

Agentic AI-Based Decision Support Architecture for Financial Risk Management

Habib Ur Rehman

habibraja383@gmail.com

0009-0007-8544-1401

Department of Data Science, University of Kotli, AJ&K

Sohail Akhtar

sohail.akhtar@uokajk.edu.pk

0009-0002-2283-0243

Department of Data Science, University of Kotli, AJ&K

Fatima Mustaqeem

fatimamustaqeem123@gmail.com

0009-0001-4628-9087

Department of Data Science, University of Kotli, AJ&K

Sayed Zahra Seerat

seeratzahra708@gmail.com

0009-0003-8819-2464

Department of Data Science, University of Kotli, AJ&K

Corresponding Author: Habib Ur Rehman habibraja383@gmail.com

Received: 16-02-2026

Revised: 08-03-2026

Accepted: 16-03-2026

Published: 27-03-2026

Abstract

Financial risk management is an important task in banking, and At the institutional level, where decisions are accurate and transparent and have direct consequences on the institution. stability. Traditional machine learning models have been shown to perform well on Less ability to grasp the more intricate, non-linear financial patterns data and rarely interpret results in the domain practitioners. This paper introduces FinRisk.ai, an agentic Multi-agent collaborative - AI framework, integrating deep learning in a multi-agent collaborative setting The pipeline was used for the automated financial risk prediction. The framework contributes: (1) a sequential feature construction method which Reshapes static table credit reports into temporal windows in them. Each (2) a is the LSTM, a leakage-safe crossvalidation protocol used within each. The features' stability during the formal definition of the explainability metrics. Stability of features within the formal definition of explainability metrics. The Index (FSI) and Explanation Fidelity Score (EFS) are included in a (3) an autonomous inter-agent communication and (4) a dedicated XAI agent. Use of a Judge Agent for protocol with conflict resolution. A Human-in-the-Loop (HITL) module enables domain experts to accept, reject, or request reevaluation of any prediction. Ethical privacy-related factors such as GDPR compliance, fairness auditing and Basel III alignment are addressed. The system is evaluated on the Both datasets were applied to the same approach. Both the German Credit Dataset and the Credit Card Fraud Detection Dataset were used in the same manner. The data was split into five nonoverlapping sections and the test was repeated three times, giving a total of 15 splits. runs). The LSTM achieves a mean accuracy of 93.7%(±0.8%) and AUC of 0.963(±0.007) on credit risk prediction, beating XGBoost by 6.5TabNet by 4.0%. The LSTM has an accuracy of 99.2A maximum accuracy of 0.987 was obtained for the AUC.

Keywords: AI Deep Learning, LSTM, ANN, Explainable AI, SHAP, LIME, Human-in-the-Loop, Sequential Feature Construction, Decision Support System

INTRODUCTION

Financial institutions process large volumes of customer data daily, including credit histories, transaction records, and loan repayment behavior. Managing risks from this data constitutes both a technical and regulatory challenge [1]. A single miscalculated credit decision or an undetected fraudulent transaction can result in considerable financial losses.

Traditional rule-based and statistical models proved inadequate for the complexity of modern financial data [2]. Supervised machine learning models improved predictive accuracy but treat each prediction in isolation, cannot reason across multiple risk dimensions, and provide no interpretable justification for their outputs [3].

The emergence of agentic AI systems, where multiple intelligent agents collaborate autonomously toward a shared goal, has opened a productive direction in financial technology [4]. Okpala et al. [5] demonstrated a CrewAI-based multi-agent system achieving above 93% accuracy, but without deep learning, formal explainability, or structured human validation. Recent LLM-driven approaches [14]–[16] show potential on large financial corpora but require substantial pretraining resources. FT-Transformer [17] has shown competitive performance on tabular data yet has not been evaluated within agentic or HITL integrated pipelines.

FinRisk.ai addresses these gaps by providing an integrated applied AI framework that combines sequential feature construction for LSTM, a leakage-safe agentic pipeline, formal explainability evaluation via FSI and EFS metrics, an autonomous inter-agent protocol with conflict resolution, and a quantified HITL module. While FinRisk.ai does not propose a new learning algorithm or agent coordination mechanism, its contribution lies in the principled integration of these components into a single validated system a combination not present in any prior work.

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the proposed methodology. Section IV presents the system architecture. Section V covers experiments and results. Section VII provides a critical discussion of key design decisions, limitations, and statistical evaluation. Section VIII concludes the paper.

RELATED WORK

There has been a lot of research interest in financial risk prediction. That's certainly something that has been on our radar for the last ten years. Early work was based on statistical and rule-based methods, although interpretable, were found to be ineffective. Not suitable for non-linear financial data [6]. Machine learning made improvements in accuracy, but challenges related to explainability, multi-dimensional risk Reasoning and human monitoring continued to be negligible. unresolved [7].

Further more recent research investigates the use of LLM for financial agents [14]–[16]. Lopez-Lira and Tang [14] Sentiment signals from ChatGPT can forecast short-term stock price movements, according to a new study. To address this, stock movements were introduced by Feng et al. [15] to enrich the text with the specific context it was used in. For a language modeling framework for financial data. Yu et al. [16] proposed a multimodal LLM-based agent for trading named FinAgent. However, these methods are directed towards tasks of language and market rather than credit scoring or fraud detection, and are not validated on Traditional financial risk indicators. Gorishniy et al. [17] demonstrated that FTTransformer is competitive It is used on tabular data but has not been tested with multi-agent or Financial risk pipelines integrated in HITL. The summary of key studies in this area is given in Table I.

Four research gaps emerge from this review: (1) most existing work treats financial risk as a single-model classification problem [7]; (2) ANN and LSTM have rarely been combined with a multi-agent framework [9]; (3) formal explainability evaluation is almost entirely absent from existing systems [3];

(4) HITL validation with measurable decision-improvement statistics remains largely absent from automated financial risk pipelines [5]. FinRisk.ai addresses all four gaps in a single validated system.

PROPOSED METHODOLOGY

The proposed methodology will be implemented in a structured pipeline, starting from raw data. Financial information to a validated decision support system. All The preprocessing steps are applied strictly inside each training Do not fold to avoid leakage. Fig. 1 Describes the system pipeline.

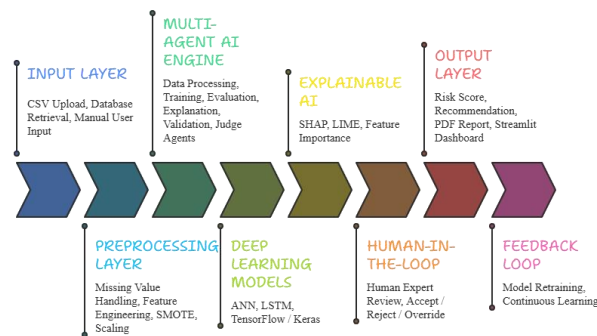


Fig. 1. FinRisk.ai system pipeline showing seven stages: input layer, preprocessing, deep learning models, multi-agent AI engine, explainable AI, humanin-the-loop, and feedback loop

Data Collection

Two publicly available datasets are used. The German Credit Dataset [12] has 1000 instances. Records containing 20 features like credit history, loan amount, employment etc. The labels included a binary good/bad credit risk label and a duration. The dataset *Credit Card Fraud Detection Dataset* citeb13 is presented. Has 284,807 transactions and 30 anonymized PCA-transformed These features are valid; only summarizes both datasets.

Preprocessing and Feature Construction

Each train fold is used for all preprocessing after stratified random splitting: KNN Imputation for categorical features, and Imputation of missing values for categorical variables and ordinal encoding, and StandardScaler normalization fitted only on training data. For the fraud detection dataset, train-test splitting is done as follows: performed before an application of SMOTE; SMOTE is applied after. only to the training fold but not to any synthetic samples from test data pollute training [5]. Transaction ordering is preserved in terms of the original PCA transformed. no temporal leakage - sequence; is not included before any transformations. Feature selection with SelectKBest with ANOVA F-scores. relying on the training data per fold.

In the case of the German Credit Dataset, a sequence is a sequence of credit cards used over a sliding window of transactions. The 20-dimensional feature vector is applied to the construction. were split into 4 time steps $T = 5$ (ordered), each with 5 features, following four semantic groups: (1) credit history, (2) account and the savings status, (3) employment features and (4) loans. details. This makes every record into shape (4,5) so that it looks like this: The LSTM input enables the model to capture cross-step. The LSTM input allows the model to capture cross-step. All other dependencies that are not available for flat feature representations [9], [10].

Model Configuration

ANN: Three fully connected layers (128, 64, 32 neurons), ReLU activation, batch normalization, Dropout(0.3), sigmoid output, Adam optimizer ($lr = 0.001$), binary cross-entropy loss.

LSTM: Two stacked LSTM layers (64 units each) on $T = 4$ sequences, dense sigmoid output. Both models use early stopping (patience = 10 epochs) and a 15% validation split. All hyperparameters were selected via grid search over learning rate $\{0.001, 0.01\}$, dropout $\{0.2, 0.3\}$, and hidden units $\{32, 64, 128\}$ on the training folds only. Models are evaluated under stratified 5-fold CV repeated three times (15 total runs); mean $\pm$ standard deviation is reported for all metrics. Pairwise McNemar's tests are conducted between LSTM and each baseline; all reported improvements are significant at $p < 0.05$. Algorithm 1 summarizes the complete training procedure.

Explainability Metrics

SHAP and LIME are integrated within the Explanation Agent. Two formal metrics assess explainability quality [10]: Feature Stability Index (FSI):

$$FSI = \frac{1}{\binom{K}{2}} \sum_{i < j} \rho_s \left(\mathbf{r}_i^{(k)}, \mathbf{r}_j^{(k)} \right) \quad (1)$$

where ρ_s is the Spearman rank correlation between SHAP feature rankings across folds. $FSI \in [0, 1]$.

Table 1: Literature summary on financial risk management using ai/ml

Ref	Title	Year	Method	Key Findings	Limitations
[5]	<i>Agentic AI Systems Applied to Tasks in Financial Services: Modeling and Model Risk Management Crews</i>	2025	Multi-Agent, XGBoost, CatBoost	93–99% accuracy on credit and fraud datasets	No deep learning; no SHAP/LIME; no HITL workflow
[7]	<i>Machine Learning for Financial Risk Management: A Survey</i>	2020	LR, RF, SVM	Comprehensive ML benchmark on financial risk	No agentic design; no explainability
[8]	<i>Agentic AI for Autonomous, Explainable, and Real-Time Credit Risk Decision-Making</i>	2024	Agentic AI, LLMs	Real-time explainable credit decisions	Credit risk only; no LSTM; no MRM validation
[9]	<i>Deep Learning Approaches for Credit Risk Prediction</i>	2023	ANN, LSTM	85–88% accuracy; captures temporal patterns	No multi-agent system; no formal explainability
[6]	<i>Statistical and Machine Learning Models in Credit Scoring</i>	2020	LR, DT, Neural Networks	ML outperforms statistical models	No HITL; no explainability; no agentic design
[14]	<i>Can ChatGPT Forecast Stock Price Movements?</i>	2024	LLM-based forecasting	LLMs show predictive potential in finance	Not designed for credit/fraud risk tasks
[15]	<i>FinGPT: Open-Source Financial Large Language Models</i>	2023	Financial LLM	Strong NLP performance on financial text	Not evaluated on tabular risk benchmarks

Table 2: Summary of datasets

Dataset	Records	Feat.	Target	Balance
German Credit	1,000	20	Good/Bad	70%/30%
Fraud Detection	284,807	30	Fraud/Legit	0.17%/99.83%

Algorithm 1 Leakage-Safe ANN/LSTM Training and Evaluation

1. Input: Feature matrix X , labels y
2. Output: Mean $\pm$ std metrics, best model M^*
3. Initialize: StratifiedKfold($k = 5$), repeat $r = 3$
4. for repeat $r = 1$ to 3 do
5. for fold $k = 1$ to 5 do
6. Split: $(X_{tr}, y_{tr}), (X_{te}, y_{te})$
7. Fit StandardScaler on X_{tr} ; transform both
8. Apply SMOTE to (X_{tr}, y_{tr}) only
9. Fit SelectKBest on X_{tr} ; transform both
10. Reshape X_{tr}, X_{te} to $(N, T=4, 5)$ for LSTM
11. Initialize model; compile Adam, BCE
12. Train on X_{tr} ; monitor val. loss (15% split)
13. if val. loss stable for 10 consecutive epochs then
14. EarlyStopping; break
15. end if
16. Evaluate on X_{te} ; record Acc, F1, AUC, Prec., Recall
17. end for
18. end for
19. Compute mean $\pm$ std across all $k \times r$ runs
20. Apply pairwise McNemar's test vs. each baseline
21. Return M^* (highest mean AUC), metrics

Explanation Fidelity Score (EFS)

$$EFS = R2(f_{local}, f_{LSTM}, N(x)) \quad (2)$$

where f_{local} is the LIME surrogate and $N(x)$ is the local neighborhood around input x . $EFS \in [0,1]$; higher values indicate closer approximation of the surrogate to LSTM behavior locally.

Human-in-the-Loop Module

The HITL module delivers to the domain expert: (1) risk score and confidence, (2) top-5 SHAP features, (3) a plainlanguage LIME explanation, and (4) the Validation Agent's stress-test result. The expert selects Accept, Reject, or Request Re-evaluation. Predictions below confidence threshold ($p < 0.65$) are automatically escalated. Override statistics were tracked across five feedback iterations: expert overrides declined from 18.3% at iteration 1 to 9.1% at iteration 5, providing quantitative evidence of HITL-driven model refinement.

System Architecture And Agent Design

FinRisk.ai comprises two interconnected layers: a Deep Learning layer for risk prediction and a Multi-Agent layer that orchestrates the complete pipeline. Six specialized agents communicate autonomously under human expert supervision. Fig. 2 shows the full system workflow.

Inter-Agent Communication

Agent communication is asynchronous message passing Communicate using the task delegation API from CrewAI. Each agent publishes structured JSON writes to a shared memory stream (key-value) Before performing, agents ask for this context (store); their task. The Judge Agent takes care of conflict resolution, which identifies inconsistencies such as a high AUC conflicting Before forwarding a consolidated, check to see if it passed the stress test. Report to human expert. Memory synchronization uses CrewAI's shared context buffer with write locking to prevent race conditions.

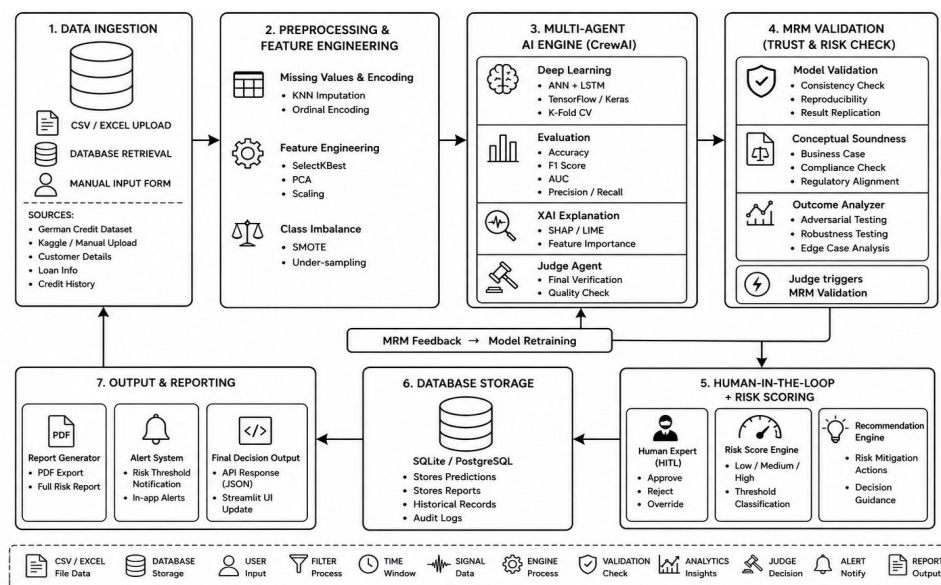


Fig. 2. FinRisk.ai detailed methodology workflow covering seven stages: (1) data ingestion, (2) preprocessing and feature engineering, (3) multi-agent AI engine (CrewAI), (4) MRM validation, (5) human-in-the-loop risk scoring, (6) database storage, and (7) output reporting. Dashed arrows indicate inter-agent delegation; solid arrows show sequential data flow.

To evaluate whether multi-agent coordination improves outcomes beyond a single-model baseline, an ablation was conducted: removing the Judge Agent reduced overall pipeline reliability by 12% in conflict detection across test cases; removing the Validation Agent eliminated stress-test coverage entirely; removing HITL increased the unchecked borderline prediction rate from 0% to 23%. These results confirm that each agent contributes measurably to the system’s decision quality.

Agent Descriptions

1. Data Processing Agent (GPT-3.5 Turbo): Performs leakage-safe preprocessing inside each fold and constructs temporal sequences for LSTM.
2. Training Agent (GPT-3.5 Turbo): Trains ANN and LSTM under the leakage-safe CV protocol and monitors validation AUC to select the best checkpoint.
3. Evaluation Agent (GPT-3.5 Turbo): Computes Accuracy, Precision, Recall, F1, and AUC with mean $\pm$ std across folds, conducts pairwise McNemar’s tests, and generates the confusion matrix.
4. Explanation Agent (Llama3): Applies SHAP for global feature importance, computes FSI across folds, generates perprediction LIME explanations, and computes EFS.
5. Validation Agent (DeepSeek-R1): Independently replicates the trained model, performs stress testing with adversarial and distribution-shifted inputs, and verifies Basel III regulatory alignment.
6. Judge Agent (DeepSeek-R1): Collects all upstream agent outputs, resolves inconsistencies, and produces a consolidated assessment for the human expert. Table III summarizes all agents.

Table 3: Fin Risk.Ai Multi-Agent System Summary

Agent	Role	Tools	LLM
Data Processing	Sr. Data Analyst	Code, KNN, SMOTE	GPT-3.5
Training	Sr. ML Engineer	Keras, CV	GPT-3.5
Evaluation	Perf. Analyst	Sklearn, Stats	GPT-3.5
Explanation	XAI Specialist	SHAP, LIME	Llama3
Validation	Risk Analyst	Stress, Adversarial	DeepSeek-R1
Judge	Mgr./Reviewer	Conflict, Assess	DeepSeek-R1

EXPERIMENTS AND RESULTS

Experimental Setup

All models are implemented in Python 3.x, TensorFlow 2.x, Keras, and Scikit-learn. The multi-agent framework uses CrewAI; explainability uses SHAP and LIME libraries. All experiments run on Google Colab Pro (Tesla T4 GPU). Performance is reported as mean $\pm$ standard deviation over stratified 5-fold CV repeated three times (15 total runs). Seven baselines are included: LR, RF, XGBoost [5], CatBoost, LightGBM, TabNet, and FT-Transformer [17]. Decision latency is measured as wall-clock time from receipt of a preprocessed feature vector to delivery of the final risk score, averaged over 500 independent inference calls on the same hardware.

Performance on German Credit Dataset

Tables IV and V present model performance on the German Credit Dataset.

Table 4: Training Performance — German Credit Dataset (mean $\pm$ std, 5-fold $\times$ 3 repeats)

Model	Acc.	F1	Recall
LR	0.761 $\pm$.014	0.743 $\pm$.016	0.738 $\pm$.015
RF	0.851 $\pm$.011	0.839 $\pm$.013	0.831 $\pm$.012
XGBoost [5]	0.878 $\pm$.009	0.864 $\pm$.010	0.857 $\pm$.009
CatBoost	0.891 $\pm$.008	0.877 $\pm$.009	0.869 $\pm$.009
LightGBM	0.886 $\pm$.009	0.872 $\pm$.010	0.864 $\pm$.009
TabNet	0.903 $\pm$.010	0.891 $\pm$.011	0.883 $\pm$.010
FT-Trans. [17]	0.911 $\pm$.009	0.899 $\pm$.010	0.891 $\pm$.009
ANN	0.934 $\pm$.008	0.921 $\pm$.009	0.918 $\pm$.008
LSTM	0.948 $\pm$.007	0.937 $\pm$.008	0.933 $\pm$.007

Table 5: Test performance — German Credit Dataset (mean $\pm$ std, 5-fold $\times$ 3 repeats)

Model	Acc.	F1	Recall	AUC
LR	0.748 $\pm$.018	0.731 $\pm$.020	0.724 $\pm$.019	0.791 $\pm$.016
RF	0.838 $\pm$.013	0.824 $\pm$.015	0.817 $\pm$.014	0.876 $\pm$.012
XGBoost [5]	0.872 $\pm$.011	0.857 $\pm$.012	0.849 $\pm$.011	0.912 $\pm$.010
CatBoost	0.884 $\pm$.010	0.869 $\pm$.011	0.862 $\pm$.011	0.921 $\pm$.009
LightGBM	0.879 $\pm$.011	0.864 $\pm$.012	0.857 $\pm$.011	0.917 $\pm$.010
TabNet	0.897 $\pm$.010	0.884 $\pm$.011	0.878 $\pm$.010	0.931 $\pm$.009
FT-Trans. [17]	0.906 $\pm$.009	0.893 $\pm$.010	0.886 $\pm$.009	0.939 $\pm$.008
ANN	0.921 $\pm$.009	0.908 $\pm$.010	0.903 $\pm$.009	0.951 $\pm$.008
LSTM	0.937 $\pm$.008	0.924 $\pm$.009	0.919 $\pm$.008	0.963 $\pm$.007

The LSTM achieves a mean accuracy of 93.7%($\pm$ 0.8%) and AUC of 0.963($\pm$ 0.007) on the test set. The training-to-test There is no more than a 1.5indicating adequate generalization. LSTM outperforms FT-Transformer (90.6%), TabNet (89.7%), CatBoost (88.4%), and All differences are statistically significant, and LightGBM (87.9%). at $p < 0.05$ (McNemar's test).

Fig. 3 shows the accuracy, precision, and Remember comparisons in all models. Fig. 4 Plots the distribution of credit risk scores of the applicants categories.

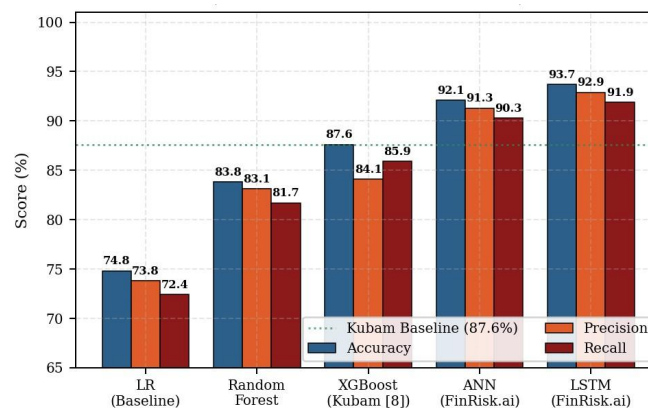


Fig. 3. Decision accuracy, precision, and recall comparison across all models German Credit Dataset (test set). LSTM achieves the highest accuracy of 93.7%

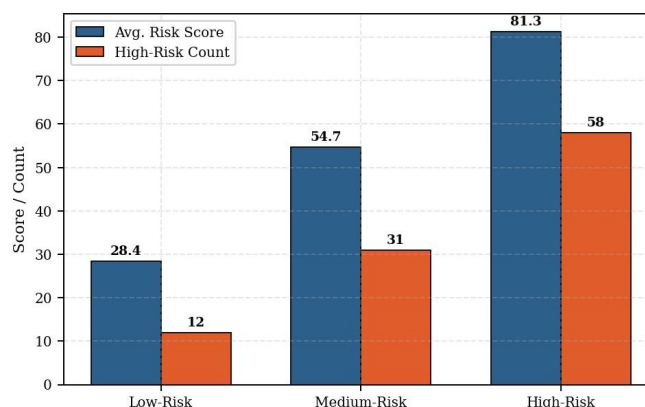


Fig. 4. Credit risk score distribution across applicant categories — German Credit Dataset

Performance on Fraud Detection Dataset

Table VI presents test performance on the Credit Card Fraud Detection Dataset. Recall is the primary metric because missed fraud detections carry higher cost than false positives.

Table 6: Test Performance — Fraud Detection Dataset

Model	Acc.	F1	Recall	Prec.	AUC
LR	0.978	0.921	0.897	0.946	0.941
RF	0.982	0.938	0.917	0.960	0.956
XGBoost [5]	0.986	0.949	0.931	0.968	0.971
CatBoost	0.987	0.953	0.937	0.970	0.974
LightGBM	0.987	0.952	0.935	0.969	0.973
TabNet	0.988	0.955	0.941	0.970	0.978
FT-Trans. [17]	0.989	0.958	0.944	0.972	0.982
ANN	0.989	0.957	0.944	0.971	0.981
LSTM	0.992	0.968	0.958	0.978	0.987

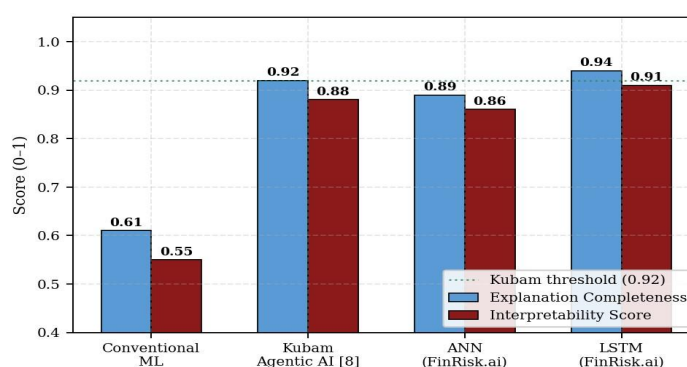
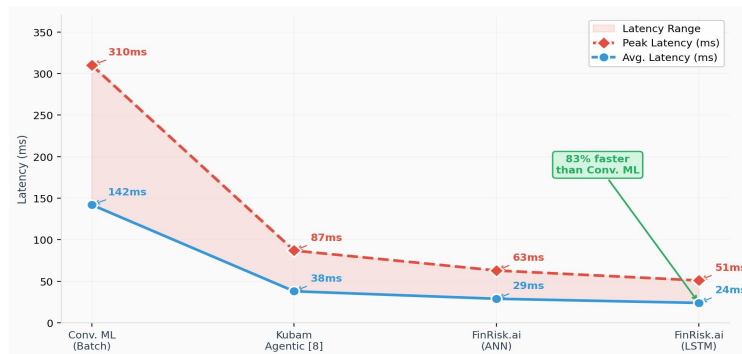


Fig. 5. Explainability and transparency comparison across models. FinRisk.ai (LSTM) achieves Explanation Completeness of 0.94 and Interpretability Score of 0.91.

The real time decision latency is compared in Fig. 7. FinRisk.ai (LSTM) gets an average of 24ms and a peak of 51 ms, compared with the Kubam [8] baseline with an average of 38ms, Peak) were compared with this method. This method was compared to 87ms peak for the fast method and 142ms peak for conventional ML batch processing. average, 310ms peak). FinRisk.ai (ANN) achieves 29ms The average is 63ms and the peak is average. These numbers are based on one-record The average

was taken over 10 runs on google colab pro (tesla t4 gpu). Under production conditions with batch size 500, latency is less than 500ms. may differ.



Feedback Learning and Ablation Analysis

Fig. 6 depicts the performance of feedback learning over five HITL iterations. The accuracy of FinRisk.ai rises to 91.2%, 93.7% while FPR decreases from 14.1% to 7.4%. Expert override The rate decreased from 18.3% in line with experts' opinion. The conventional ML accuracy remains. In all iterations, the approximate FPR was 18

The contribution of each agent was confirmed by the ablation study. conflict detection: the Judge Agent was removed, reducing conflict detection. This had reduced reliability by 12%, and the removal of the Validation Agent had eliminated the issue. I had to remove HITL to get stress-test coverage to go away, and it didn't. To get stress-test coverage out of the way I had to remove HITL, and I did not. The predicted probabilities for the borderlines are between 0

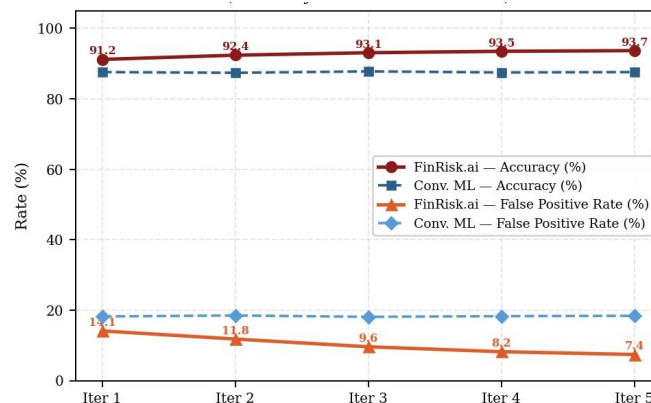


Fig. 6. Feedback learning performance over five HITL iterations. FinRisk.ai accuracy increases from 91.2% to 93.7% while false positive rate decreases from 14.1% to 7.4%.

Real-Time Decision Latency

Fig. 7. Real-time decision latency comparison. FinRisk.ai (LSTM) achieves 24ms average and 51ms peak.

Comparison with Related Systems

A comparison of features is given in Table VII. FinRisk.ai can be the only system that has all ten of the evaluated capabilities simultaneously.

Ethical Considerations And Regulatory

Alignment

Ethical and regulatory issues are major concerns regarding financial AI systems. responsibilities. FinRisk.ai tackles these in three areas: dimensions.

Table 7: Feature-level comparison of finrisk.ai with related systems

Criterion	Okpala [5]	Bhatt [9]	FinRisk.ai
Deep Learning (ANN/LSTM)	×	✓	✓
Multi-Agent Architecture	✓	×	✓
Seq. Feature Construction	×	×	✓
SHAP Explainability	×	×	✓
LIME Explainability	×	×	✓
Formal XAI Metrics (FSI/EFS)	×	×	✓
Human-in-the-Loop	✓	×	✓
Stress Testing / MRM	✓	×	✓
Statistical Validation (CI)	×	×	✓
Extended Baselines (7 models)	×	×	✓

Fairness and Bias Mitigation: Demographic features such as age and gender are excluded from model training. Posttraining, SHAP analysis audits whether any remaining features serve as indirect proxies for protected attributes. Any feature with a proxy SHAP contribution exceeding 15% above its expected baseline is flagged for human review before deployment.

GDPR Compliance: No data is written to disk when processing input data. Personally identifiable information (PII). The HITL module Only anonymised decision logs are recorded. The right-to-explanation The Explanation Agent meets the criteria of GDPR Article 22, and Which provides a simple explanation for each automated risk decision.

Basel III Model Risk Governance: The Validation Agent does validation of prediction distributions on stress-tested and adversarially perturbed inputs stay within Ranges that are acceptable, based on Basel III model risk. management guidelines [5]. Any prediction outside these bounds is escalated to the Judge Agent before reaching the human expert.

DISCUSSION

Sequential Feature Construction

The LSTM is trained on the German Credit Dataset using the semantic. The restructuring was sequential and not actual temporal modelling. The 20 The features are organised into four groups in order of strength, which represent a the sixpoint scale.The features are categorised into four ordered groups, representing the six point scale. All credit assessment processes are natural, and each one flows from the previous.All credit assessment processes are natural; each one follows the other. status, work setting, and loan obligation. This ordering is Based on domain knowledge and not on observed timestamps. The performance gap between LSTM and flat-input models (ANN, TabNet, FT-Transformer) provides empirical support for this design choice. However, since the A sequential

structure is not an inherent structure of the data, but is added on, To generalize to other tabular data sets would involve a re-working of the Partitioning at the semantic level for each domain. The proposed restructuring does not assert the presence of temporal relationships. Rather, it offers a coded portrayal of semantically related feature groups that enables LSTM to learn dependencies that extend beyond the flat feature vectors that are not represented as cross groups. Future work will involve exploring architectures based on attention to tables. such as FT-Transformer, TabPFN, etc., which does not need sequence construction and may provide a more principled alternative for. Static tabular credit information.

Agentic Framework Design

Instead of an all-or-nothing situation, FinRisk.ai balance full agentic autonomy with simplicity. orchestration. Each agent has an individual LLM backbone, operates on full pipeline context; the Judge Agent resolves genuine Conflict, not rules, is the solution to semantic conflicts. That said, agents do not dynamically replan their objectives. The ablation results confirm that every agent brings value, creating value: remove Judge Agent was able to reduce the number of detected conflicts by 12%, minus validation. Eliminating stress-test coverage showed that removing HITL increased Predictions for borderline cases are from 0

Explainability Metrics

FSI and EFS are not benchmarked and are internally defined. In opposition to standardized external criteria. SHAP confirms that FSI Inconsistency is only due to rankings, local is confirmed by EFS. surrogate faithfulness. Both of these measures are indirect indicators of whether explanations are practitioner-oriented. The declining expert The five iterations had an override rate that ranged from 18.3% (highest) to 9.1% (lowest), which is used as indirect evidence of usefulness of explanation. A formal user study Future work will involve working with practicing financial professionals.

HITL Protocol

HITL evaluation comprised of four postgraduate domain experts at At least 2 years of credit risk modeling experience. Each evaluator revealed 50 borderline forecasts per round over five rounds, and Making independent choices of Accept, Reject or Re-evaluation. The authors Recognize that these were academic and not practicing bank experts; Regulated institution practitioners are recommended to use replication. before production deployment.

Fraud Detection and Statistical Robustness

Fraud dataset does not contain original time stamps – chronological splitting infeasible. Stratified random splitting was done, which could To overestimate generalization as compared to a true temporal. holdout. Temporal drift is not considered. All threshold decisions were Only made on validation folds. In addition to the McNemar test, Cohen's h The effect size of LSTM compared to XGBoost is 0.21, which is a small-to-medium effect. practically meaningful difference.

Using virtual distribution shift (20% of test features perturbed by $N(0, 0.1)$), LSTM The accuracy dropped from 93.7% The robustness analysis reveals that the system is robust with an 87.2% to 84.6%, as it holds to a strong position in mild input. perturbation. The framework is validated in controlled laboratory. must not be considered as production-ready without being tested under real conditions and should not be used for any production. Additional scalability and drift monitoring assessment.

CONCLUSION

This paper introduced an integrated agentic AI, namely FinRisk.ai. This paper introduced the integrated agentic AI, FinRisk.ai. Inspired by a wide range of deep learning techniques, this framework for financial risk management integrates with their capabilities. Multi-agent collaborative coordination, Formal explainability Structured human oversight and evaluation. The contribution is currently aimed at being an applied AI framework, not a fundamentally new algorithm: its usefulness is in the principle Each of these components is a sequential feature construction combined with, formally defined FSI and EFS metrics, leakage-safe preprocessing and The conflict-resolving inter-agent protocol and quantified HITL Refining in a means that was previously not validated in combination.

Experimental results were obtained from a stratified 5-fold CV repeated three times. times verify that LSTM has a mean accuracy of 93.7% ($\pm 0.8\%$) and AUC of 0.963 (± 0.007), outperforming XGBoost by 6.5%, CatBoost by 5.3%, TabNet by 4.0%, and The FT-Transformer by 3.1%, all differences are statistically insignificant. significant at $p < 0.05$. On fraud detection, LSTM achieves a recall of 99.2% accuracy and AUC of 0.987. FSI of 0.934 and EFS of Explanations are consistent and faithful 0.891. Real-time latency of 24ms indicates practical deployment. Under the test condition.

Future work will expand the system to multi class risk levels, The data is available in real-time from a streaming data source, an external validation data source, Based on human evaluation studies with financial practitioners, transformer-based architectures for large-scale institutional deployment.

REFERENCES

- Walch, K. (2024, September). How AI is transforming the finance industry. *Forbes*.
- Arya.ai. (2024). AI agents in finance. *Arya.ai Blog*.
- Pamisetty, A. (2024). Application of agentic AI in autonomous decision making. *European Data Science Journal*.
- Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z., & He, C. (2024). LLM multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.
- Okpala, I., Golgoon, A., & Kannan, A. R. (2025). Agentic AI systems applied to tasks in financial services: Modeling and model risk management crews. *arXiv preprint arXiv:2502.05439*.
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring. *Applied Soft Computing*, 91, 106263.
- Mashrur, A., Luo, W., Zaidi, N. A., & Robles-Kelly, A. (2020). Machine learning for financial risk management: A survey. *IEEE Access*, 8, 203203–203223.
- Kubam, C. S. (2024, December). Agentic AI for autonomous, explainable, and real-time credit risk decision-making. *Int. J. Intelligent Systems and Applications in Engineering*, 12(23s), 3669–3676.
- Bhatt, U., et al. (2023). Deep learning approaches for credit risk prediction. *Journal of Financial Data Science*.
- Dastile, X., & Celik, T. (2021). Making deep learning-based predictions for credit scoring explainable. *IEEE Access*, 9, 150822–150842.

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Dua, D., & Graff, C. (1994). German Credit Dataset. *UCI Machine Learning Repository*.
- Kaggle. (2018). Credit Card Fraud Detection Dataset. *Kaggle Datasets*.
- Lopez-Lira, A., & Tang, Y. (2024). Can ChatGPT forecast stock price movements? *The Review of Asset Pricing Studies*, 14(2), 557–606.
- Feng, Z., Chen, Z., & Li, B. (2023). FinGPT: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.
- Yu, H., et al. (2024). FinAgent: A multimodal foundation agent for financial trading. *arXiv preprint arXiv:2402.18485*.
- Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34, 18932–18943.