Volume (1) No (10), 2025
Journal of Global Research in Multidisciplinary Studies (JGRMS)
Review Paper/Research Paper
Available online at https://saanvipublications.com/journals/index.php/jgrms/index

A Review of Machine Learning Techniques for Risk Evaluation in Healthcare and Insurance Systems

Neha Upadhyay
Assistant Professor
Department of Computer Applications
IIS University
Bhopal (M.P.)
neha.upadhyay887@gmail.com

***Abstract*—Financial institutions require an accurate estimation of the risk of loan default in order to reduce losses incurred by credit and sustain lending. This study proposes a robust stacking-based machine learning framework that integrates Knowledge Graph Embedding (KGE) for semantic feature enrichment with XGBoost as the final predictive model. The approach is evaluated on the Home Credit Default Risk (HCDR) dataset, comprising diverse financial, demographic, and behavioral attributes of loan applicants. A comprehensive preprocessing pipeline, including imputation, normalization, one-hot encoding, and correlation-based feature selection, ensures data quality and model generalizability. The proposed KGE-XGBoost model captures both structured tabular and relational semantics by transforming borrower-entity relationships into dense embeddings, which are concatenated with original features to form a unified representation. Experimental results demonstrate superior performance with 96.79% accuracy (ACC), 80.83% precision (PRE), 78.75% recall (REC), and an F1-score (F1) of 79.00%. The proposed model exhibits a strong ability to outperform the baseline models (Random Forest achieved ACC 94.20%, NN achieved ACC 89%, and DT achieved ACC 73%), particularly in scenarios with class imbalances. The KGE integration has been found to greatly contribute to feature expressiveness and it presents a scalable and promising credit risk assessment solution to real-life financial applications.**

***Keywords*—Loan Default Prediction, XGBoost, Knowledge Graph Embedding (KGE), Credit Risk Assessment, Machine Learning, Classification Models, Feature Enrichment.**

I. INTRODUCTION

Lending has been a traditional source of economic growth in the financial industry, providing people and companies with access to credit to consume, invest, and grow. Microfinance institutions, banks, and digital lending platforms raise good income by charging interests and any processing fees on loans [1][2]. Conventional credit evaluation practices, which are usually based upon manual analysis of income, collateral, and credit record, are slow, expensive [3][4]. This necessitates high-tech practices that not only guide loan distribution but also ensure companies' financial stability and limit exposure to risky borrowers.

Loan defaults are a chronic in the financial ecosystem despite strict evaluation frameworks. Defaults occur when borrowers fail to make scheduled payments[5][6][7], leading to financial losses for lenders and increased interest rates for future borrowers. Past economic downturns, such as the global financial crisis of 2008, widespread defaults could destabilize entire financial systems[8][9]. Predicting default risk is difficult due to the complex interplay of factors such as fluctuating incomes, job instability, market volatility, and unforeseen personal emergencies.

Machine Learning (ML) techniques have emerged as powerful solutions for predicting loan defaults, as they can identify hidden patterns. The ML algorithms unlike traditional ones have the ability to incorporate various variables including spending patterns, past transactions, demographic information and real-time financial history to produce more precise forecasts [10][11]. Classifying the borrowers as being high-risk and low-risk has been commonly and broadly applied using Techs, like LR, DT, RF, and Gradient Boosting [12][13]. Such predictive features can assist lenders in minimizing default instances and enhancing operational efficiency through automated decision-making.

The stacking ensemble methods solve the problem of incorporating several base models and incorporating their results with the help of a meta-learner to provide strong results. In stacking, the various classifiers can be used together (as with a meta-model) in loop [14][15] to predict loan defaults. This composite design combines the merits of alternative algorithms, which are therefore more dependable and precise in predicting defaults than individual algorithms [16][17]. With the introduction of stacking, financial institutions can have a competitive edge in terms of reduced credit risks, enhanced quality of loan portfolio, and enhanced confidence in loaning practices enhancing financial stability in the end.

A. Motivation and contribution

The research is driven by the increasing necessity of dependable and intelligent methods of loan default prediction within the financial industry, where the misestimation of risks may lead to disastrous financial losses as well as the loss of confidence in the automated loan issuers. The traditional single-model predictive methods are characterized by low levels of generalization, over-fitting, and the inability to effectively model the intricate interactions between borrower features, credit histories, and institutional policies, and finally, the effectiveness of a stacked ensemble combining XGBoost and KGE to utilize both predictive learning and semantic interpretation and hence increase model interpretability, reliability, and risk-sensitivity. Using a hybrid predictive

approach, this study aims to enhance loan default prediction, as described below:

- The study utilizes the HCDR dataset, a large-scale real-world benchmark, to design and validate a robust loan default prediction framework.
- A comprehensive data pre-processing pipeline is employed, including missing value imputation, normalization, categorical encoding, and filter-based feature selection, ensuring bias-free and high-quality inputs for model training.
- A is novel stacked ensemble architecture developed by integrating Extreme Gradient Boosting (XGBoost) with Knowledge Graph Embedding (KGE), enabling the system to capture deeper borrower–institution–loan relationships beyond conventional feature sets
- **The proposed model outperforms existing approaches**, achieving high ACC, PRE, REC, F1, ROC-AUC, demonstrating its effectiveness over baseline models like RF, DT and NN.
- **Extensive evaluation using a confusion matrix, ROC curve, and key classification metrics** validates the model's robustness and highlights its applicability in real-world financial decision-making.
- A valid and relatable hybrid model that could be implemented in real-life financial risk management and improve decision-making and confidence in automated lending systems.

B. Novelty with justification

This work is novel because it stacked Knowledge Graph Embedding (KGE) and XGBoost to enhance loan default prediction by identifying both statistical and semantic patterns. Unlike traditional models, the proposed framework embeds borrower–institution–loan relationships into dense vectors using algorithms enriching the input space with contextual information. This is justified by the relational nature of financial data, where risk depends on both individual features and entity interactions. Leveraging XGBoost robustness with KGE-based semantic signals, the model achieves superior reliability and ACC (96.79%) on the HCDR dataset, outperforming RF, Neural Networks, and Decision Trees, thereby offering a more explainable and risk-sensitive credit scoring solution.

C. Structure of the paper

This research is structured according to the following scheme: Section II is a review of the related literature on ML and the prediction of loan defaults and hybrid ML. The proposed methodology, such as preprocessing, knowledge graph embedding, and stacking architecture, is described in Section III. Section IV gives the results of the experiment and comparative analysis, and, Section V, gives conclusions to the study and future research directions.

II. LITERATURE REVIEW

This section presents a comprehensive literature review on loan default prediction using ML and hybrid semantic models. It explores existing approaches including conventional classifiers, ensemble techniques, and knowledge-driven learning frameworks. Additionally, Table I provides a summarized comparison of the reviewed studies discussed in this section:

Ramachandra and Vaithiyanathan 2025, This privacy-preserving approach uses a differentially private optimizer layer in the WGAN architecture to protect against membership inference attacks, which can compromise individuals' privacy. Deep-learning algorithms that classify loan defaulters require abundant data, which could be compromised through federated learning. The goal is to improve loan defaulter prediction without directly exchanging sensitive loan data. A remarkable 95% ACC is achieved by the suggested system, as shown by Federated WGAN, synthetic dataset populations, and privacy budgets [18].

Khan *et al.* (2025) study presents a novel strategy of integrating the HITL with the XGBoost, a single of the most powerful ensembles learning algorithms to provide a well-balanced model of loan default prediction. By analyzing loan data of Lending Club, this approach has reached a 99.4% ACC and a high PRE and high REC, hence indicating balanced performance of the model [19].

Zhou (2024) trained an individual stacked model for each loan client based on personalized features. The data used contains information about fifteen million loan applicants, their default status, and 468 features in all. 41 of the features that can be quantitatively analyzed are selected according to the feature importance output by a RF model. The stacked model consists of two layers, in which a LGBM classifier is the base learner, and an LR model is the meta learner. The stacked model outperforms the individual Logistic Regression model but performs nearly the same as the individual LGBM Classifier. the stacked models trained with personalized features result in AUC=0.772 and F1=0.188[20].

Kumar Jain *et al.* (2024) an analytics model that uses machine learning to improve the banking industry's ability to anticipate loan defaults using open P2P loan data from Lending Club. The model employs cutting-edge ML techniques, including RF, to enhance the ACC and reliability of predictions. Credit history, loan purpose, and debt status are among the crucial borrower variables the algorithm considers when identifying high-risk borrowers. With parameters like 89% accuracy, 99.5% sensitivity, and 80.3% specificity, RF emerged as the most successful ML model among those that were tested [21].

Pathak *et al.* (2023) a thorough examination of the Lending Club dataset for the purpose of predicting loan defaults. Data cleansing, EDA, and standardization are all a part of the research, as is the use of foundational ML models and an ensemble model. In the initial stage of the dataset, factors significant to loan default prediction are identified through data filtering and EDA approaches. Basic ML models like as DT, LR, and RF are trained using preprocessed data in order to forecast loan default. An ACC of about 73% is attained by these models. Employing an ensemble learning approach further enhances prediction accuracy. The ensemble model, which pooled the results of multiple separate models, increased ACC by 76.8% [22].

Gao, Yang and Wang (2023) This research aims to use ML methods to sift through mountains of data on bank loan defaults. The goal is to develop a BP neural network-based loan default prediction model that can identify potential default risks early on and help mitigate them proactively. After evaluating the model's performance, ten variables used as input features. Finally, a 2-hidden-layer BP neural network determined. Multiple K-value models are assessed, concluding with a K-value of 4 for the KNN algorithm validates the two models BP neural network and KNN

algorithm, analyzing ACC, REC, and PRE. The BP neural network model achieves an ACC of 70% or more, a REC rate exceeding 55%, and a PRE rate surpassing 60%[23].

Lakshmanarao *et al.* (2023) developed a plan to use ML and DL models to forecast when loans go into default. Loan default data from Lending Club is used in this analysis. To obtain a preprocessed dataset, the dataset is subjected to multiple data preprocessing programs. Future ML methods suggested included DT, LR, K-NN, and feed forward NN. Experiment findings showed that the proposed feed forward NN had a 91% success rate in forecasting repayment defaults. [24].

TABLE I. SUMMARY OF THE RELATED WORK FOR LOAN DEFAULT PREDICTION USING MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

Author	Dataset	Methodology	Results / Analysis	Advantages	Limitations	Future Work
Ramachandra & Vaithiyanathan (2025)	Loan data	Federated WGAN with differentially private optimizer to generate synthetic data for privacy-preserving	Achieved 95% accuracy while protecting against membership inference attacks	Protects data privacy; synthetic dataset generation ensures security	Requires large computation; synthetic capture real-world complexity	Explore scalability of Federated WGAN
Khan et al. (2025)	Lending Club dataset	HITL + XGBoost ensemble	99.4% accuracy, high precision, high recall	Balanced, robust model; incorporates human-in-the-loop for validation	HITL integration increases complexity and dependency on human feedback	Extend HITL with deep learning real-time deployment in banking systems
Zhou (2024)	15 million loan applicants,	Stacked model (LGBM as base + Logistic Regression as meta)	AUC = 0.772, F1 = 0.188; outperforms Logistic Regression but comparable to LGBM	Personalization using client-specific features; interpretable stacking	Low F1 score; minimal improvement	Improve feature engineering; hybrid stacking with DL models
Kumar Jain et al. (2024)	Lending Club dataset (open P2P loan data)	Random Forest compared with LR, DT, SVM	RF: 89% accuracy, 99.5% sensitivity, 80.3% specificity	Identifies high-risk borrowers effectively; superior sensitivity	Accuracy lower compared to newer DL methods; feature limitations	Explore deep ensemble models; include external borrower behavior data
Pathak et al. (2023)	Lending Club dataset	Data filtering, EDA, standardization → ML models (LR, DT, RF) + Ensemble	Individual ML 73% accuracy; Ensemble: 76.8% accuracy	Strong preprocessing pipeline; ensemble improves performance	Limited accuracy compared to advanced models; weak generalization	Apply hybrid ML-DL models; add socio-economic contextual features
Gao, Yang & Wang (2023)	Bank loan default data	BP Neural Network & KNN (K=4)	BP: 70% accuracy, recall >55%, precision >60%	Neural network captures non-linearity; early risk detection	Accuracy relatively low; limited feature selection	Improve with feature explore CNN/LSTM for time-series pattern
Lakshmanarao et al. (2023)	Lending Club dataset	ML models (DT, RF, LR, KNN) + DL (Feedforward NN)	Feedforward NN: 91% accuracy	High performance with DL; comparative study with ML	Overfitting risk due to high reported accuracy; dataset bias	Extend to cross-platform datasets; explore federated DL

A. Research Gap

Despite notable progress in ML and DL for loan default prediction, several key challenges remain unresolved. Many existing approaches rely on single models or basic ensemble techniques, offering limited generalization and reduced reliability in complex financial environments. Although methods like SMOTE and ADASYN have been used to address class imbalance, they do not fully capture the complex relationships between borrower attributes and default behavior in diverse datasets, such as Lending Club and HCDR. Furthermore, knowledge-driven feature enrichment methods like Knowledge Graph Embedding (KGE) remain underexplored, and their integration into advanced ensemble frameworks is rare. In particular, few studies have investigated the combination of KGE with robust meta-learners such as XGBoost within a stacked architecture. This gap underscores the need for a reliable, high-ACC predictive framework that leverages both semantic enrichment and stacking techniques to improve interpretability, robustness, and real-world applicability in financial risk management.

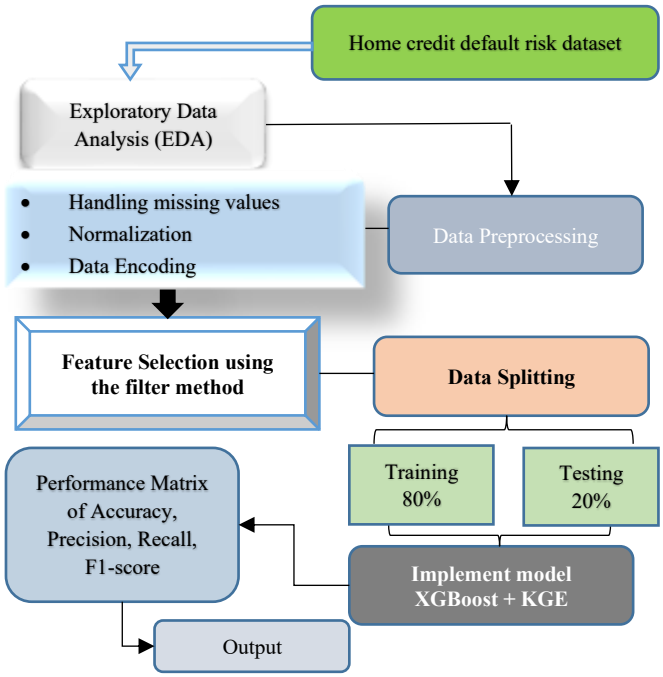

Fig. 1. Flowchart for Loan default prediction using stacking models

III. METHODOLOGY

The process of predicting loan defaults is an important undertaking of financial risk management because it helps lenders identify high-risk borrowers before loans are disbursed. Figure 1 shows the methodology that was used to conduct this study. First, a dataset of HCDR was acquired and put through EDA to understand the distribution of features, identify outliers, and determine class imbalance. The next step involved data preprocessing, including tasks such as classifying categorical variables, normalizing numerical variables, and handling missing values. To improve the model's efficiency and decrease dimensionality, the filter-based technique was used to select features while retaining the most important predictors. Stratification was used to keep the class proportions consistent, and the processed data was then split into training and testing sets of 80 and 20, respectively. During the model-building phase, stacking has been deployed, where different base learners, including XGBoost and KGE were trained Final, the performance of the proposed stacking framework was measured using standard measures, i.e., ACC, PRE, REC, and F1 structured statistical properties and relational semantics to enhance predictive ACC and stability, especially under the condition of class imbalance.

A. Data acquisition

The data for this study is sourced from the HCDR dataset, specifically utilized for loan default prediction. It comprises anonymized information on applicants, including their demographic details, financial history, previous loans, monthly installments, and repayment records, making it well-suited for modeling and predicting the likelihood of loan repayment failures. Table II presents the description and key properties of the dataset obtained from various sources.

TABLE II. DATASET DESCRIPTION AND KEY ATTRIBUTES

Data Source	Description	Key attributes
application.csv	Information about the loan applicant (anonymized) and loan at application time	Demographic information (i.e., age, gender, family status, etc.), employment type, years in business/employment, income, loan information (loan type, requested amount), external source's data
bureau.csv	Borrower's previous credits provided by other banks and financial institutions	Number of loans active/close, total loan exposure, total overdue amount, remaining term, number of defaults.
bureau_balance .csv	Borrower's monthly data of prior credits in the bureau	Monthly status of availed credits, i.e., regular/overdue payment
previous_applic ation.c	Information on previous loan applications and their status for the applicants	Loan amount, loan type, loan duration, decision (approve/reject)
POS_CASH_b alance.csv	Monthly data on previous sales or cash loans by current customers	Monthly balance, term of cash loan, loan status
credit_card_ balalance.csv	Monthly balance snapshots of previous credit cards	Credit limit, utilized amount, receivable amount, payments
installments_pa	Payment history for previous loans	Installment size, last paid amount, overdue amount, status

B. Data visualization

Data visualization was used to analyze correlations, demographic patterns, and occupational distributions in the HCDR dataset. Insights from heatmaps and bar charts revealed multicollinearity, gender-based repayment risks, and occupation-linked credit behaviors, guiding reliable feature selection and risk profiling and the correlation heat map, as shown in Fig. 2.

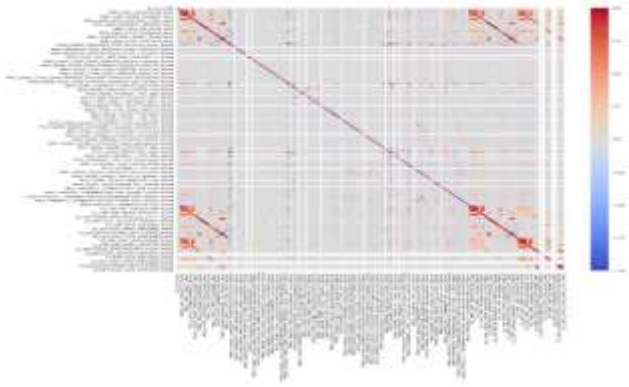

Fig. 2. Correlation Metrix of Home Credit Default Risk Dataset

Figure 2 shows a heatmap of the correlation matrices for the HCDR dataset, which was obtained from the Seaborn library. It illustrates the linear relationships among numerical features from the previous application table. Highly correlated features are shown in red or blue, indicating strong positive or negative correlations, respectively. This aids in multicollinearity detection and dimensionality reduction for improved model reliability.

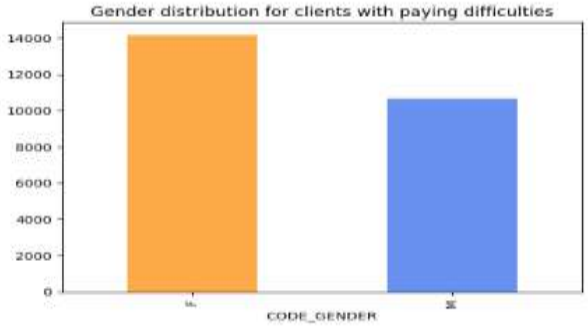

Fig. 3. Gender Distribution for Client based on the paying difficulties

Figure 3 is a bar chart that depicts the gender distribution of clients from the HCDR dataset who are having trouble repaying their loans. It reveals that females (F) exhibit a higher incidence of payment issues compared to males (M), with approximately 14,000 cases for females and 11,000 for males. This suggests potential gender-related risk factors in credit default behavior.

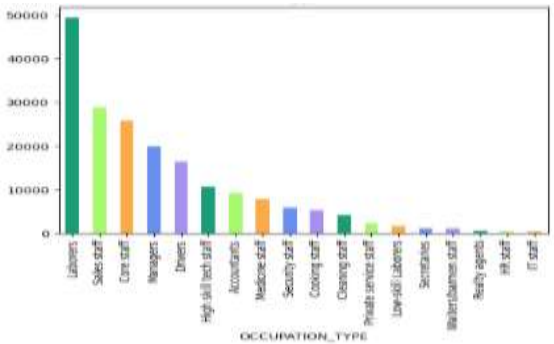

Fig. 4. Occupation TYPE for members without difficulties in the home credit risk dataset

Figure 4 presents the occupational distribution of clients without repayment difficulties in the HCDR dataset. The highest count is for Laborers (49,000), followed by Sales staff (28,000), Core staff (26,000), and Managers (19,000). Occupations like IT staff and HR staff have the lowest counts (500–1,000), highlighting occupational stratification relevant to credit risk profiling.

C. Data preprocessing

Data quality and integrity are guaranteed by the use of data pretreatment techniques including normalisation and missing value imputation [25]. A number of preprocessing procedures are necessary to ensure that ML models used to predict loan defaults have a low error rate. This process enhances model performance, generalizability, and reliability by reducing noise, inconsistencies, and bias in the dataset. Further Preprocessing Steps are briefly explained below:

- **Missing Value:** The dataset contains missing values that can be either filled in or omitted. As a missing value technique, use the mode, median, or mean [5]. Consequently, this study employed a Scikit-learn pipeline to imputationally determine the mode for categorical data and the mean for numerical variables.

D. Data Normalization with Standard Scale

The features were standardized using the StandardScaler from the scikit-learn package, a widely used tool for feature normalization in machine learning[26]. The StandardScaler ensures that all features are on the same scale by removing their means and scaling them to unit variance. Following the steps outlined in equation (1), can determine the average score for each attribute:

$$z = \frac{x-u}{s} \quad (1)$$

feature's standard deviation (s), feature's mean (u) over training data, and feature's value (x) are all variables in this context. Divide the data by the standard deviation to scale it and eliminate the mean. This is the usual behaviour of the StandardScaler.

E. Feature encoding

A common requirement for neural network processing is the transformation of categorical data used in loan default prediction models, such as the borrower's marital status, job status, and loan type [27]. The inability of neural networks to process categorical input necessitates the employment of other encoding methods, such as Label Encoding and One-Hot Encoding. The One-Hot Encoding method uses binary columns to represent each category. The category is present if the value is "1" and absent if the value is "0." This assists the neural network in treating every category separately, thus making no false presumptions of order or precedence. Although One-Hot Encoding can expand the number of features in the model, particularly on variables containing a lot of categories, it also guarantees higher ACC of a model.

F. Feature selection using filter method

Feature selection is a crucial step in enhancing neural network performance by eliminating irrelevant, redundant, or noisy features[28]. Model correctness, overfitting reduction, and computing efficiency can all be achieved through the use of various strategies. These techniques are generally classified into three categories: Three Approaches: Filtering, Embedding, and Hybrid. For high-dimensional datasets, filter methods excel computationally because they assess the importance of each feature separately from any learning algorithm. This makes them ideal for loan default prediction, which involves high-dimensional financial, demographic, and behavioral attributes. With just the most important features kept, were able to simplify the model and speed up training.

G. Data splitting

The dataset is divided into 80:20 stratified train-test partitions with a random state of 10, which guarantees consistency in class distribution and repeatability across partitions. It is possible to maintain the ratio of default to non-default classes in the training and testing sets using stratification.

H. Proposed stacked XGBoost + KGE model for loan default prediction

The model is a stacking-based ensemble designed to enhance classification performance on structured **tabular data** by incorporating semantic knowledge through **Knowledge Graph Embeddings (KGE)** and using **XGBoost** as the final prediction engine. It begins with a traditional **dataset**, represented as feature vectors $x_i \in \mathbb{R}^m$, where m denotes the number of engineered numerical and categorical features[29]. Normalization, label encoding, and missing value imputation are some of the common pre-processing steps used on this data. Parallel to this, a **Knowledge Graph** is constructed to represent real-world relationships among entities such as borrowers, institutions, accounts, and geographic locations. The format (h, r, t) triples are embedded using a KGE approach such as TransE, DistMult, or ComplEx. In this case, a head entity (h), a relation (r), and a tail entity (t). This results in an embedding $h_i \in \mathbb{R}^d$ with a fixed length for every entity. The **tabular features** and **semantic embeddings** are then **concatenated** to form a unified stacked feature representation is equation (2):

$$z_i = [x_i \parallel h_i] \in \mathbb{R}^{m+d} \quad (2)$$

This packed input z_i is inputted into an XGBoost classifier, which creates an ensemble of the gradient-boosted decision trees that used to factor in complex interactions and correct imbalance of classes. This model has its logistic objective function in binary classification whereby the output probabilities are generated in the form of eqn (3).

$$P(y_i = 1 \mid z_i) = \frac{1}{1+\exp(-f(z_i))} \quad (3)$$

where f(·) represents the forecast of XGBoost ensemble. XGBoos binary logistic objective is used to do training and maximize the binary cross-entropy loss in eqn (4).

$$L = -\sum_{i=1}^{n}\left[y_i \log\left(P(y_i)\right) + (1-y_i)\log\left(1-P(y_i)\right)\right] \quad (4)$$

The ACC, PRE, REC, and F1 metrics are used to assess the model's efficacy. As for other metrics. The F1-score is prioritized due to the data set's inherent imbalance. Such tabular learning and semantic embedding stacking-based integration enables the model to learn not only the statistical cues but also the relational information and thus becomes extremely useful when it comes to structured prediction tasks, including loan default prediction or fraud detection.

I. Performance metrics

A model evaluation is of the utmost significance in any predictive modelling attempt. Binary classification is a sort of classification that uses the prefixes "P" and "N" to separate

data into two groups. With this sorting, get two right categories (TP and TN) and two incorrect ones (FP and FN). Table III provides the confusion matrix for these four outcomes, which are utilised to assess the binary classifier.

TABLE III. CONFUSION METRICS

	Predicted Positive	Predictive Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

1) Accuracy

A popular metric, accuracy shows what percentage of the model's predictions were correct out of all the possibilities [26]. The total number of instances, TP, and TN are added together to achieve this ratio. More specifically, the following formula (5) determines the accuracy:

$$Accuracy = \frac{TP+TN}{Total\ instances\ of\ data} \quad (5)$$

2) Precision

The ACC, consistency, and reliability of the model's predictions is the percentage of accurate predictions. The expression for it is equation (6):

$$Precision = \frac{TP}{TP+FR} \quad (6)$$

3) Recall

The recall measures a model's ability to identify actual positive instances; it is calculated as the ratio of the total number of real positives to the number of right positive predictions. The calculation is shown in equation (7):

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

4) F1 Score

The F1-score is a well-rounded assessment that considers both the positive and negative consequences. It is calculated by harmonically adding the PRE and REC ratings. It particularly comes in handy when there is an imbalance in the classes, since it does not make the model rate itself on one metric, performance of classifying the positive class. The F1-score is calculated as in equation (8):

$$F1-\text{score} = \frac{2\times recall\times precision}{recall+precision} \quad (8)$$

5) ROC Curve

One tool for visualizing how a binary classification model performs at different threshold levels is the ROC curve. The curve is a plot of the TPR also known as the REC versus the FPR at different thresholds. The FPR is derived as in equation (9):

$$FPR = \frac{FP}{FP+TN} \quad (9)$$

These indicators are used to evaluate the model's efficacy. Below, Shows the results of the evolutionary models:

IV. RESULT ANALYSIS AND DISCUSSION

This section describes the experimental findings to predict loan default by stacking machine learning structure on HCDR dataset. The suggested model combines the representation of semantic features using Knowledge Graph Embedding (KGE) and XGBoost as the ultimate classifier. Performance is measured based on such key metrics as ACC, PRE, REC, F1. It was implemented with Python programming language and in a Jupiter Notebook setting on Google Colab. The libraries, including Keras, TensorFlow, NumPy, Pandas, Seaborn, and Matplotlib, were used. The experiments have been run on a hardware with an NVIDIA GTX 1660i (8 GB VRAM) + 16 GB RAM, which has adequate computational memory to run the training and testing of stacked XGBoost + KGE. The next parts would elaborate on the performance consequences of the suggested method.

TABLE IV. XGBOOST + KGE MODEL PERFORMANCE ON HOME CREDIT RISK DATASET

Measure	XGBoost + KGE
Accuracy	96.79
Precision	80.83
Recall	78.75
F1-score	79.00
ROC AUC	83.60

Table IV provides the quantitative analysis of the suggested XGBoost + Knowledge Graph Embedding (KGE) model used with HCDR dataset, which is expected to improve the reliability of loan default prediction. The model achieves a high classification accuracy of 96.79, indicating high overall correctness in binary classification. An ACC of 80.83% indicates that the model is very strong at reducing FP, which is very important in financial risk scenarios. The F1 of 79.00% ensures that the model performs balancedly in terms of REC and ACC, while a score of 78.75% shows that the model is highly sensitive in predicting actual defaulters. The model's capacity to distinguish between default and non-default classes is further demonstrated by its remarkable ROC AUC of 83.60%. The integration of KGE significantly contributes to improved feature representation, enabling more accurate credit risk stratification.

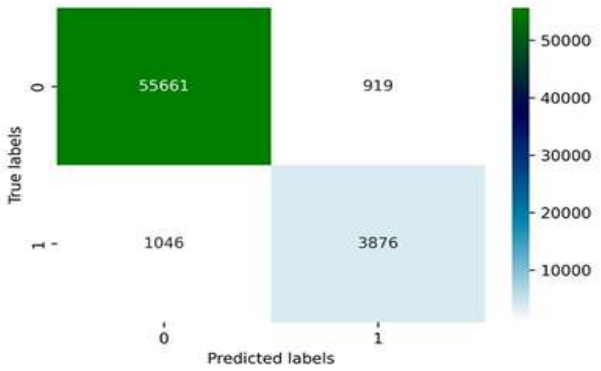

Fig. 5. Confusion matrix of XGBoost + KGE model

The XGBoost with KGE model's predictions on loan defaults using the HCDR dataset are summarized in Figure 5, the confusion matrix. A full range of outcomes, including correct, incorrect, and non-existent results, are detailed in the matrix. The diagonal values (55,661 and 3,876) represent correctly classified non-defaulters and defaulters, respectively. The off-diagonal values indicate misclassifications: 919 non-defaulters were incorrectly predicted as defaulters, and 1,046 defaulters were misclassified as non-defaulters.

The XGBoost + KGE ROC curve on the HCDR data is illustrated in figure 6. The curve attains AUC of 0.836 with a high level of discrimination of defaulters and non-defaulters. Its sharp ascent to the left upper quadrant indicates a good classification achievement when there is an imbalanced collection of data.

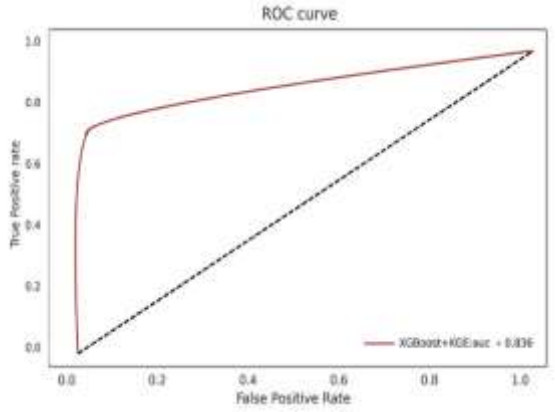

Fig. 6. Roc curve of XGBoost + KGE model

A. Discussion

This section gives a comparative analysis of loan default prediction based on the HCDR dataset. The suggested XGBoost + Knowledge Graph Embedding (KGE) model is compared with the current models such as RF, NN, and DT as presented in Table V The evaluation is performed in terms of the main key performance measurements such as accuracy.

TABLE V. COMPARISON BETWEEN PROPOSED XGBOOST + KGE MODEL AND EXISTING MODELS FOR LOAN DEFAULT PREDICTION

Measure	Accuracy
Proposed XGBoost + KGE	**96.79 %**
Random Forest[30]	94.20 %
Neural Network model[31]	89.0 %
Decision Tree[32]	73.0 %

The performance comparison of the proposed stacking-based XGBoost + Knowledge Graph Embedding (KGE) model with the traditional models, which are RF, NN, and DT, as presented in Table V. The results indicate that the combination of XGBoost and KGE showed the highest ACC of 96.79, which easily outperforms other baseline models. In comparison, the RF classifier attained an ACC of 94.20%, which, although competitive, remained lower than the proposed model. Similarly, the NN model yielded 89.0%, highlighting its limited capability in capturing complex feature interactions within the dataset. The DT classifier reported the lowest ACC of 73.00%, indicating its susceptibility to overfitting and inability to generalize effectively. These results show that the suggested stacking model is more accurate in predicting loan defaults than other methods.

The integration of KGE for semantic feature augmentation and XGBoost as a predictive engine reflects a stacking-inspired hybrid architecture, where knowledge-driven feature learning is layered beneath gradient-boosted decision-making. This makes the model more generalizable, which improves its ability to depict intricate linkages in borrowers' actions. Especially in risk-sensitive settings, its balanced performance proves it is beneficial for loan default prediction. Nonetheless, the architecture introduces computational overhead due to embedding construction and model complexity, suggesting future scope for optimizing the stacking pipeline or incorporating lightweight meta-learners for scalability.

V. CONCLUSION AND FUTURE WORK

Loan default prediction is one of the most important issues in financial risk management, where prompt and accurate identification of high-risk borrowers can go a long way to minimize losses and to provide better credit allocation traditional models in modeling complex borrower relationships, this paper suggested a stacked machine learning framework that integrates XGBoost with Knowledge Graph Embedding (KGE) to predict them better. The combination of tabular financial characteristics and semantic descriptions based on the real-life entity relationships enables the model to utilize both statistical cues and contextual information the proposed stacked XGBoost + KGE model was tested on HCDR dataset and showed high predictive ACC according to various evaluation metrics. The model has an ACC of 96.79%, F1 of 79.00%, and ROC-AUC of 83.60, which is better than the baseline models such as the RF, NN, and DT. The equal ACC and REC metrics suggest that the model is strong in overcoming the issue of class imbalance, which is a usual problem with credit scoring data. The knowledge graph element led to better representation of the features, whereas the XGBoost gradient-boosted decision tree offered good classification and generalization. Further research should be done to simplify this complexity in future work by examining lightweight KGE methods, Real-time deployment and explainability framework, e.g. SHAP, or LIME, can also be explored in order to enhance transparency, regulatory compliance and trust in automated lending systems. Altogether, stacked architecture provides a scalable and smart platform of next-generation loan default prediction systems.

REFERENCES

[1] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction," *Mathematics*, vol. 12, p. 3423, 2024, doi: 10.3390/math12213423.

[2] C. Franco, P. Madrazo-Lemarroy, J. Beltrán, J. A. Núñez Mora, and P. Moncayo, "Loan Default Prediction: A Complete Revision of LendingClub," *Rev. Mex. Econ. y Finanz. Nueva Epoca*, 2023, doi: 10.21919/remef.v18i3.886.

[3] B. Chaudhari, S. Verma, and S. Somu, "A Review of Secure API Gateways with Java Spring for Financial Lending Platforms," vol. 14, pp. 315–326, 2024, doi: 10.56975/ijcsp.v14i4.303090.

[4] B. Chaudhari, S. C. G. Verma, and S. R. Somu, "Transforming Financial Lending: A Scalable Microservices Approach using AI and Spring Boot," *Int. J. Sci. Res. Mod. Technol.*, pp. 72–81, Aug. 2024, doi: 10.38124/ijsrmt.v3i8.527.

[5] A. Egwa, B. Habeeb, A. Ajiya, and M. Suleiman Bizi, "Default Prediction for Loan Lenders Using Machine Learning Algorithms," pp. 1–12, 2022.

[6] X. Zhu, Q. Chu, X. Song, P. Hu, and L. Peng, "Explainable prediction of loan default based on machine learning models," *Data Sci. Manag.*, 2023, doi: 10.1016/j.dsm.2023.04.003.

[7] J. D. Turiel and T. Aste, "Peer-to-peer loan acceptance and default prediction with artificial intelligence," *R. Soc. Open Sci.*, 2020, doi: 10.1098/rsos.191649.

[8] V. Padimi, S. T. Venkata, and D. N. Devarani, "Applying machine learning techniques to maximize the performance of loan default prediction," *J. Neutrosophic Fuzzy Syst.*, vol. 2, no. 2, pp. 44–56, 2022.

[9] V. Singh, "Predicting Loan Default Risk in P2P Lending Platforms: A Study of Lending Club Borrowers," *Int. J. Sci. Res.*, vol. 12, no. 11, pp. 2255–2260, 2023.

[10] J. Chen, "Research on Financial Loan Default Prediction Based on Multi-Model Ensemble and Custom Thresholds," *Trans. Comput. Sci. Intell. Syst. Res.*, vol. 7, pp. 666–674, 2024, doi: 10.62051/7dnjhn18.

[11] Bhushan Chaudhari, S. Chitraju, and G. Verma, "Synergizing Generative AI and Machine Learning for Financial Credit Risk Forecasting and Code Auditing," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 2, pp. 2882–2893, Apr. 2025, doi: 10.32628/CSEIT25112761.

[12] Q. Zhu, W. Ding, M. Xiang, M. Hu, and N. Zhang, "Loan Default

Prediction Based on Convolutional Neural Network and LightGBM," *Int. J. Data Warehous. Min.*, 2022, doi: 10.4018/IJDWM.315823.

[13] J. C. Alejandrino, J. P. Bolacoy, and J. V. B. Murcia, "Supervised and unsupervised data mining approaches in loan default prediction," *Int. J. Electr. Comput. Eng.*, 2023, doi: 10.11591/ijece.v13i2.pp1837-1847.

[14] V. Verma, "Security Compliance and Risk Management in AI-Driven Financial Transactions," *Int. J. Eng. Sci. Math.*, vol. 12, no. 7, pp. 1–15, 2023.

[15] W. Yin, B. Kirkulak-Uludag, D. Zhu, and Z. Zhou, "Stacking ensemble method for personal credit risk assessment in Peer-to-Peer lending," *Appl. Soft Comput.*, 2023, doi: 10.1016/j.asoc.2023.110302.

[16] Z. Kun, F. Weibing, and W. Jianlin, "Default Identification of P2P Lending Based on Stacking Ensemble Learning," in *Proceedings - 2020 2nd International Conference on Economic Management and Model Engineering, ICEMME 2020*, 2020. doi: 10.1109/ICEMME51517.2020.00203.

[17] W. Han, X. Gu, and L. Jian, "A multi-layer multi-view stacking model for credit risk assessment," *Intell. Data Anal.*, 2023, doi: 10.3233/IDA-220403.

[18] P. Ramachandra and S. Vaithiyanathan, "Fed-DPSDG-WGAN: Differentially Private Synthetic Data Generation for Loan Default Prediction via Federated Wasserstein GAN," *IEEE Access*, vol. 13, pp. 52069–52084, 2025, doi: 10.1109/ACCESS.2025.3552487.

[19] S. A. Khan, R. Salman, A.-H. M. Majid, M. Mythili, M. B. Alazzam, and others, "Enhancing Loan Default Prediction with Human-in-the-Loop and XGBoost Ensemble Learning," in *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2025, pp. 1–6.

[20] L. Zhou, "Loan Defaults Prediction Based on Stacked Models Trained by Personalized Features," *Highlights Business, Econ. Manag.*, vol. 40, pp. 422–428, 2024, doi: 10.54097/rd657111.

[21] Y. Kumar Jain, P. K. Mannepalli, K. Kaur, A. Maheshwari, J. Singh, and M. Ranka, "Effective Machine Learning-Based Predictive Analytics for Loan Default Prediction in Banking Sector," in *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*, Dec. 2024, pp. 1–6. doi: 10.1109/CCIS63231.2024.10931843.

[22] P. Pathak, A. Jain, M. Bansal, and P. S. Rana, "SentiNet: Empowering Robust Loan Default Prediction through Ensemble Modeling," in *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, 2023, pp. 1–6. doi: 10.1109/CVMI59935.2023.10464518.

[23] B. Gao, X. Yang, and Z. Wang, "An Empirical Study of BP Neural Network and KNN for Bank Loan Default Prediction," *IEEE Xplore*, 2023, doi: 10.1109/ICEMI58056.2023.10471076.

[24] A. Lakshmanarao, C. Gupta, C. S. Koppireddy, U. V Ramesh, and D. R. Dev, "Loan Default Prediction Using Machine Learning Techniques and Deep Learning ANN Model," in *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS)*, 2023, pp. 1–5. doi: 10.1109/AICERA/ICIS59538.2023.10420221.

[25] A. M. Abidemi, M. D. Ajegbile, Y. B. Ajegbile, J. Adedeji, and C. Dada, "A Deep Learning Prediction Model For Loan Default," *GSJ*, vol. 11, no. 7, 2023, [Online]. Available: www.globalscientificjournal.comwww.globalscientificjournal.com

[26] M. A. Kheneifar and B. Amiri, "A Novel Hybrid Model for Loan Default Prediction in Maritime Finance Based on Topological Data Analysis and Machine Learning," *IEEE Access*, vol. 13, no. April, pp. 81474–81493, 2025, doi: 10.1109/ACCESS.2025.3566066.

[27] J. Olusegun, "The Impact of Data Preprocessing and Attribute Scaling on Neural Network Accuracy for Loan Default Prediction," 2025.

[28] B. John, "Feature Selection Techniques for Enhancing Neural Network Performance in Loan Default Prediction," 2025.

[29] D. A. Agustina Pertiwi, K. Ahmad, T. L. Nikmah, Alamsyah, B. Prasetiyo, and M. A. Muslim, "Combination of Stacking with Genetic Algorithm Feature Selection to Improve Default Prediction in P2P Lending," in *2023 5th International Conference on Cybernetics and Intelligent Systems, ICORIS 2023*, 2023. doi: 10.1109/ICORIS60118.2023.10352271.

[30] I. R. Berrada, F. Barramou, and O. B. Alami, "Towards a Machine Learning-based Model for Corporate Loan Default Prediction," vol. 15, no. 3, pp. 565–573, 2024.

[31] E. S. Jayaram, "Machine Learning-Based Loan Default Prediction: Models, Insights, And Performance Evaluation In Peer-To-Peer Lending Platforms," *Educ. Adm. Theory Pract.*, vol. 30, no. 5, pp. 12975–12989, 2024, doi: 10.53555/kuey.v30i5.5637.

[32] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1022, no. 1, p. 012042, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012042.

