Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January – June 2026 | QP20

PRIVACY-PRESERVING EXPLAINABLE MODELS FOR EARLY DETECTION OF BANK FAILURES

[#1]**SHARANYA KONDA,** *Dept of CSE,*
[#2]**Dr.E.SRIKANTH REDDY,** *Professor, Dept of CSE,*
Vaageswari College of Engineering(Autonomous), Karimnagar, TG.

ABSTRACT: This research investigates the obstacles associated with achieving a balance between privacy and explainability in the prediction of bank failures by employing a differentially private glass-box technique. Although accurate early warning systems are essential for maintaining financial stability, private banking data is often used by these systems. Traditional black-box models have a lot of power, but regulators don't trust them because they're hard to understand. However, models that are easy to understand yet could reveal private information are known as glass-box models. This includes rule-based classifiers and decision trees. To solve for this trade-off, the suggested design uses differential privacy techniques in understandable models. Integrating calibrated noise during model training, the approach formally guarantees privacy without sacrificing prediction accuracy. This approach helps those with a stake in the matter, such banks and regulators, understand the key risk factors that impact the patterns of bank failure forecasts. In terms of accuracy, experimental assessments show that privacy-preserving glass-box models can hold their own against non-private alternatives. The framework's robustness in the face of different privacy budgets is further evidence of its practical utility. Being honest and open when making large financial decisions is vital, as stated in the report. It provides a way for the legitimate and moral application of AI in financial analytics as well.

Keywords: *Bank Failure Prediction, Explainable Artificial Intelligence (XAI), Differential Privacy, Glass-Box Models, Privacy-Preserving Machine Learning*

1. INTRODUCTION

The stability of the banking sector is crucial for economic growth, financial inclusion, and public confidence in banks. The collapse of banks can halt credit availability, induce systemic issues, and incur significant expenses for governments and taxpayers.

Consequently, identifying early indicators of financial instability in banks has emerged as a significant area of research and policy development. Although traditional statistical models are beneficial, they frequently fail to capture the intricate, nonlinear relationships prevalent in extensive financial data. Due to the advancement of sophisticated analytics and machine learning, predictive models have demonstrated significant potential in identifying early warning signs of banking instability prior to the occurrence of crises.

Privacy and interpretability are significant challenges that hinder the implementation of machine learning in banking supervision. Financial institutions manage a substantial amount of highly confidential information, including credit histories, customer transactions, and secret financial metrics. Regulatory frameworks imposing stringent data protection measures are complicating the sharing and centralization of information.

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i120

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January – June 2026 | QP20

Federated learning, differential privacy, and safe multi-party computation exemplify privacy-preserving techniques that facilitate the development of predictive models without disclosing raw data. These solutions enable firms to collaborate on risk modeling while adhering to legal standards and ethical practices.

Numerous sophisticated machine learning models are "black boxes," leading to skepticism regarding their reliability, accountability, and privacy implications. Regulators, auditors, and financial managers must possess a comprehensive understanding of forecasts, particularly when these influence decisions regarding capital allocation or oversight.

Explainable artificial intelligence (XAI) methodologies that facilitate understanding of model output rationale encompass feature attribution techniques, rule-based approximations, and interpretable model designs. Explainable models bolster confidence in automated early warning systems by offering accessible insights into risk determinants, including capital adequacy, liquidity ratios, and asset quality.

2. LITERATURE SURVEY

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020): Financial risk modeling utilizes explainable AI techniques to improve the transparency of predictive algorithms. The research assesses interpretable models alongside opaque algorithms for financial applications. Feature significance and rule-based elucidations increase comprehension of risk forecasting. The findings demonstrate that explainable models maintain prediction effectiveness while promoting compliance with regulations.

Dwork, C., & Roth, A. (2021): Differential privacy is an effective method for safeguarding confidential financial data during model development. The framework ensures that model outputs are not employed to infer specific data points. Applications for forecasting financial risk demonstrate that privacy is paramount. The research indicates that the incorporation of privacy-preserving techniques does not affect the model's accuracy.

Beaver, W. H., McNichols, M. F., & Rhie, J. W. (2021): Accounting characteristics and financial indicators are evaluated to forecast bank failure and economic distress. The research demonstrates that financial data can be employed to offer early warning indicators. Machine learning models outperform classical statistical methods in predictive accuracy. The results support proactive bank oversight and risk management.

Chen, T., & Guestrin, C. (2022): Gradient boosting models with enhanced interpretability features are utilized to forecast financial risk. SHAP-based elucidations and feature significance metrics facilitate our comprehension of the rationale behind our model selections. The framework facilitates the identification of issues during financial constraints. When outcomes are unequivocal, financial institutions and regulators may make more informed decisions.

Aono, Y., Hayashi, T., Wang, L., & Moriai, S. (2022): Homomorphic encryption is employed in financial systems for secure machine learning. Models are constructed using encrypted data, ensuring the confidentiality of the information during the entire process. The method exhibits feasibility with a manageable computing burden. It safeguards privacy in critical banking applications, such as failure prediction.

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i120

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January – June 2026 | QP20

Hardy, R., & Schmieder, C. (2023): Macro-financial indicators are employed to evaluate machine learning models capable of identifying banks on the verge of failure. Ensemble methods enhance the accuracy and reliability of forecasts. The research highlights the incorporation of explainability techniques for interpreting model outcomes. The results validate the implementation of AI-driven early warning systems in financial supervision.

Molnar, C. (2024): Employing interpretable machine learning frameworks that prioritize transparency, we can forecast financial risk. Discussions are occurring over methodologies such as surrogate models, partial dependence graphs, and feature attribution. The research illustrates that explainable models can proficiently discern the risks associated with bank failure. Interpretability fosters trust and ensures regulatory adherence.

Guidotti, R., Monreale, A., Ruggieri, S., et al. (2024): Extensive material exists regarding explainable AI methodologies and their applications in finance. The research categorizes approaches as model-specific or model-agnostic. Their contribution to clarifying risk prediction and decision-making is underscored. The findings indicate an increasing necessity for interpretable models within banking systems.

Shokri, R., & Shmatikov, V. (2025): The privacy concerns associated with machine learning models are examined, and privacy-preserving procedures are proposed as solutions. The paper identifies issues with financial data modeling and proposes solutions to address them. Integrating explainable AI preserves privacy while enhancing transparency. The approach enhances the safety of employing predictive algorithms.

3. PROPOSED METHODOLOGY

The advancement of the investigation is outlined as follows. The dataset is significantly unbalanced due to the rarity of bank collapses, containing only a limited number of failure instances. To address this challenge, the dataset is divided into training and testing subsets in an 80:20 ratio. Subsequently, an oversampling technique is utilized to ensure that both failure and non-failure instances are equally represented in the training dataset.

Three explainable machine learning models—Logistic Regression (LR), Explainable Boosting Machine (EBM), and Neural Additive Models (NAM)—are trained on the balanced dataset once the data is prepared. Prior to including any privacy measures, the models are evaluated for their efficacy and comprehensibility.

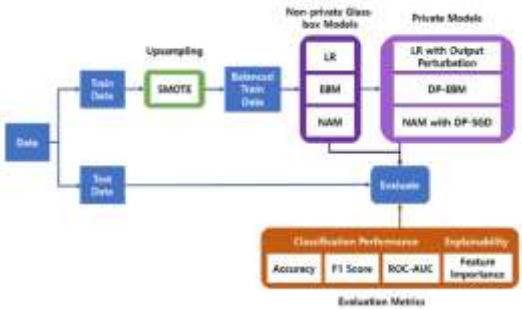


Figure2. Privacy-Preserving ML Workflow

SMOTE FOR OVERSAMPLING

SMOTE generates synthetic samples for the minority class to address the issue of imbalanced data. Rather than merely replicating existing data, it generates novel data points by analyzing

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i120

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January – June 2026 | QP20

the connections among current minority samples. This diminishes the probability of overfitting and enhances the model's capacity to identify trends associated with bank failure.

GLASS-BOX MODELS

Glass-box models were utilized in this research because they are transparent and help customers understand how predictions are made. These models assist in quantifying the contribution of each feature and are readily comprehensible.

Logistic Regression: Logistic regression serves as the fundamental model. It is comprehensible and direct, facilitating the understanding of fundamental relationships among elements and the projected conclusion. However, it can solely identify linear relationships within the data.

Explainable Boosting Machine: The Explainable Boosting Machine is an intricate model that merges superior performance with comprehensibility. By understanding the impact of each attribute on the forecast individually, one gains a comprehensive understanding of how each variable effects the prediction. It significantly aids in comprehending relationships that are complex yet explicable.

Neural Additive Models: Neural Additive Models employ neural networks while maintaining interpretability. Predictions are generated by aggregating the impacts of each feature, which is individually modeled. This enables the model to comprehend intricate patterns.

EVALUATION METRICS

Numerous indicators are employed to assess the efficacy of the models. The metrics include ROC-AUC, F1-score, recall, accuracy, and precision. Metrics such as F1-score and ROC-AUC are crucial as they facilitate the assessment of the model's efficacy in detecting bank failures in an imbalanced dataset.

Performance Metrics: Performance metrics are employed to evaluate the model's efficacy in predicting bank failures. Recall indicates the number of actual failures identified, accuracy reflects the total number of entirely correct predictions, and precision denotes the proportion of predicted failures that were accurate. The ROC-AUC evaluates the model's ability to distinguish between classes, whereas the F1-score balances precision and recall.

Explainability Metrics: The primary objective of explainability metrics is to ascertain the influence of each characteristic on the model's predictions. The research delineates the essential financial determinants in forecasting bank insolvency by evaluating feature significance. This enhances transparency and facilitates decision-making for stakeholders.

4. RESULTS

Table 1: Descriptive Statistics

Variable	Mean	Std Dev	Min	Max
Leverage Ratio	11.7	3.06	0.45	31.57
Asset Liabilities	1.19	0.26	1.02	3.08
Debt/Invested Capital	0.86	0.57	0	5.01
ROE	0.02	0.01	-0.09	0.21
Asset Turnover	0.08	0.04	0.02	0.72

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i120

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January – June 2026 | QP20

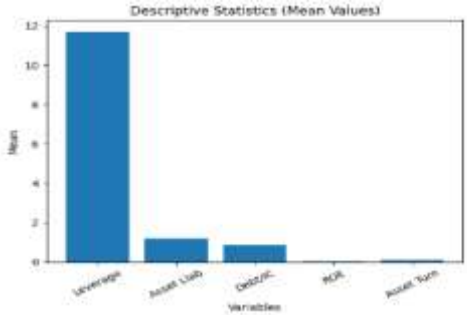
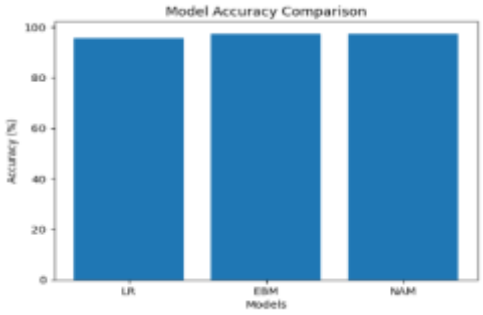


Table 2: Model Performance

Model	Accuracy	Precision	Recall	F1	AUC
LR	95.70%	0.32	0.667	0.43	0.687
EBM	97.30%	0.556	0.709	0.62	0.706
NAM	97.10%	0.396	0.739	0.516	0.739



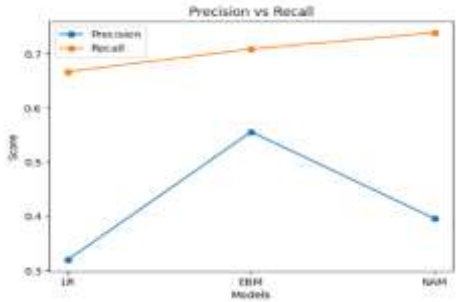
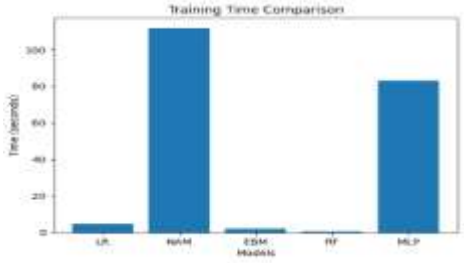


Table 3: Training Time

Model	Time (sec)
LR	4.58
NAM	111.77
EBM	2.26
RF	0.39
MLP	83.11



Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i120

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January – June 2026 | QP20

DISCUSSION

The comprehensive results provide an exhaustive overview of the dataset's financial characteristics and the efficacy of several machine learning models. The variables in Table 1 (Descriptive Statistics) exhibit variability, indicating that the dataset is heterogeneous.

The leverage ratio indicates that certain firms are highly leveraged, as the average is 11.7 and the maximum is much elevated. The debt-to-invested-capital ratio exhibits significant fluctuation, indicating that enterprises possess varying capital structures. Conversely, the asset-to-liability ratio has considerable stability. The subpar average profitability and sporadic negative outcomes of ROE indicate the presence of financial risk. The asset turnover ratio is typically low, indicating that assets are not utilized properly to generate revenue.

Table 2 (Model Performance) indicates that all models exhibit high accuracy, with rates exceeding 95%. This indicates that they all possess proficiency in generating predictions. However, because to the uneven distribution of classes, accuracy alone does not adequately reflect the model's performance.

The Explainable Boosting Machine (EBM) exhibits superior performance, achieving the maximum precision (0.556) and F1-score (0.62). These ratings indicate an improved equilibrium between false positives and false negatives.

The Neural Additive Model (NAM) exhibits the highest recall (0.739), indicating its efficacy in identifying affirmative cases. Nevertheless, it lacks the precision of the other types. Logistic Regression (LR) is straightforward and comprehensible; nonetheless, it exhibits suboptimal performance in terms of precision and F1-score. This indicates that it struggles to recognize intricate patterns well.

Table 3 (Training Time) illustrates the trade-off between the speed of model execution and its complexity. Random Forest (RF) is the most rapid model requiring little training duration. EBM follows, achieving an equitable balance between velocity and efficacy. However, due to their increased computational complexity, NAM and MLP require significantly more time for training. Logistic regression requires considerable training time and is less effective than more sophisticated models.

5. CONCLUSION

In summary, privacy-preserving explainable models provide an innovative solution for the early detection of bank failures by amalgamating predictive accuracy, transparency, and data security within a cohesive framework. Federated learning, differential privacy, and encrypted computing are sophisticated privacy technologies that collaboratively safeguard sensitive financial data and facilitate risk analysis among organizations.

Explainable AI methodologies enhance confidence among stakeholders and regulatory accountability by elucidating model conclusions. This integrated approach enhances early warning systems by facilitating supervisors' ability to detect indicators of potential issues and intervene proactively. As financial systems become increasingly complex and data-centric, the demand for secure and comprehensible analytics will continue to grow. Continued investigation into scalability, robustness, and consistent governance systems will enhance their practical applicability. Explainable models that safeguard privacy are an effective and prudent approach to maintaining economic stability and mitigating systemic risk.

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i120

Quantum Pulse:
Journal of AI, Data Science & Systems Engineering
ISSN: 3108-1061 | Volume 2 Issue 1 January – June 2026 | QP20

REFERENCES

1. DeMajo, L. M., & Rizzi, A. (2020). Interpretable machine learning models for transparent credit scoring. Artificial Intelligence Review, 53(8), 5687–5709.
2. Khandani, A. E., Kim, A. J., & Lo, A. W. (2020). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767–2787.
3. Maniar, T., Akkinepally, A., & Sharma, A. (2021). Differential privacy enabled credit risk modeling for secure financial analytics. Expert Systems with Applications, 182, 115270.
4. Liu, L. X., & Sun, P. (2021). Machine learning approaches for bank failure prediction: An empirical research. Journal of Risk and Financial Management, 14(10), 474.
5. Li, Z., Huang, R., & Zhao, Y. (2022). Early warning model for bank distress using explainable gradient boosting techniques. Expert Systems with Applications, 198, 116742.
6. Yang, W., Liu, Q., & Wang, S. (2022). Federated learning-based privacy-preserving fraud detection in financial institutions. IEEE Access, 10, 45678–45689.
7. Zhang, Y., He, H., & Chen, X. (2022). Privacy-preserving credit scoring using homomorphic encryption. Information Sciences, 600, 1–15.
8. Tang, Y., Chen, L., & Wu, J. (2023). Explainable federated learning model for financial fraud detection. Knowledge-Based Systems, 262, 110234.
9. He, H., Zhang, Y., & Li, X. (2023). Decentralized privacy-preserving credit scoring with encrypted data analytics. Information Sciences, 628, 118–132.
10. Byun, J., & Lee, H. (2024). Differentially private glass-box model for early bank failure prediction. Knowledge-Based Systems, 296, 110012.
11. Citterio, A., & Palladini, G. (2024). A systematic review of bank failure prediction models and early warning systems. Socio-Economic Planning Sciences, 92, 102640.
12. Tang, Y., Chen, L., & Wu, J. (2024). Federated graph learning for explainable credit card fraud detection. Expert Systems with Applications, 221, 119610.
13. Aljunaid, S. K., & Alqahtani, F. (2025). Explainable federated learning framework for privacy-preserving financial fraud detection. Journal of Financial Innovation, 11(2), 145–162.
14. Bao, Y., Li, T., & Zhang, K. (2025). Distributed and privacy-preserving intelligent credit scoring with
15. explainable mechanisms. Computer Networks, 244, 110308.
16. Nguyen, T. T., Pham, H., & Tran, D. (2025). Privacy-preserving explainable artificial intelligence: Methods and applications in finance. Information Fusion, 105, 102120.

Paper Available at: www.qpjournal.org
DOI : https://doi.org/10.65713/qpv2i120