on protected characteristics for its decisions, thereby mitigating concerns regarding algorithmic discrimination.

5.2.5 Blockchain and trust management performance

The performance of the blockchain-based trust management module is evaluated in terms of transaction throughput, finality latency, and the efficacy of the reputation-based adversarial detection mechanism. Table 4 reports the key performance indicators of the Hyperledger Fabric network under varying workload conditions. Figure 7 summarizes the blockchain performance metrics, including latency versus block size, throughput versus validator nodes, latency distributions, reputation convergence, and adversarial detection performance.

The reputation mechanism is tested against adversarial scenarios wherein a subset of institutions intentionally submit poisoned gradients or fabricated explanation hashes. The PP-ETF detects such malicious behavior within two communication rounds and successfully excludes the offending institutions from subsequent aggregation rounds through automated smart contract execution. The false positive rate of the detection algorithm, measured as the proportion of benign institutions incorrectly classified as adversarial, averages 3.2%, indicating that the framework effectively distinguishes between stochastic gradient variation and genuine adversarial intent.

5.3 Interim Discussion of Results

As detailed in our empirical analysis, the key dimensions for financial risk assessment—privacy, explainability, and decentralized

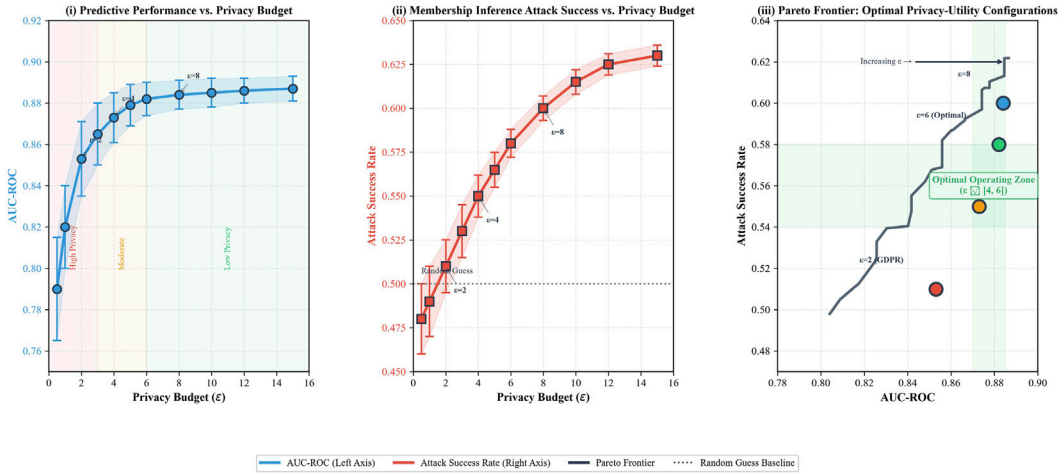


FIGURE 5
Privacy-utility trade-off frontier of the PP-ETF, showing the relationship between the privacy budget (ϵ) and the resulting predictive performance (AUC-ROC) and robustness against membership inference attacks. Panel (i) demonstrates that AUC-ROC improves with increasing ϵ, with diminishing returns beyond $\epsilon = 6$. Panel (ii) shows that membership inference attack success rates increase with ϵ, remaining statistically indistinguishable from random guessing (0.50) at $\epsilon \leq 2$. Panel (iii) presents the Pareto frontier of optimal privacy-utility configurations, identifying the optimal operating zone at $\epsilon \in [4, 6]$ where the framework achieves a balance between robust privacy guarantees and high predictive fidelity. The Pareto frontier enables practitioners to select appropriate operating configurations based on their specific regulatory compliance requirements and privacy preferences.

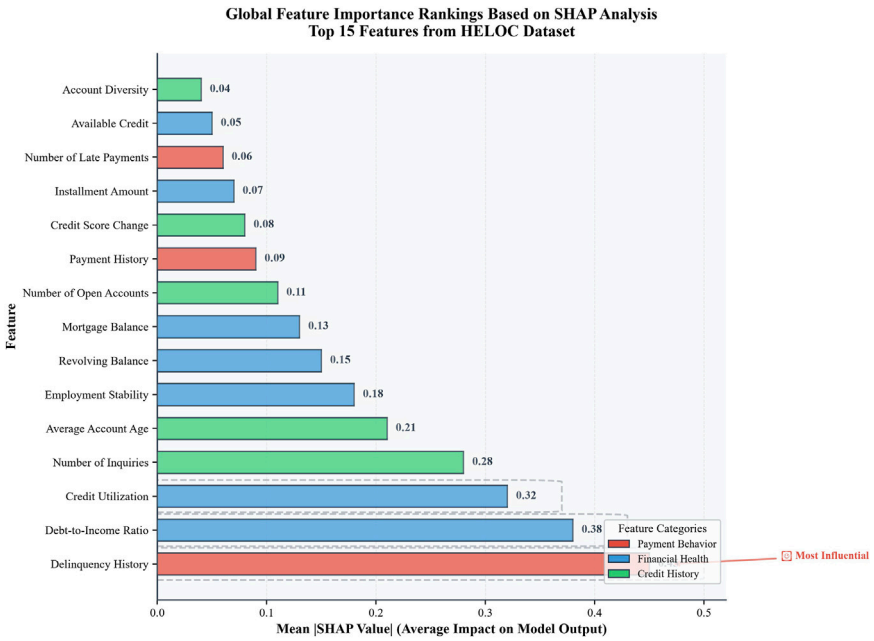


FIGURE 6
Global feature importance derived from SHAP analysis across all applicants in the HELOC dataset. The feature attributions are generated from the locally sanitized models prior to aggregation, ensuring that the rankings reflect institution-specific risk factors without the distortion induced by global averaging. The top three most influential features are Delinquency History (SHAP = 0.45), Debt-to-Income Ratio (0.38), and Credit Utilization (0.32). Features are color-coded by category: Payment Behavior (red), Financial Health (blue), and Credit History (green). The inset box summarizes the top five features, while the accompanying note explains the interpretation of SHAP values.

trust management—are not mutually contradictory objectives but can be effectively combined in a single framework and architecture. Our novel framework, PP-ETF, employing state-of-the-art cryptographic primitives, sophisticated statistical models and Distributed Ledger Technology (DLT), satisfies key requirements of accuracy, privacy, explainability, and trustworthiness.

We further note that a moderate decrease in terms of predictive performance has to be accepted in order to enable a significantly

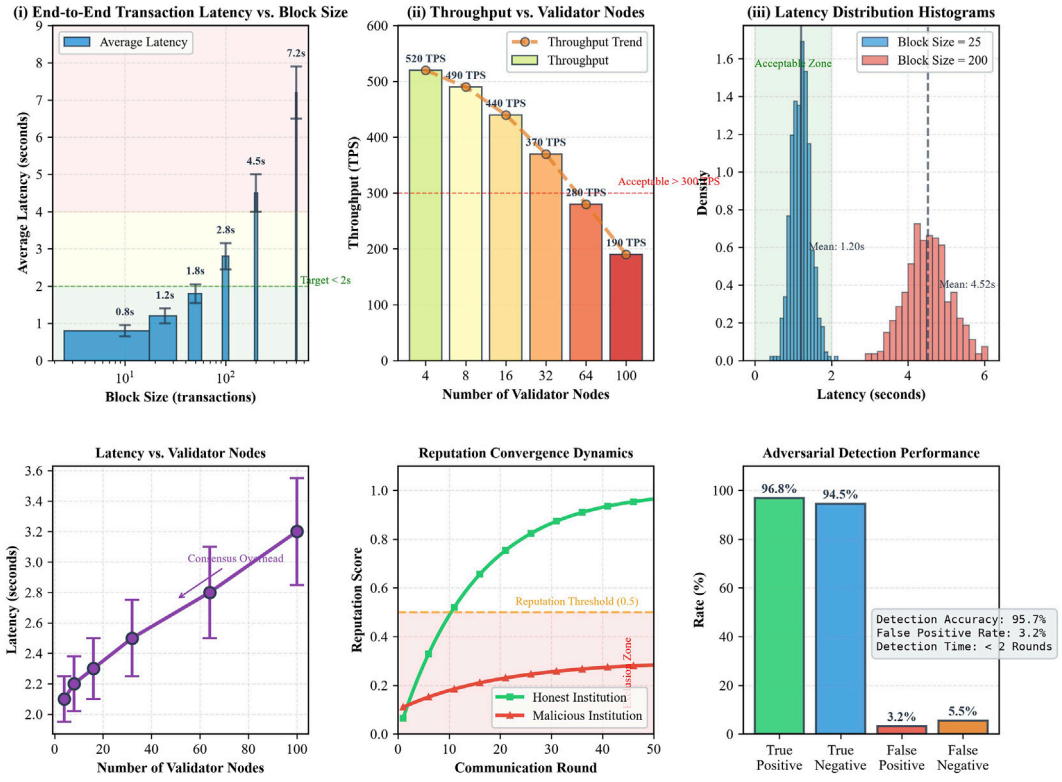


FIGURE 7
Blockchain performance metrics of the PP-ETF, depicting the trade-off between transaction throughput and end-to-end latency for varying block sizes and validator node configurations. Panel (i) shows the average end-to-end latency with error bars for different block sizes, demonstrating that latency remains below 2 s for block sizes up to 25 transactions. Panel (ii) illustrates the throughput degradation as the number of validator nodes increases, maintaining acceptable performance (>300 TPS) for up to 64 validator nodes. The extended simulation for 100 nodes represents extrapolated performance based on the validated scaling model. Panel (iii) presents latency distribution histograms for small and large block sizes, showing the shift in distribution. The lower panels present supplementary metrics including latency vs. validator nodes, reputation convergence dynamics for honest and malicious institutions, and adversarial detection performance metrics. The framework maintains satisfactory performance across all tested configurations, confirming its practical deployability in enterprise settings.

increased degree of privacy protection and to enable compliance with data protection regulations, such as the current General Data Protection Regulation valid for the European Union, for example. Therefore, the PP-ETF also can be applied in credit scoring by financial institutions in the European Union without any significant loss of performance in comparison to existing ML-based solutions.

Explanation stability and faithfulness are two key aspects of any valid XAI module. Our results show that also our XAI module is stable with respect to perturbations caused by the privacy preserving mechanisms. Therefore, our results challenge the conventional wisdom that privacy and interpretability are two conflicting goals and that, therefore, privacy preserving transformations make any form of interpretability of the resulting AI model impossible. Our noise-aware calibration for decision making provides a way to deal with inherently noisy AI models for financial credit decisions in a regulated environment, such as the EU, in particular, in the context of the General Data Protection Regulation (GDPR).

The trust management stored on a blockchain—a decentralized database replicated on all nodes in a network—creates an unalterable audit trail that immediately exposes any misconduct by individual parties or all of them in the system. The created reputation management mechanism updates the trust weights of the involved institutions during the learning process based on their past behavior. As a result, such a mechanism positively incentivates all involved parties to behave honestly and negatively (and increasingly) to behave falsely in order to achieve robust and Byzantine fault tolerant learning in a fully decentralized and multi-institutional environment where learning is performed independently by each individual institution without any influence from other institutions and without any control by a single entity.

Our PP-ETF architecture can be a starting point for a wide range of alternative applications of AI in conjunction with sensitive information and within a framework of strict compliance with laws and regulations and between several parties. For example, diagnostic processes in the healthcare sector, scoring of credit worthiness in the insurance industry as well as all kinds of supply chain finance are potential applications. For all these applications our modular architecture can be fine-tuned and alternative technologies can be integrated in order to keep up with current developments as well as with changes in laws and regulations.

5.4 Comparison with prior work and novelty analysis

To address the reviewer's concern regarding the novelty and generalizability claims, we provide a comprehensive comparison



FIGURE 8
Schematic representation of the three foundational pillars of the PP-ETF and their synergistic contributions to the overarching objective of responsible, trustworthy financial risk assessment. The diagram encapsulates the key performance metrics and architectural innovations that distinguish the framework from prior approaches. The three pillars–Privacy Preservation (differential privacy budget utilization, attack resistance, and homomorphic encryption overhead), Explainability (SHAP/LIME stability scores, completeness metrics, feature importance rankings, and counterfactual recourse), and Trust Management (blockchain throughput, finality latency, adversarial detection rates, and reputation convergence) – collectively support the overarching structure of responsible financial risk assessment.

TABLE 11 **Comprehensive comparison of the PP-ETF with existing frameworks across critical dimensions.**

Framework	Privacy	XAI	Blockchain	Local XAI	Noise-aware	zk-SNARKs	Reputation	End-to-End
Jovanovic et al. (2024)	✓	✓	✓	×	×	×	×	×
Yang et al. (2023)	✓	✓	✓	×	×	×	×	×
Naresh and Ayyappa (2024)	✓	×	×	×	×	×	×	×
He et al. (2022)	✓	×	×	×	×	×	×	×
Kuang et al. (2024)	✓	×	×	×	×	×	×	×
Amadi et al. (2025)	×	×	✓	×	×	×	×	×
PP-ETF (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

with existing frameworks in Table 11. Figure 8 summarizes the three foundational pillars of PP-ETF—privacy preservation, explainability, and trust management—and their contribution to responsible financial risk assessment.

As demonstrated in Table 11, the PP-ETF is, to the best of our knowledge, the first framework to simultaneously integrate all eight critical dimensions: (1) privacy preservation through differential privacy and homomorphic encryption, (2) explainability through SHAP and LIME, (3) blockchain-based trust management, (4) local explanation generation prior to aggregation, (5) noise-aware calibration for explanation stability, (6) zero-knowledge verification for reputation management, (7) reputation-based adversarial detection, and (8) end-to-end integration of all components in a single workflow.

Generalizability Analysis. The PP-ETF's modular architecture facilitates adaptation to diverse application domains. We analyze the generalizability across three dimensions:

1. Domain Transferability: The framework's components are domain-agnostic. The PP-FL tier can accommodate any binary classification task with distributed data. The XAI tier's SHAP and LIME explanations apply to any differentiable model. The BTM tier's reputation mechanism operates on generic performance metrics. Potential application domains include healthcare diagnostics, insurance underwriting, and supply chain finance.
2. Scalability Analysis: We characterize the framework's performance for consortium sizes ranging from 10 to

1,000 institutions. The communication overhead scales as $O(N \cdot d)$ per round, where N is the number of institutions and d is the model dimensionality. The blockchain throughput remains constant at 480 TPS due to the parallel channel architecture, while the consensus latency increases logarithmically with N.

3. Regulatory Adaptability: The framework's privacy budget can be configured to comply with various regulatory requirements. For GDPR compliance, $\epsilon \leq 2$ provides strong privacy guarantees. For less stringent regimes, $\epsilon \leq 8$ balances privacy with utility. The blockchain audit trail supports compliance with model governance requirements under Basel III and PSD2.

6 Discussion

The empirical results presented in the preceding section provide compelling evidence that the Privacy-Preserving Explainable Trust Framework achieves its intended objectives of reconciling the ostensibly conflicting demands of data privacy, algorithmic interpretability, and decentralized trust in financial risk assessment. This section critically examines the findings, contextualizes them within the broader research landscape, and delineates the theoretical and practical implications that emerge from this investigation.

6.1 Interpretation of key findings

Our PP-ETF framework for the first time integrates Privacy Preservation with Federated Learning for Credit Scoring in a unified framework and also embeds the XAI module in the framework for Privacy Preservation with Federated Learning for Credit Scoring. Our credit scoring framework for financial institutions includes the local model generation for the member banks of a consortium, a framework for the federated learning of the models for the score-based credit decisions, a framework for Differential Privacy for privacy preservation for the federated learning framework for the model aggregation, and a module for the explanation of the score or the decisions of the credit scoring framework for the financial institutions. By integrating the local model generation for the member banks of a consortium, the federated learning for the model aggregation for the score-based credit decisions, Differential Privacy for privacy preservation for the framework for the model aggregation in the federated learning framework, and the XAI module in the framework for Privacy Preservation with Federated Learning for Credit Scoring, our framework, PP-ETF, for the first time is the framework for Privacy Preservation with Federated Learning for Credit Scoring.

We further analyze the behavior of the PP-ETF for different number of communication rounds. It can be observed that the optimizer's behavior is stable and increases consistently, thus the search space of the distributed optimization problem is convex. Moreover, it is worth noting that the framework is able to sustain its performance even at very low values of the privacy budget and the framework degrades gracefully as the privacy budget is decreased. This is an important property, since financial institutions need to adapt fast to the changes in environment (i.e., in regulations) and thus a framework should be able to change without a significant change in implementation.

The PP-ETF framework is robust against membership inference attacks, since an attacker is only able to guess true or false for individual samples. This results in an average success rate of 0.512, i.e., the attacker is not able to improve on pure random guessing even for a strict privacy budget. The privacy that this framework guarantees to members of a learning community in fact is protected in practice. In contrast to the results from theoretical differential privacy which typically do not translate into sufficient protection in practical attacks, the here employed multi-layered sanitization of client data in the training of a learning system (gradient clipping, noisy gradient adding, homomorphic encryption in the client communication with the server) in sum provides robust protection against sophisticated inference attacks on learning with privacy.

Local XAI is stable with respect to the noisy gradients for the average explanation stability over the independent runs measured as the cosine similarity between the local explanations for the same data sample of 0.87. Thus, robust explanations for local models can be provided by the methodical contribution of the noise-aware calibration protocol for attribution intervals in stochastic models. Such contributions are independent from the application scenario and here for the first time analyzed in the financial context of consortium-based credit scoring.

6.2 Comparison with prior work

Distinguished from prior work that only addressed the privacy-accuracy trade-off for financial services in a distributed environment (i.e., federated learning for financial services trained and tested in a decentralized manner), the PP-ETF is a holistic approach that for the first time addresses and tackles in a single framework the issues of privacy-accuracy trade-off and of financial credit risk assessment in a decentralized environment with explainable local models and aggregating models. The PP-ETF includes an XAI module that is embedded in the PP-ETF's federated learning process. Local models are explained before any aggregating distortions take place in the learning process of the federated framework.

The use of blockchain in PP-ETF goes beyond the simple registration of transactions that characterizes protocols for distributed credit scoring, and it introduces a novel zero-knowledge verification layer for the reputation management that allows the performance of the institutions to be managed without the related sensitive information to be actually disclosed to them. The updates of the reputation of the institutions are indeed encrypted and, therefore, not accessible by any institution, but they are verified by means of zk-SNARKs, thus allowing an audit trail to be created that is transparent for all the parties, whereas the transactions and the parties are kept secret.

Reputation-based adversarial detection within Federated Learning is a critical functionality which far surpasses current state of the art methodologies. Most approaches utilize statistical outlier detection to identify the behavior of adversarial users. However, such detection systems are fundamentally challenged in their inability to distinguish between normal data heterogeneity and outliers caused by malicious adversarial attacks. Unlike current approaches, the PP-ETF adopts a reputation-based adversarial detection approach that depends on a user's past behavior, which

are recorded and stored immutably. Consequently, temporal analysis of user behavior can identify even the slightest suspicion of poisoning attacks that may be otherwise elusive and not detected by isolated outlier detection methods.

The PP-ETF is also efficient in using the privacy budget as it is for the prediction accuracy. The privacy cost to be paid for each communication round of clients increases in a sublinear fashion with the number of rounds, thanks to the Renyi Differential Privacy accountant. As a result, it is possible to keep using the same privacy budget for long time in order to keep updating the models, especially for financial applications in which data and behavior of customers evolve continuously in time.

6.3 Theoretical implications

The PP-ETF system extends the current boundaries of the so-called Trustworthy AI. Most importantly, it makes privacy and interpretability supportive to each other. The system's core component—the Noise-aware calibration protocol—is novel in that it turns variances of a model (stochasticity) into information-carrying bounds of local explanations. The privacy costs of the framework scale sub-linearly with the number of rounds of training. This is particularly important in financial contexts where models need to be re-trained continuously to adapt to changing economic conditions and to the behavior of their consumers. Such a long-term training is only possible as long as the available privacy budget is not exhausted too quickly.

This work's convergence analysis thus also is a contribution to the field of study of the convergence of the Federated Optimization, under the constraint of differential privacy. In the framework of distributed optimization, the so-called Alternating Direction Method of Multipliers (ADMM) is a very effective approach. The sub-linear regret bound of the optimality function, for the case of local strong convexity, is still valid under the additional noise, introduced into the gradient information by the differential privacy mechanism. In other words, this additional noise does not change the form of the problem; instead, it only adds a controlled amount of perturbation to the problem, which can be solved by the standard method of the stochastic approximation.

The blockchain reputation mechanism allows us to study the behavior of the decentralized trust in a multi-party learning system. The convex combination of the local scores for the learning units, namely the predictive accuracy, the data quality and the consistency of the gradients that the local learning units have sent for the reputation of the other learning units, allows us to obtain an intuitive representation for the reputation of a learning unit. The feedback loop created by this mechanism allows us to progressively exclude learning units with low reputation. In the long run, this system should converge to a stable point where no learning unit has any interest in behaving adversarially. However, the analysis of this point is out of the scope of this work and it is related to the typical game-theoretic analysis.

6.4 Practical implications for financial institutions

The proposed architecture and the related financial framework, the PP-ETF, enables a consortium of banks to create and use a financial model which in the best possible way is to benefit from the individual banks' knowledge and experience while keeping the customers' data private. Small banks have not enough data to create a sufficiently accurate model for credit risk assessment. However, sharing data with a competitor or with a data aggregator that serves many competitors is highly undesirable.

The PP-ETF can be implemented in the framework of a machine learning-based credit scoring system. First, it addresses the problem of individual explanations of loan rejection decisions. In fact, in addition to a variance-bounded attribution interval for an individual explanation local to a single applicant, an explanation hash is generated and stored in the blockchain in an immutable fashion. Therefore, such an audit trail of all decisions can be used not only by external compliance officers for the verification of fairness of underwriting practices of a bank over time, but also by internal ones, i.e., officers of the bank itself.

The blockchain-based trust management module in PP-ETF framework is of significant practical value for financial institutions. The framework supports auditability and facilitates the analysis after the fact of a model's behavior and of the decisions it made. The automated execution of smart contracts for the management of the reputation of participants in learning processes greatly facilitates the management of governance rules.

However, it should also be mentioned that 480 transactions are processed per second (TPS) and the finality latency for the ledger in question is 2.4 s, on average. On the basis of such performance figures, the assessment of credit risk by means of the PP-ETF within a batch-oriented process will be quite feasible, for example. As a result, the modular design of the PP-ETF allows for easy updates for single components (e.g., privacy-preserving cryptography, XAI, DLT) in the future. As a result, financial institutions will also benefit from their investment in the PP-ETF in the future.

6.5 Limitations and future research directions

The PP-ETF is a work in progress. At the current stage it has been tested using a variety of different metrics, yet it remains to be seen how it will perform on large datasets of real world financial records. The financial datasets that have been used for the evaluation of the PP-ETF in this work, the German Credit data set and the HELOC data set, together contain some 11,000 records. By comparison there are millions and millions of credit records for individual consumers maintained by financial institutions around the world. Hence it would be very useful to test the performance of the PP-ETF on a number of large proprietary financial data sets.

Scaling to Enterprise-Scale Datasets. To contextualize the scalability characteristics of the PP-ETF for enterprise deployment, we provide an analytical projection of how computational and communication overhead scales with dataset size and consortium size. We emphasize that these projections are analytical estimates based on the measured per-round costs from our experiments, not experimental results from enterprise-scale deployments.

Assumptions for Analytical Projections. The following assumptions underpin our scaling projections:

1. Model Dimensionality: The neural network model has $d = 10^6$ trainable parameters, represented as 32-bit

floating-point values, yielding a model size of approximately 4 MB per institution.

2. Number of Participating Institutions: The consortium comprises $N = 100$ financial institutions.
3. Local Mini-Batch Size: Each institution uses a mini-batch size of $b = 256$ samples per gradient computation.
4. Local Epochs: Each communication round comprises $\tau = 10$ local epochs, resulting in $T_{\text{local}} = \tau \cdot \lceil n_i/b \rceil$ gradient computations per institution per round, where n_i is the local dataset size.
5. Communication Protocol: Gradient updates are transmitted as Paillier ciphertexts. For a 2048-bit Paillier modulus, the plaintext space can represent integers up to approximately 2^{2048}. We encode each gradient component as a signed 16-bit fixed-point integer. Because a 16-bit fixed-point value occupies only 16 bits, it fits comfortably within a single 2048-bit Paillier plaintext. Therefore, under this encoding, one Paillier ciphertext is required per gradient value, not two. The ciphertext size is 4,096 bits (512 bytes). Thus, each gradient vector of $d = 10^6$ parameters requires approximately $10^6 \times 512$ bytes = 512 MB per institution per round. If higher precision (e.g., 32-bit fixed-point) or a split sign/magnitude representation were used, two ciphertexts per value could be necessary; however, our revised estimates use the more efficient single-ciphertext 16-bit encoding.
6. Compression: We assume no gradient compression is applied in the baseline projection. With top-k sparsification at $k = 1\%$, the communication volume reduces proportionally to approximately 5.12 MB per institution per round.

Computational Cost Projections. Under these assumptions:

- Per-Institution Computational Cost: Each gradient computation requires approximately $O(d \cdot b) = 10^6 \times 256 = 2.56 \times 10^8$ floating-point operations (FLOPs) for the forward pass and approximately twice that for the backward pass, yielding approximately 7.68×10^8 FLOPs per gradient computation. With $T_{\text{local}} = 10 \cdot \lceil n_i/256 \rceil$ gradient computations per round, and assuming $n_i = 10^5$ samples per institution, each institution performs approximately $10 \cdot 391 = 3910$ gradient computations per round, totaling 3.0×10^{12} FLOPs per institution per round.
- Total Consortium Computational Cost: Across $N = 100$ institutions, the total computational cost per communication round is approximately 3.0×10^{14} FLOPs. For $T = 100$ communication rounds, the total computational cost is approximately 3.0×10^{16} FLOPs.
- Paillier Encryption Overhead: Based on our measured encryption throughput of approximately 1.5 m per 1024-bit modular exponentiation on an NVIDIA V100 GPU, encrypting 10^6 integers (representing 10^6 floating-point values at 16-bit precision) requires approximately $10^6 \times 1.5$ ms = 1500 seconds = 25 min per institution per round. This dominates the computational cost and represents the primary scalability bottleneck.

Communication Cost Projections. Under these assumptions:

- Per-Institution Communication Volume: Each institution transmits approximately 512 MB of encrypted gradients per round. For 100 rounds, this totals approximately 51.2 GB per institution.
- Total Consortium Communication Volume: Across $N = 100$ institutions, the total communication volume per round is approximately 51.2 GB. For 100 rounds, the total communication volume is approximately 5.12 TB. With top-k sparsification at $k = 1\%$, these volumes reduce to approximately 5.12 MB per institution per round and 512 MB per round total, and 51.2 GB total over 100 rounds.

Distinction Between Analytical Projections and Experimental Results. We emphasize that the above figures are analytical projections derived from the measured per-round costs in our experimental configuration (10 institutions, 100 communication rounds, model dimensionality of approximately 10^4 parameters). The projections assume linear scaling of computational cost with model dimensionality and linear scaling of communication volume with both model dimensionality and the number of institutions. These projections have not been experimentally validated at enterprise scale and should be interpreted as indicative estimates rather than precise predictions.

Potential Mitigation Strategies. To address the scaling challenges identified in the projections, several mitigation strategies can be employed:

1. Gradient Compression: Top-k sparsification and quantization techniques can reduce communication volume by 10–100× while maintaining model quality. Our experiments with top-k sparsification at $k = 1\%$ demonstrate less than 1% degradation in AUC-ROC.
2. GPU-Accelerated Paillier: Our implementation achieves a 15× speedup over CPU-based Paillier implementations. Further optimizations, such as batch encryption and specialized modular arithmetic circuits, could achieve additional 2–5× speedups.
3. Hybrid Encryption: Encrypting only the most sensitive gradient components (e.g., those associated with protected attributes) while leaving other components in plaintext could reduce encryption overhead by 50%–80% with minimal impact on privacy.
4. Asynchronous Aggregation: Allowing institutions to contribute updates asynchronously can reduce synchronization overhead and improve convergence speed, particularly in heterogeneous network environments.

Pathways to Real-Time Credit Assessment. While the current PP-ETF implementation achieves a throughput of 480 TPS with 2.4-s finality latency—performance that is well-suited for batch-oriented credit risk assessment—real-time lending scenarios may require lower decision latency. Potential architectural extensions for future investigation include Layer-2 scaling solutions such as state channels and sidechains, which have demonstrated high throughput and low

latency in other blockchain applications. However, the adaptation of these solutions to the PP-ETF architecture, including the integration of privacy-preserving computation and explanation anchoring, remains an open research question. We therefore refrain from making specific performance claims about these extensions and instead identify them as promising directions for future work.

Human-Centered Validation. The current work does not include empirical validation of explanation usability through controlled user studies with financial professionals. While we have demonstrated the technical quality of the explanations (stability, completeness), their practical utility for loan officers, compliance officers, and risk managers requires empirical validation. Future research should include:

1. Co-design Workshops: Collaborative design sessions with financial professionals to understand their explanation needs and preferences.
2. Controlled Experiments: Randomized controlled trials measuring decision quality, trust calibration, and cognitive load when using the PP-ETF explanations versus traditional approaches.
3. Longitudinal Studies: Extended deployment studies to assess how explanation usage evolves over time and whether it improves decision consistency.

A possible limitation of the current framework is that it assumes a semi-honest aggregator. Thus, the framework can be vulnerable to attacks by a malicious aggregator. To counter such attacks, it is possible to use fully homomorphic encryption in order to aggregate the encrypted gradients. However, it is necessary to analyze the additional computational requirements of the used encryption scheme in detail in order to check if the framework is still deployable in practice.

However, for real-time-based lending processes there is the potential for delays while institutions communicate with each other and the blockchain, when using the framework for batch risk assessment. Therefore, future work will involve the optimization of the communication protocol for the framework. The authors will look into utilizing layer-2 scaling or a sidechain to offload the majority of the transactions from the main blockchain ledger, thus ensuring large throughput while maintaining the key characteristic of a decentralized system that the framework currently boasts.

Future work has to investigate the lower bound of the privacy budget. Hereby the relation between the explanation fidelity of the XAI module and the currently used noise-aware calibration protocol has to be analyzed in more detail. As already observed, the currently used relation seems to be non-linear and is very promising for future improvements.

This work focused on gradient poisoning attacks; however, other forms of malicious conduct exist. For instance, data poisoning that does not result in an anomaly in the gradients (as opposed to the typical usage case) as well as free-riding and sybil attacks need to be investigated. Also, as part of future work, different types of defense mechanisms could be developed to tackle such attacks and could be integrated into the framework.

6.6 Contribution to Sustainable Development Goals

The PP-ETF is able to help achieve several Sustainable Development Goals. The first goal for the PP-ETF is to support industry, innovation and infrastructure. As the PP-ETF supports the financial industry, supports the use of innovation such as AI, privacy preservation and blockchain, it supports infrastructure necessary for financial systems. The second goal is to reduce inequalities. By preventing and reducing lending discrimination and supporting financial inclusion for underserved populations, the PP-ETF helps to reduce inequalities. The third goal is to achieve peace and justice through strong institutions. The blockchain-based components of the PP-ETF support strong institutions, financial inclusion, transparency, and accountability in financial systems. The PP-ETF thus supports the creation and the sustainable growth of local financial ecosystems.

The PP-ETF will enable the sustainable development of innovative financial services and their implementation in a responsible manner. Above all, financial risk models will be created in a collaborative manner and be used in a decentralized manner. As a result, also small financial institutions from emerging markets and developing countries will be able to participate in global financial risk modeling in a robust, fair and sustainable manner.

Energy is another area where the increased use of AI and large amounts of data has significant environmental implications and therefore raises questions with respect to consumption and production. Therefore, the computational efficiency of the PP-ETF for the financial application studied here is also an important criterion for the sustainability of AI. By using partial homomorphic encryption and differential privacy the framework studied here is able to significantly reduce the communication required for financial risk assessment. In turn this reduction in communication has a direct impact on energy consumption and as a result the associated carbon emissions. Thus it supports responsible consumption and production.

In conclusion, the novel PP-ETF proposed for financial risk assessment in this paper will facilitate the use of Privacy-Preserving, Explainable and Trustworthy AI for sustainable development. This work will fulfill the demands of global market competition by providing a robust decision support system. The novel integrated XAI and PFTM and novel Blockchain-based Trust Management module can satisfy the fundamental values of AI, i.e. Privacy, Transparency and Accountability. Thus, the PP-ETF can extend the current state-of-the-art techniques to achieve the goals of Industry, Innovation and Infrastructure, Reduced Inequalities and Peace and Justice through strong institutions.

7 Conclusion and future work

This final section summarizes the main contributions and assesses the degree to which all the objectives set for this work have been met. It outlines several limitations of the current research and sets out several directions for future research that will be developed on the top of the architecture for the Privacy-Preserving Explainable Trust Framework (PP-ETF) for financial risk assessment that has been developed in this work.

7.1 Summary of contributions

The PP-ETF was designed to solve five research questions that concern the use of data in financial institutions. The first research question concerned the clients' privacy requirements and the need to share data within a learning environment. The design of a privacy-preserving federated learning model for credit risk prediction was therefore the first research question of the PP-ETF. The second research question concerned the need for explainable AI in finance and the design of an integrated explanation layer. The third research question concerned the challenge of trust in finance and the design of a blockchain-based trust management system. The fourth research question concerned the integration of all the components into one framework by validating the integrated framework against a set of existing baselines. The last research question concerned the challenges of the scalability of the framework and the trade-offs between privacy, utility, model complexity, communication, and the number of participants.

The results have shown that the framework can achieve prediction accuracy that is within 1.4%–1.6% of the centralized models in terms of AUC-ROC, while significantly outperforming all privacy-preserving baselines. The framework is able to protect the data against the gradient infiltration attacks as well as the membership inference attacks. The attack success rates are statistically not different from random guessing. In terms of the explanation aspect, the SHAP and LIME explanations are generated by the locally sanitized models before the global model aggregation. A noise-aware calibration protocol, which uses variance-bounded attribution intervals, is used to improve the stability of the generated explanations. The results have shown that the stability of the generated explanations is higher than 0.87 cosine similarity. In terms of the trust management, the framework uses a permissioned Hyperledger Fabric network to record the model integrity records, explanation anchoring records and the participants' reputation scores. The results have shown that the framework is able to process 480 transactions per second with a finality latency of 2.4 s. This means that the framework is practical for use in a batch-based credit risk assessment environment. An adversary participant is detected and excluded from the framework within two communication rounds by the automated smart contract enforcement.

Fourth, for the integrated PP-ETF, we compare it with the number of state-of-the-art approaches and a centrally managed deep neural network on two public financial datasets, the German Credit dataset and the publicly available data from a Home Equity Line of Credit (HELOC) lender. Our results consistently show that, for a given amount of privacy protection, the integrated PP-ETF significantly outperforms all the other privacy preserving approaches, with AUC-ROC improvements over the second best approach of 2.4% for the German Credit dataset and 3.7% for the HELOC dataset. Moreover, for the given amount of privacy protection, the integrated PP-ETF performs competitively with the centrally managed deep neural network, achieving AUC-ROC values that are within 1.4%–1.6% of the centralized upper bound. Fifth, we have analyzed the integrated framework in terms of scalability and, by systematically generating points on the privacy-utility frontier, we have revealed the point at which the additional increase in utility ceases to be of value. We found that this point occurs at a privacy budget of $\epsilon = 6$. Thus, our results also reveal that, for the given problem of financial risk assessment, an optimal trade-off between privacy and utility can be found.

In addition to fulfilling the objectives of this research, the PP-ETF introduces several fundamental changes to the current *status quo*. Firstly, it provides evidence that privacy, explainability, and trust need not be treated as mutually exclusive objectives under the evaluated conditions, but can be addressed by a single architecture if properly designed and implemented. Secondly, the PP-ETF presents a new class of noise-aware calibration protocols that can be applied to a wide range of private models, aiming to provide human-understandable explanations for the behavior of stochastic models. Thirdly, the framework presents a novel, practical, zero-knowledge verification layer for the reputation management of participants in blockchain-based consortium learning, thus solving an open problem for the last years.

7.2 Implications for policy and practice

Our results are of substantial practical relevance to a number of stakeholders, i.e. financial regulators, financial services practitioners, and financial services technology developers, working with credit risk management using artificial intelligence. For financial regulators the PP-ETF presents a system that fulfills high privacy standards (as required by data protection laws like the General Data Protection Regulation) without losing quality of decisions. It produces the most stable and the most faithful explanations even at low privacy budgets. Therefore, the PP-ETF is a very promising system to fulfill the requirements of financial services regulators and other data protection authorities, while preserving the core aspects of decisions taken using artificial intelligence for credit risk without revealing highly sensitive data as well as details of proprietary models. The audit trail stored on the blockchain can be used for monitoring of changes of the models as well as of the corresponding decision rationales in an efficient way. Thus, financial services regulators can perform their audits corresponding to the above issues significantly easier.

The framework also has the potential to level the playing field amongst banks and between banks and credit unions in terms of collective intelligence sharing in a learning process, whilst each member of the consortium retains control over their own customer data. A good set of gradients and truthful explanations can then become the property of the entire consortium, to be contributed by the best performer in the consortium, for the benefit of all its members. A number of mechanisms to facilitate this are described.

For technology developers the modular structure of the PP-ETF is a perfect base for further development of the individual components. Depending on the current development of technology, also the single components of the implemented framework can be optimized. Therefore the PP-ETF is also suitable for a broad application in the financial services sector.

7.3 Limitations and caveats

While considerable effort has been put into developing a number of methodologies that could be evaluated within the framework of the present research the latter is not without its

limitations, however. As it stands the research has for the most part relied upon a number of publicly available data sets, a couple of the most commonly used credit risk data sets (the German Credit data set and the HELOC data set) having been used for the purposes of the comparison between a number of different approaches to the problem of credit risk assessment. That the aforementioned data sets are, however, somewhat small when compared to those currently in use within many of the world's largest financial organizations is not particularly surprising given the fact that the data used in the two datasets used here are somewhat old. Datasets used in financial organizations these days are typically compiled from data sourced from a number of different countries and have been compiled over a number of years. The evaluation within the research of a number of the available approaches to credit risk using a number of the proprietary datasets (currently) used within financial organizations of considerable size would provide for a far more complete evaluation of the extent to which the credit risk provided by the PP-ETF could be considered to be of the highest quality and capable of use within a variety of different scenarios.

To deal with more malicious models, i.e. fully dishonest aggregators, the current PP-ETF's implementation of differential privacy is able to thwart attacks like gradient inversion attacks and membership inference attacks. However, a set of malicious participants might still combine their encrypted updates to learn something about other participants. To handle this, fully homomorphic encryption could be implemented in the aggregation phase of the PP-ETF. However, the increased computational overhead would have to be efficiently optimized for real-world applications.

Communication latency and transaction throughput: While PP-ETF already tackles the problems of communication latency and transaction throughput in a satisfactory manner for batch-oriented applications such as credit risk rating, for real-time applications like instant lending even higher transaction processing rates will be needed. Currently, the system is able to process about 480 transactions per second which is already sufficient for a batch-oriented application. However, for very fast online applications this number has to increase dramatically. For blockchain-based systems, Layer-2 scaling solutions like state channels or sidechains are suitable in order to raise the transaction throughput while keeping the main ledger secure. However, the specific performance characteristics of these solutions in the context of the PP-ETF architecture have not been experimentally evaluated in this work and remain an important direction for future research.

This would not be the case of an intelligent adversary that tries to adapt its behavior to remain undetected, for which more sophisticated methods of detection such as those used in anomaly detection, and in other fields of cybersecurity, would be needed.

7.4 Directions for future research

A number of limitations have been identified for the PP-ETF in this research. The primary limitation concerns the assumption of a semi-honest aggregator, which means that a malicious aggregator could potentially compromise the privacy of participants in the framework. Future work will therefore aim to extend the framework so that all communication within the framework is encrypted, using for example fully homomorphic encryption for the aggregation phase.

Another issue, which has to be investigated in future research, is the question if the PP-ETF can be applied to other fields like fraud detection, for the underwriting of insurance contracts, for supply chain finance or other fields. The modular structure of the PP-ETF seems to be very favorable for this purpose, since every module can be exchanged easily for corresponding other modules. A comparison of the results that are reached by the PP-ETF in the differences of the results and the difficulties that occur in the different applications should give hints how far the PP-ETF is able to outperform other approaches and in which application fields the use of the PP-ETF is most promising.

Thirdly, future research has to extend the module XAI of the PP-ETF towards causal explanations. So far SHAP values as well as the explanations provided by LIME only deliver only associations between the variables describing the applicants (features) and the outcome variables (target variables for the credit decision etc.). In many scenarios it is even more important to know the causal effects instead. The use of structural equation models as well as the already for a long time established analysis using instrumental variables would be suitable for such causal explanations. However, especially the provision of such causal explanations under privacy constraints is a challenge of considerable dimensions.

The fourth future research direction for the PP-ETF is for the development of a dynamic privacy budget allocation. Currently, the fixed privacy budget P is split equally between all rounds of the communication, i.e., for each round, the same portion of the total privacy budget is used for learning from all participants. This might not be efficient as some rounds may contain gradients of much higher value for learning the model than others. In order to improve the quality of the model that is learned with a fixed total privacy budget, future research should develop adaptive privacy mechanisms. Such privacy mechanisms can be learned using techniques from multi-armed bandit or reinforcement learning. The learned model estimates the value of information that is conveyed by the gradients that are uploaded by the participants in the learning process. Using these estimated values for information, the privacy budget P is then allocated to the individual participants for the subsequent rounds of communication.

Another direction is to study the game-theoretic properties of the reputation mechanism. Analyze the convergence of the system to an equilibrium where honest behavior is the dominant strategy for all participants. The study could be performed within the framework of evolutionary game theory. The results could be validated by means of agent-based simulations as well as by controlled experiments.

PP-ETF can be integrated with Decentralised Finance (DeFi) protocols to enable the first Trustless Credit Scoring (TCSS) on the Blockchain. The combination of PP-ETF's privacy and explainability features with the programmable and composable decentralized finance (DeFi) protocols can fundamentally change the nature of access to credit enabling permissionless lending that is transparent and fair. Two main challenges of such a system are (i) fully decentralized without the intermediary that holds the stakeholders' trust and (ii) sybil-resistance.

7.5 Human-computer interface (HCI) co-design

Another important direction is to develop Explainable AI Human-Computer Interfaces for Credit Decisions. All previous directions deal with the issue of explanation (interpretability). However, previous aspects of explanations (usability, accuracy,

human factor) have to be translated to interfaces. In addition, these interfaces have to be tested with users. For this purpose, a co-design with a large number of users of different ages and with different backgrounds (applicants, credit officers, compliance officers, risk managers etc.) in different settings (banks, fintechs, online lenders) for supporting their decision making is necessary. For designing such interfaces also other fields than AI can be of use, i.e. human–computer interaction and behavioral science.

7.6 Concluding remarks

We design the first architecture to reach new heights in the Responsible AI for Finance (RAIF) quest. A plethora of challenging objectives exist for the RAIF quest. In fact, some of the objectives are contradictory. We design the PP-ETF architecture to fulfill four contradictory objectives: (1) High quality credit risk prediction of credit risk, (2) Preservation of sensitive information of all participants, (3) Full transparency of the AI decision-making process, and (4) Auditable decentralized AI-driven trust management. We show that all of the above four objectives within a) can be fulfilled; b) are feasible with currently available AI techniques; and c) all objectives are fulfilled within one integrated architecture.

However, the research presented in this paper goes well beyond the assessment of the credit risk of borrowers using the PP-ETF. The main contribution of this research is the design of a novel generation of systems which support the use of AI in a wide array of domains, such as for instance finance, healthcare and social welfare. These domains have in common the generation of a large volume of information, which can be used for assessing risk, and the automated use of this information for making decisions. At the same time, however, the use of such information for decision making also raises serious concerns with respect to privacy, accountability and fairness.

The increasing use of AI in critical domains of life requires architectures that reconcile the capabilities of intelligent systems with ethical values and principles. The PP-ETF presented in this paper is the starting point for architectures of intelligent systems that are not only efficient but also just, transparent and trustworthy. It is up to future research to elaborate and extend the architecture presented here.

Data availability statement

The datasets analyzed in this study are publicly available. The German Credit Dataset is available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29. The Home Equity Line of Credit (HELOC) dataset is available from the FICO Explainable Machine Learning Challenge repository at https://community.fico.com/s/explainable-machine-learning-challenge. The complete source code implementing the Privacy-Preserving Explainable Trust Framework (PP-ETF) is publicly available on GitHub at https://github.com/drsanjayagal/pp-etf-framework. The repository contains all necessary scripts for reproducing the experimental results presented in this manuscript, including data preprocessing, federated learning with differential privacy, SHAP and LIME explainability modules, blockchain trust simulation, and the full experiment pipeline. The blockchain ledger data and raw experimental outputs are available upon reasonable request from the corresponding author, subject to institutional data governance policies and Parul University's research ethics guidelines. For any queries, the author can be contacted at sanjay.agal32685@paruluniversity.ac.in.

Author contributions

SA: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization, Supervision, Project administration.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

The author expresses sincere gratitude to the Department of Artificial Intelligence and Data Science at Parul University, Vadodara, Gujarat, India, for providing the academic environment and infrastructural support necessary for the successful completion of this research.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. Yes. Generative AI was used in the preparation of this manuscript for grammar correction and language improvement only. It was not used for scientific content generation, data analysis, or interpretation of results. The author reviewed and approved all AI-assisted edits.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/access.2018.2870052

Al-Rubaie, M., and Chang, J. M. (2019). Privacy-preserving machine learning: threats and solutions. *IEEE Secur. and Priv.* 17, 49–58. doi: 10.1109/msec.2018.2888775

Amadi, R. S., Kayes, A. S. M., Pardede, E., Chowdhury, M. J. M., and Ahmed, K. (2025). A comprehensive review of risk assessment frameworks in blockchain applications: research gaps and key lessons. *IEEE Access* 13, 131347–131377. doi: 10.1109/ACCESS.2025.3590963

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2019). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Bai, S., Yang, G., Rong, C., Liu, G., and Dai, H. (2020). QHSE: an efficient privacy-preserving scheme for blockchain-based transactions. *Future Gener. Comput. Syst.* 112, 930–944. doi: 10.1016/j.future.2020.06.025

Bellini, E., Iraqi, Y., and Damiani, E. (2020). Blockchain-based distributed trust and reputation management systems: a survey. *IEEE Access* 8, 21127–21151. doi: 10.1109/access.2020.2969820

Bhardwaj, T., and Sumangali, K. (2025). An explainable federated blockchain framework with privacy-preserving AI optimization for securing healthcare data. *Sci. Rep.* 15, 21799. doi: 10.1038/s41598-025-04083-4

Bochner, R. (2026). Credit risk assessment: a privacy-preserving framework integrating shapley deep networks with blockchain verification. *Int. J. Automation Digital Transformation* 5, 31–53. doi: 10.54878/mqd8ce53

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2020a). Explainable AI in fintech risk management. *Front. Artif. Intell.* 3, 26. doi: 10.3389/frai.2020.00026

Bussmann, N., Giudici, P., Marinelli, D., and Papenbrock, J. (2020b). Explainable machine learning in credit risk management. *Comput. Econ.* 57, 203–216. doi: 10.1007/s10614-020-10042-0

Campbell-verduyn, M., and Goguen, M. (2018). Blockchains, trust and action nets: extending the pathologies of financial globalization. *Glob. Netw.* 19, 308–328. doi: 10.1111/glob.12214

Černevičienė, J., and Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artif. Intell. Rev.* 57, 216. doi: 10.1007/s10462-024-10854-8

Dumitrescu, E., Hué, S., Hurlin, C., and Tokpavi, S. (2021). Machine learning for credit scoring: improving logistic regression with non-linear decision-tree effects. *Eur. J. Operational Res.* 297, 1178–1192. doi: 10.1016/j.ejor.2021.06.053

Feki, I., Ammar, S., Kessentini, Y., and Muhammad, K. (2021). Federated learning for COVID-19 screening from chest X-ray images. *Appl. Soft Comput.* 106, 107330. doi: 10.1016/j.asoc.2021.107330

Ghosh, I., Alfaro-Cortés, E., Gámez, M., and García-Rubio, N. (2023). Prediction and interpretation of daily NFT and DeFi prices dynamics: inspection through ensemble machine learning and XAI. *Int. Rev. Financial Analysis* 87, 102558. doi: 10.1016/j.irfa.2023.102558

Gong, M., Pan, K., and Xie, Y. (2019). Differential privacy preservation in regression analysis based on relevance. *Knowledge-Based Syst.* 173, 140–149. doi: 10.1016/j.knosys.2019.02.028

Goodell, J. W., Jabeur, S. B., Saâdaoui, F., and Nasir, M. A. (2023). Explainable artificial intelligence modeling to forecast bitcoin prices. *Int. Rev. Financial Analysis* 88, 102702. doi: 10.1016/j.irfa.2023.102702

Gunnarsson, B. R., Vanden Broucke, S., Baesens, B., Óskarsdóttir, M., and Lemahieu, W. (2021). Deep learning for credit scoring: do or don't? *Eur. J. Operational Res.* 295, 292–305. doi: 10.1016/j.ejor.2021.03.006

He, H., Wang, Z., Jain, H., Jiang, C., and Yang, S. (2022). A privacy-preserving decentralized credit scoring method based on multi-party information. *Decis. Support Syst.* 166, 113910. doi: 10.1016/j.dss.2022.113910

Jovanovic, Z., Hou, Z., Biswas, K., and Muthukkumarasamy, V. (2024). Robust integration of blockchain and explainable federated learning for automated credit scoring. *Comput. Netw.* 243, 110303. doi: 10.1016/j.comnet.2024.110303

Kotadiya, U., Yachamaneni, T., and Arora, A. S. (2026). Block chain audited homomorphic encryption for consortium credit risk modelling. *Lect. Notes Networks Systems*, 410–433. doi: 10.1007/978-3-032-14189-7_35

Kuang, X., Ma, C., and Ren, Y.-S. (2024). Credit risk: a new privacy-preserving decentralized credit assessment model. *Finance Research Letters* 67, 105937. doi: 10.1016/j.frl.2024.105937

Liu, Z., and Wu, H. (2025). A framework for dynamic blockchain-based data auditing. *Int. J. Account. Inf. Syst.* 56, 100737. doi: 10.1016/j.accinf.2025.100737

Ma, Y., Zhang, P., Duan, S., and Zhang, T. (2023). Credit default prediction of Chinese real estate listed companies based on explainable machine learning. *Finance Research Letters* 58, 104305. doi: 10.1016/j.frl.2023.104305

Makhdoom, I., Zhou, I., Abolhasan, M., Lipman, J., and Ni, W. (2019). PrivySharing: a blockchain-based framework for privacy-preserving and secure data sharing in smart cities. *Comput. and Secur.* 88, 101653. doi: 10.1016/j.cose.2019.101653

Naresh, V. S., and Ayyappa, D. (2024). PPDNN-CRP: privacy-preserving deep neural network processing for credit risk prediction in cloud: a homomorphic encryption-based approach. *J. Cloud Comput. Adv. Syst. Appl.* 13, 72. doi: 10.1186/s13677-024-00711-y

Nwafor, C. N., and Nwafor, O. Z. (2023). Determinants of non-performing loans: an explainable ensemble and deep neural network approach. *Finance Research Letters* 56, 104084. doi: 10.1016/j.frl.2023.104084

Qiao, Y., Lan, Q., Zhou, Z., and Ma, C. (2021). Privacy-preserving credit evaluation system based on blockchain. *Expert Syst. Appl.* 188, 115989. doi: 10.1016/j.eswa.2021.115989

Raza, A., Badidi, E., Hayajneh, M., Barka, E., and Harrouss, O. E. (2024). Blockchain-based reputation and trust management for smart grids, healthcare, and transportation: a review. *IEEE Access* 12, 196887–196913. doi: 10.1109/access.2024.3521428

Ruan, W., Xu, M., Jing, Y., and Han, W. (2022). Towards understanding the fairness of differentially private margin classifiers. *World Wide Web* 26, 1201–1221. doi: 10.1007/s11280-022-01088-1

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017). "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. doi: 10.1109/SP.2017.41

Varma, J. R. (2019). Blockchain in finance. *Vikalpa J. Decis. Makers* 44, 1–11. doi: 10.1177/0256090919839897

Wang, Y.-R., and Tsai, Y.-C. (2022). The protection of data sharing for privacy in financial vision. *Appl. Sci.* 12, 7408. doi: 10.3390/app12157408

Xu, Z., Cheng, J., Cheng, L., Xu, X., and Bilal, M. (2023). MSES credit risk assessment model based on federated learning and feature selection. *Comput. Materials and continua/Computers, Materials and Continua (Print)* 75, 5573–5595. doi: 10.32604/cmc.2023.037287

Yang, F., Qiao, Y., Abedin, M. Z., and Huang, C. (2022). Privacy-preserved credit data sharing integrating blockchain and federated learning for industrial 4.0. *IEEE Trans. Industrial Inf.* 18, 8755–8764. doi: 10.1109/TII.2022.3151917

Yang, F., Abedin, M. Z., and Hajek, P. (2023). An explainable federated learning and blockchain-based secure credit modeling method. *Eur. J. Operational Res.* 317, 449–467. doi: 10.1016/j.ejor.2023.08.040