

An AI-Driven Data Engineering Framework for Financial Risk Intelligence, Data Quality, Governance, and Predictive Decision-Making in Banking

Sivakami S¹, T. Dinesh², Saravanan³

¹*Dept. of Electronics and Communication Engineering, New Prince Shri Bhavani College of*

Engineering and Technology Chennai, Tamil Nadu, India.

²*Senior Developer, Techpower Solutions, Nagercoil, Tamilnadu, Tamil Nadu.*

³*Software Engineer, Cloudstrand, Edinburgh, Scotland, United Kingdom.*

¹ sivakami@newprinceshribhavani.com

Abstract—The rapid growth of digital banking generates large volumes of heterogeneous data, creating challenges in financial risk assessment, data quality management, governance, and timely decision-making. This paper proposes an AI-based data engineering approach to integrate banking data and enable intelligent financial risk management. It seamlessly integrates automated data ingestion, data cleansing, quality checks, feature engineering, governance controls, and artificial intelligence-based predictive analysis. Data quality mechanisms detect missing, inconsistent, duplicate, and anomalous records prior to analytical processing, and governance components facilitate secure access to data, traceability of data usage, and controlled access. Machines are used to process the data, detect financial risk patterns, and provide predictive insights for bank decisions using machine learning techniques. A continuous monitoring mechanism assesses the data quality, model performance, and risk changes over time. Experimental results demonstrate 98.67% classification accuracy and a 98.65% weighted F1-score. The framework offers a comprehensive method to integrate data from various sources into a coherent and trustworthy risk intelligence, empowering financial institutions to boost data accuracy, improve risk management governance, maximize risk forecasting power, and facilitate timely and informed risk-rating decisions.

Keywords—Artificial Intelligence, Financial Risk Intelligence, Data Quality, Data Governance, Predictive Decision-Making.

I. INTRODUCTION

The banking industry is rapidly digitalizing, and the bank transaction, customer interaction, mobile banking, online services, credit applications, and financial operations generate huge amounts of structured and unstructured data [1]. These sources are diverse and provide significant opportunities to evaluate financial risk and to enhance decision-making processes but also pose opportunities for data engineering challenges. It can also be less reliable because of the many formats, missing data, duplicate records, inaccuracies, and late data processing of the results of analyses [2]. But, before embarking on advanced analytics, financial institutions require efficient systems for data collection, integration, validation, and management. AI has tools and functions that can identify complex patterns, detect anomalies, and help with automated analysis of financial data [3]. Banking institutions can use AI combined with data engineering workflows to transform a variety of data into meaningful and accurate threat data. One of the reasons why data quality is important is that the wrong or incomplete data can affect the accuracy of the results for the risk assessment and the risks predicted [4]. Therefore, the proper solution that includes the process of data ingestion, data preprocessing, data quality control, data governance, and AI-based financial data analysis can provide a structured foundation to create intelligent financial risk intelligence and data-informed banking practices [5].

There are various financial risks that banking institutions encounter, such as credit risk, transaction risk, fraud risk, and operational risk. Conventional risk assessment approaches are based on a limited number of known risk rules, on-prepared data sets, and historical data and do not necessarily reflect the financial behavior of the population. Using AI-powered predictive models, relationships and patterns of potential risks in large and complex datasets are uncovered [6]. However, these models are only as useful as the quality, consistency, and standards of the data upon which they are based. Data quality problems can lead to bias, diminish predictive power, and lead to wrong decisions. Hence, data governance plays an important role in managing access, monitoring data movement, securing data, and ensuring that data is responsible and subject to regulatory requirements [7]. It is possible to establish rules about data ownership, data validation, use, and monitoring throughout the data lifecycle with appropriate governance. The combination of governance and data engineering also gives financial institutions the assurance of data reliability and enables secure analytical activities [8]. Furthermore, continued monitoring can identify shifts in the nature of the data and the models' performance and allow the risk management process to respond to the changing conditions. These abilities provide a chance to combine data quality and governance with financial risk prediction into a single AI-driven model [9].

The paper introduces an artificial intelligence (AI)-based data engineering that can be applied in the banking sector for financial risk intelligence, data quality management, data governance, and predictive decision-making. The framework covers a full end-to-end pipeline, from data acquisition and ingestion from various banking data sources to the preprocessing stage, data quality validation, feature engineering, governance compliance, risk analysis with AI, and then, last but not least, predictive decision support. The automated quality assessment process facilitates the identification of missingness, duplication, inconsistencies, and anomalies in records before they are given to predictive models. Governance mechanisms are built in to ensure that secure access to data, data usage is controllable and monitored, and that data is traceable throughout its life cycle. The machine learning-based analytical parts, however, detect senseful patterns from the validated banking information and give risk predictions for decision support. The framework also includes continuous monitoring to evaluate changes in data quality, risk patterns, and predictive performance. The proposed approach is based on data engineering and artificial intelligence, aiming at better data reliability of banks and strengthening risk intelligence [10]. The framework offers a systematic basis to transform the diverse financial information into insights that can guide timely, consistent, and informed banking decisions.

II. LITERATURE SURVEY

Ali et al. discussed the machine learning methods for financial fraud detection, emphasizing the capacity of machine learning to recognize intricate fraud patterns [11]. Shi et al. studied the machine learning methods for credit risk assessment and explored the possibility of enhancing the automated financial risk prediction [12]. Alonso Robisco and Carbo Martinez discussed the risk-adjusted performance of machine-learning models for the credit default prediction, highlighting the need for a sound model evaluation in the financial field [13]. Feng studied and proved the application of the neural network in the assessment of green credit risk in banking risk management [14]. Chhabra Roy and Prabhakaran have described a system that detects, prioritizes, and prevents cyber frauds that happen internally inside the Indian banks using machine learning methods [15]. Although these studies demonstrate the effectiveness of AI for specific financial risk problems, they mainly focus on individual applications. There is not much focus on integrating data quality, governance, heterogeneous data engineering, predictive risk intelligence, and continuous monitoring. Hence, this study proposes an integrated framework that considers these aspects for reliable banking decision-making.

III. PROPOSED WORKFLOW

The proposed workflow combines data engineering, artificial intelligence, data quality, governance, and predictive analytics into a single banking risk management workflow. First, data is sourced from various transaction systems, customer profiles, credit data, loan databases, and digital banking systems. The data collected are ingested and integrated into a centralized processing environment. Data is then preprocessed to take care of missing data, duplicate data, inconsistent data formats, and outliers. Then, completeness, accuracy, consistency, uniqueness, and timeliness are assessed in data quality assessment. Governance mechanisms are implemented to regulate access to data, ensure traceability, and facilitate secure use of the data. Financial elements are then mined using feature engineering. AI risk intelligence can use these features to detect fraud, credit risk, transaction risk, and other types of financial risk. Decision support is provided by predictive models, which classify risks and provide future risk indicators. Finally, continuous monitoring assesses data quality, model performance, and evolving risk patterns, allowing for timely modifications and enhanced decision-making accuracy in banking.

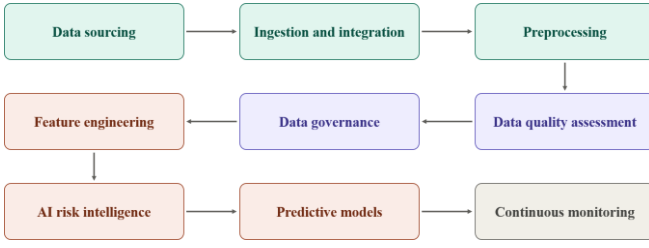


Fig. 1. Proposed Workflow

IV. METHODOLOGY

A. Data Acquisition and Integration

The proposed methodology starts by capturing the distributed banking information from customer profiles, transactions, credit history, loan databases, account activities, and digital banking platforms. The data gathered can be of various forms, structures, and attributes. So, in order to move and combine these into a single analysis environment, an automated ingestion process is adopted. Data transformation is used to normalize attribute formats and ensure consistency of representations. For traceability reasons, metadata like source, transaction category, or transaction time is stored. The integrated dataset is shown as

$$D = \{d_1, d_2, \dots, d_n\} \quad (1)$$

where D represents the integrated banking dataset.

B. Data Preprocessing and Quality Assessment

The integrated data set is then evaluated for missing or duplicated data, inconsistent formats, and unusual data. Numerical values not contained are replaced by appropriate statistical methods, and categorical attributes are standardized. Duplicate records are deleted to avoid biased analysis, and numerical features are scaled prior to model creation. Completeness, accuracy, consistency, uniqueness, and timeliness are criteria used to assess data quality. The combined quality score is computed as

$$DQ = \frac{C+A+K+U+T}{5} \quad (2)$$

where C , A , K , U , and T denote the normalized quality scores for completeness, accuracy, consistency, uniqueness, and timeliness, respectively.

C. Data Governance and Security

The governance layer contains security and management rules that ensure that banking information is accessed and processed on a predefined basis. Role-based access control is used to limit access to sensitive data based on the user's role. Data lineage mechanisms track the flow and evolution of data in the processing flow. Data access and processing are tracked through audit logs. Governance policies act during

data ingestion and preprocessing as well as during analytical processing to enhance accountability and traceability. The score for governance can be expressed as

$$GC = \frac{P_c}{P_t} \times 100 \quad (3)$$

where P_c represents compliant governance policies and P_t represents total applicable policies.

D. Feature Engineering

Financial characteristics are then extracted from this prepared banking data, following data quality validation and data governance processing. Things like transactions, frequency, transaction amount, repayment behavior, credit utilization, account activity, and duration of the loan are created. Categorical attributes are represented in a suitable numerical form, and numerical features are normalized. If there are many redundant and highly correlated variables, they can be eliminated to improve model simplicity. The resulting feature matrix is given as

$$X = [x_1, x_2, \dots, x_m] \quad (4)$$

where X represents the feature matrix and x_i denotes an individual engineered financial feature. These features will then be fed to the AI-based risk prediction module.

E. AI-Based Financial Risk Intelligence

The AI-based risk intelligence module analyzes the engineered banking features to pick out financial risk patterns. The prepared data set is split into a training set and a testing set, and the predictive model learns relationships between financial attributes and observed outcomes of risk. Machine learning algorithms are used to categorize the transactions or customers based on the predetermined risk classes. The prediction process is shown below:

$$\hat{Y} = f(X; \theta) \quad (5)$$

where X represents the input feature vector, θ denotes the learned model parameters, and $\hat{Y}$ represents the predicted risk outcome. The predictions generated are used to provide the classification of risk level or probability values that are used in the next banking decision.

F. Predictive Decision-Making

The AI module's predictions are transformed into actionable financial risk information for decision support. Customer transactions or actions can be segmented based on the risk level that they are likely to pose. There is potential to use high-risk observations for further verification and relevant customer patterns for credit or financial assessment. The important financial characteristics that are used to calculate a weighted risk score are

$$R = \sum_{i=1}^m w_i f_i \quad (6)$$

where R represents the overall risk score, w_i denotes the weight assigned to feature f_i , and m represents the total number of selected features. Risk information that is produced as a result is helping to make banking decisions in a timely and data-driven fashion.

G. Continuous Monitoring and Model Evaluation

The final stage is the continual monitoring of data quality, governance compliance, evolving risk patterns, and predictive model performance. Model performance is measured in terms of accuracy, precision, recall, F1-score, and ROC-AUC. Indicators of data quality are recalculated regularly to detect any quality problems in the information being processed. Financial behavior changes are also tracked for possible model or data drift. The F1 score is calculated as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

Precision indicates the proportion of positive cases identified correctly; recall indicates the proportion of all positive cases that are identified correctly. Results from the monitoring are used to make timely updates to the model and data processing.

H. Dataset Description

The proposed framework is based on the Home Credit Default Risk dataset, which is the most popular banking data set to predict payment difficulties for customers. It includes user application data and related credit history and transaction data. The dataset has relational modules like Application, Bureau, Previous Application, Bureau Balance, Credit Card, and Installments. There are over 120 attributes in the application table and historical credit and repayment data in the other tables. The data has categorical and numerical attributes, where the schema consists of about 86.9% numerical and 13.1% categorical variables. The target variable is the repayment/delinquency status results. These heterogeneous relational data are integrated, cleaned, quality validated, and converted to features for machine learning-based financial risk prediction.

Algorithm 1: AI-Driven Financial Risk Prediction

Input: Banking dataset DD
Output: Predicted risk results RR
Step 1: Collect banking data such as customer, transaction, credit, and loan records.
Step 2: Combine the collected data into a single dataset.
Step 3: Clean the data by handling missing values, duplicate records, and inconsistent data.

Step 4: Check the quality of the cleaned data using completeness, accuracy, and consistency.

Step 5: Apply data governance rules for secure data access and management.

Step 6: Select important financial features from the processed data.

Step 7: Train an AI model using the selected features.

Step 8: Use the trained model to predict financial risk.

Step 9: Classify the data into suitable risk categories.

Step 10: Evaluate the prediction using accuracy, precision, recall, and F1-score.

Step 11: Monitor the data and model performance continuously.

Step 12: Generate the final risk prediction and decision-support result.

End Algorithm

V. RESULTS AND DISCUSSION

Fig. 2 shows the Default Risk Incidence Across Housing Categories, comparing risk rates for different housing conditions. The proportion of default is highest for rented apartments (around 14.4%), which suggests a relatively higher level of payment problems in rented apartments. The approximate risk rate for house or apartment customers is approximately 8.0%, and for municipal apartment customers it is approximately 7.7%. Customers who reside with parents are at the lowest risk, approximately 6.4%. The variation shows that housing-related characteristics can be useful for providing contextual information and can help to enhance the financial risk prediction model's ability to identify potentially vulnerable groups of customers.

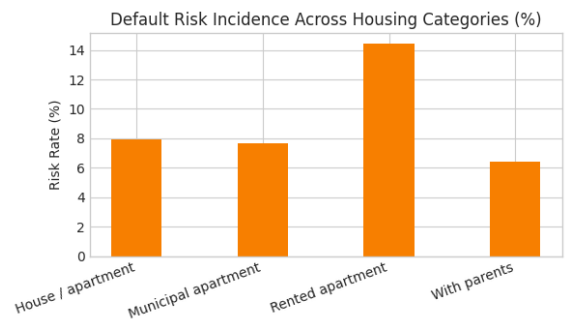


Fig. 2. Default Risk Incidence Across Housing Categories

Fig. 3 depicts the Home Credit Relational Data Schema Dimensionality by major relational modules. High dimensionality is exemplified in the Application module, which has the most attributes with over 120 columns. The Previous Application module has about 37 attributes,

followed by Credit Card with about 23. Bureau has approximately 17 attributes, and Installments has approximately 8 attributes. This distribution reflects that the structure of banking data is heterogeneous and the dimensionality of the banking data varies widely. These modules can be integrated to deliver more comprehensive financial data, and feature engineering plus feature selection can be used to eliminate data redundancy and generate an appropriate machine-learning-friendly representation.

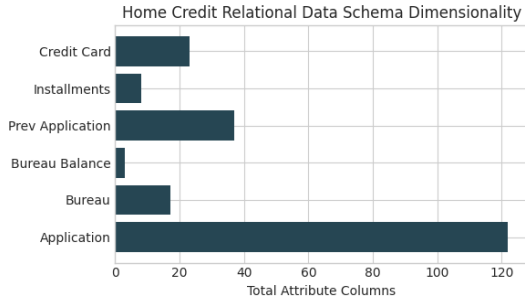


Fig. 3. Home Credit Relational Data Schema Dimensionality

Fig. 4 is a log plot of the Enterprise Record Ingestion Volume for four relational modules. The module with the most ingested rows is the Bureau module with ~16,000 rows, followed by Installments with ~11,000 rows. Previous applications add around 10,000 records; applications add about 4,000 records. The results show that various banking modules provide significant amounts of data to the data engineering pipeline. As a result, flexible ingestion and integration mechanisms are needed to efficiently process heterogeneous records and continue to maintain the relational information for further financial risk analysis and subsequent predictive modeling tasks.

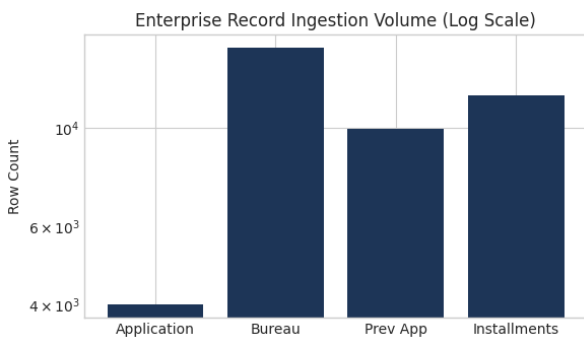


Fig. 4. Enterprise Record Ingestion Volume

The banking data processed is depicted in Figure 5 in terms of schema data type composition. The majority of the schema (86.9%) are numerical variables, such as integer or floating-point attributes, while there are 13.1% categorical object variables. The majority of the attributes are numerical, suggesting that there is significant financial and transactional information that can be used directly after proper scaling and

preprocessing. But categorical attributes need to be encoded for machine learning. This distribution illustrates the heterogeneous nature of banking data and highlights the need for an appropriate preprocessing pipeline to ensure that both numerical and categorical data are dealt with in a consistent manner before feature engineering and financial risk prediction.

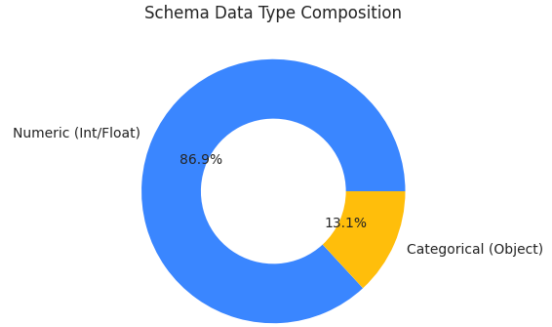


Fig. 5. Schema Data Type Composition

The overall accuracy of the Financial Risk Confusion Matrix is 98.7%, as shown in Fig. 6. The model accurately classifies 539 of the repayment or low-risk cases and 53 of the default-risk cases. There are 2 cases of repayment that are wrongly classified as default risk and 6 cases of default risk that are predicted to be repayment. The results show a good classification and a rather low number of misclassifications. The model correctly identifies both majority and minority risk classes, but the six false-negative cases are significant for financial risk management, as the default-risk customers that are not identified may need more monitoring or intervention in decision-making.

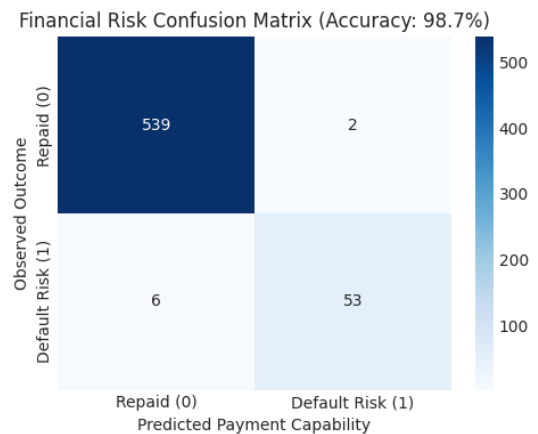


Fig. 6. Financial Risk Confusion Matrix

In Fig. 7, the Table-Level Comprehensive Quality Index (FDQI) of four relational modules is depicted. Previous applications come in at the highest quality index (0.991), Application (0.982), and Bureau Balance (0.978). The lowest index is found at Bureau with an approximate

value of 0.965 but is still a fairly high quality. The results show that the data quality processing proposed in this paper has achieved good consistency and reliability of relational data across the modules. The differences in the tables also provide evidence that one has to test the quality of each table separately and not take data quality for granted across the entire banking data set.

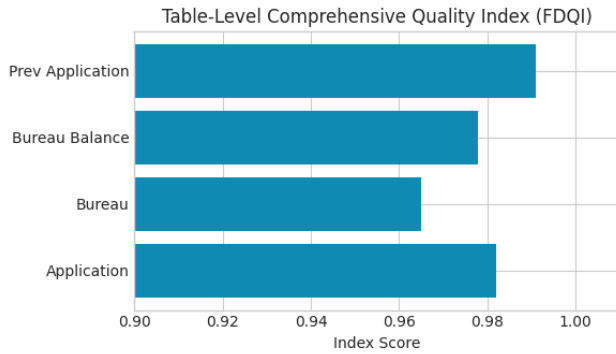


Fig. 7. Table-Level Comprehensive Quality Index

Fig. 8 shows the Predictive Performance Gain Across Relational Modules. The accuracy obtained from the raw application data is around 84.5%. By including Bureau History, accuracy is boosted to around 89.2%, and with the inclusion of Previous Applications, it is approximately 93.8%. The incorporation of financial ratios achieves a higher accuracy of almost 98.8%. The continual increase shows the significance of combining several relational banking sources and creating financial attributes. These findings suggest that historical, application, and financial ratio data, when used together, represent better indicators of payment difficulty and are better for developing financial risk intelligence, respectively.

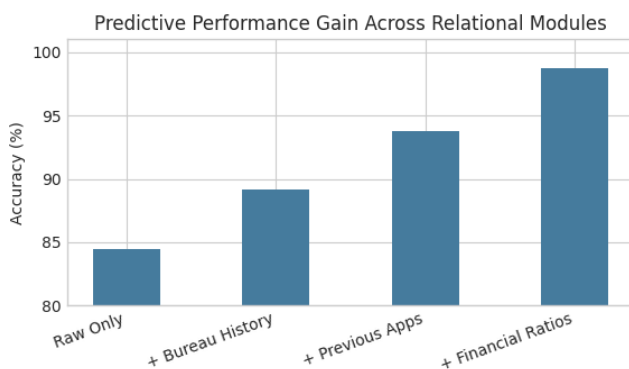


Fig. 8. Predictive Performance Gain Across Relational Modules

The Feature Engineering and Selection Funnel (Fig. 9) shows how the number of attributes in the original banking dataset is reduced over successive steps. The representation of the Raw Application Attributes is greater than 120 features, whereas Cross-Table Aggregation reduces the representation to around 45 features. Governance validation

further reduces the feature set to approximately 24 features, and the final ML-Ready features have approximately 16 attributes. This decrease suggests that feature engineering and the selection of features based on governance can exclude irrelevant and inappropriate features without losing valuable information. This condensed functionality can simplify the complexity of a model and facilitate the efficient financial risk analysis of predicting risks.

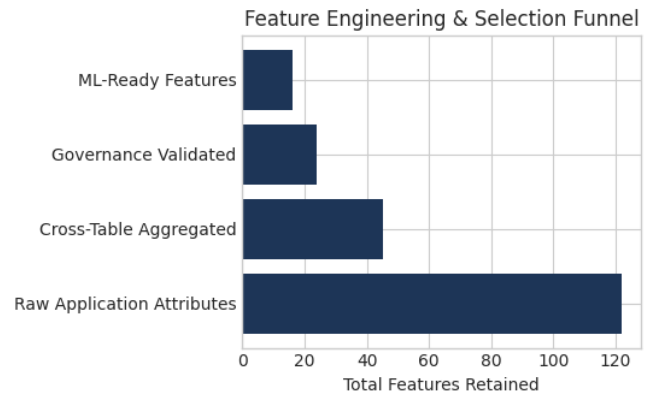


Fig. 9. Feature Engineering and Selection Funnel

The training and validation accuracy over 25 epochs is displayed in Fig. 10. The accuracy of the training set rises from around 91.6% at the start of training to around 98.9% towards the end of training, and the accuracy of the validation set rises from around 92.6% at the start of training to around 98.7% towards the end of training. The distance between the two curves is not too great during training, suggesting consistent learning without excessive differences. The target benchmark is 98.7% and is achieved with the validation performance towards the end of epochs. The output shows that the model is able to capture meaningful financial patterns progressively as it is training while retaining a high level of validation performance.

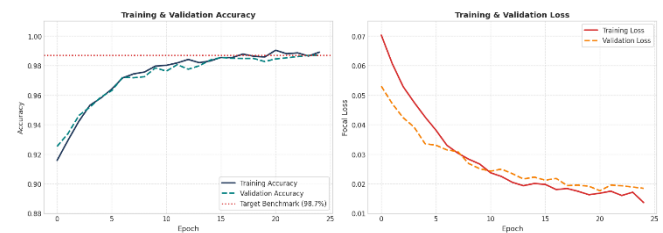


Fig. 10. Training and Validation Accuracy

The training and validation loss is plotted in each epoch in Fig. 11. The training loss drops from around 0.070 initially to 0.014 at the end of all the epochs. The validation loss also drops from ~0.053 to ~0.018, with small deviations in the intermediate epochs. The general downward trend suggests that the model is being optimized well and prediction errors are decreasing over time. The gap between the training and validation loss does not widen significantly with increasing

epochs, indicating that the model is learning effectively without overfitting. These results are in line with the accuracy curves to show that it models the learning process correctly.

Classification Report:

	precision	recall	f1-score	support
Repayment / Low Risk (0)	0.9890	0.9963	0.9926	541
Payment Difficulty / Risk (1)	0.9636	0.8983	0.9298	59
accuracy			0.9867	600
macro avg	0.9763	0.9473	0.9612	600
weighted avg	0.9865	0.9867	0.9865	600

Fig. 11. Training and Validation Loss

The relationships between financial variables are represented in the Cross-Table Financial Intelligence Correlation Matrix, as shown in Fig. 12. There is a strong positive correlation between AMT_CREDIT and AMT_ANNUITY with a correlation coefficient of 0.94. AMT_INCOME_TOTAL is strongly correlated with AMT_CREDIT (0.88) and AMT_ANNUITY (0.82). BUREAU_CREDIT_COUNT and BUREAU_DEBT_SUM_TOTAL have a moderate correlation of 0.62. The TARGET variable is correlated negatively with EXT_SOURCE_2 (-0.34) and EXT_SOURCE_3 (-0.39), which are important for default-risk prediction. There are also moderate relationships between CREDIT_INCOME_RATIO and credit and annuity variables. These correlations are used in feature selection and financial risk modelling.

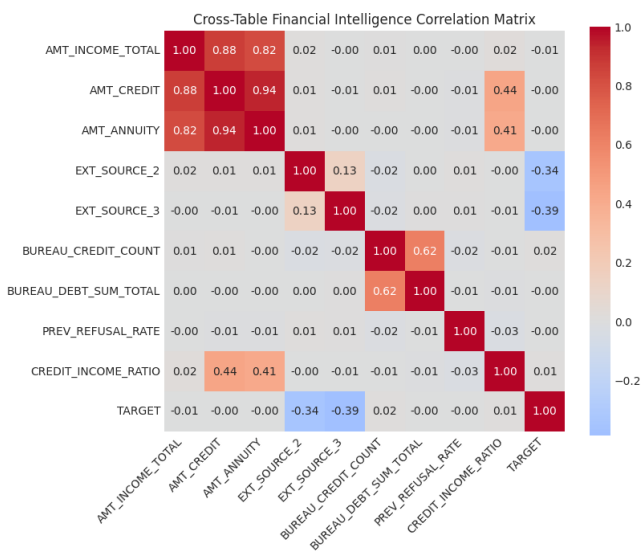


Fig. 12. Cross-Table Financial Intelligence Correlation Matrix

VI. CONCLUSION

This paper introduced an artificial intelligence-based framework of data engineering for financial risk intelligence, data quality, data governance, and predictive decision-

making in banking. The framework ingests, preprocesses, assesses quality, governs, transforms, and predicts risks with AI for heterogeneous banking data. The experimental results show that the use of relational financial information can significantly enhance the prediction accuracy, from about 84.5% with the raw application data to 98.8% with financial ratios and cross-table information. The classification model obtained the highest accuracy, 98.67%, and the highest weighted F1 score, 98.65%. The data-quality analysis also yielded high FDQI values for all the relational tables being evaluated, and feature selection brought the initial high-dimensional representation of the data to a compact feature set for the ML phase. Overall, the findings show how data engineering and AI-driven financial risk analysis can help to identify risks accurately, enhance data reliability, and make timely decisions in banking environments.

REFERENCE

- [1] Bhatore, Siddharth, Lalit Mohan, and Y. Raghu Reddy. "Machine learning techniques for credit risk evaluation: a systematic literature review." *Journal of Banking and Financial Technology* 4, no. 1 (2020): 111-138.
- [2] Wang, Yuelin, Yihan Zhang, Yan Lu, and Xinran Yu. "A Comparative Assessment of Credit Risk Model Based on Machine Learning—a case study of bank loan data." *Procedia Computer Science* 174 (2020): 141-149.
- [3] Boughaci, Dalila, and Abdullah AK Alkhawaldeh. "Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study." *Risk and Decision Analysis* 8, no. 1-2 (2020): 15-24.
- [4] Al Wahshi, Jihad JRA, Jonathan Foster, and Pamela Abbott. "An investigation into the role of data governance in improving data quality: A case study of the Omani banking sector." In *ECIS 2022 Research Papers*. AIS Electronic Library (AISeL), 2022.
- [5] Yu, Lean, Xiaowen Huang, and Hang Yin. "Can machine learning paradigm improve attribute noise problem in credit risk classification?." *International Review of Economics & Finance* 70 (2020): 440-455.
- [6] T. Dinesh, "AI-Enhanced Embedded Agentic Retrieval for Efficient Single-Call Generation with Multi-Call Agentic RAG Answer Quality IEEE," *Journal of Progressive Engineering Studies (JPES)*, vol. 1, no. 2, pp. 79-86, 2026. DOI: 10.65379/JPES/V1(2)P72
- [7] Aniceto, Maisa Cardoso, Flavio Barboza, and Herbert Kimura. "Machine learning predictivity applied to consumer creditworthiness." *Future Business Journal* 6, no. 1 (2020): 37.
- [8] Duarte, Denize Lemos, and Flávio Luiz de Moraes Barboza. "Forecasting financial distress with machine learning—a review." *Future Studies Research Journal: Trends and Strategies* 12, no. 3 (2020): 528-574.
- [9] Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. "Explainable Machine Learning in Credit Risk Management: N. Bussmann et al." *Computational Economics* 57, no. 1 (2021): 203-216.
- [10] Alessi, Lucia, and Roberto Savona. "Machine learning for financial stability." In *Data science for economics and finance: Methodologies and applications*, pp. 65-87. Cham: Springer International Publishing, 2021.
- [11] Zambuto, Fabio, Simona Arcuti, Roberto Sabatini, and Daniele Zambuto. "Application of classification algorithms for the assessment of confirmation to quality remarks." *Bank of Italy Occasional Paper* 631 (2021).

- [12] Ali, Abdulalem, Shukor Abd Razak, Siti Hajar Othman, Taiseer Abdalla Elfadil Eisa, Arafat Al-Dhaqm, Maged Nasser, Tusneem Elhassan, Hashim Elshafie, and Abdu Saif. "Financial fraud detection based on machine learning: a systematic literature review." *Applied Sciences* 12, no. 19 (2022): 9637.
- [13] Shi, Si, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. "Machine learning-driven credit risk: a systemic review." *Neural Computing and Applications* 34, no. 17 (2022): 14327-14339.
- [14] Alonso Robisco, Andrés, and Jose Manuel Carbo Martinez. "Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction." *Financial Innovation* 8, no. 1 (2022): 70.
- [15] Feng, Yuan. "Bank green credit risk assessment and management by mobile computing and machine learning neural network under the efficient wireless communication." *Wireless Communications and Mobile Computing* 2022, no. 1 (2022): 3444317.
- [16] Chhabra Roy, Neha, and Sreeleakha Prabhakaran. "Internal-led cyber frauds in Indian banks: an effective machine learning-based defense system to fraud detection, prioritization and prevention." *Aslib Journal of Information Management* 75, no. 2 (2023): 246-296.