Associating Deep Learning and Logistic Regression for Credit Card Default Prediction

Marcus Vinicius de Oliveira
Financial Risks Department
Petróleo Brasileiro SA (PETROBRAS)
Rio de Janeiro, Brazil
marcus_voliveira@petrobras.com.br

Gilson Alexandre Ostwald Pedro da Costa
Mathematics and Statistics Department
Rio de Janeiro State University (UERJ)
Rio de Janeiro, Brazil
gilson.costa@ime.uerj.br

***Abstract*—This paper proposes an hybrid approach that integrates Deep Learning models with logistic regression to enhance the prediction of credit card defaults while maintaining the explainability of final classification using a partially explainable methodology. We compare the performance of this proposed methodology with conventional classification techniques as well as a pure Multilayer Perceptron model. The primary goal of our architecture is to improve the interpretability of the classification results, enabling users to critically evaluate the algorithm's predictions and identify the variables that have the most significant impact. This understanding aims to facilitate the development of more effective business strategies.**

***Index Terms*—Credit Default, Machine Learning, Classification, Interpretability**

I. Introduction

Default prediction for credit cards is crucial for financial institutions, as they need to balance credit approvals and risk management. Default occurs when a customer cannot meet financial obligations, leading to substantial losses that affect banks' profitability. Thus, developing accurate models for predicting defaults is essential for making informed credit decisions.

Default prediction aids institutions by enhancing collection operations efficiency. By pinpointing at-risk customers, companies can apply more effective strategies like proactive monitoring and restructuring offers. This boosts credit recovery and fosters better customer relations, making them feel supported during challenges.

Quantitative techniques are crucial for default prediction, enabling analysis of extensive historical data to detect patterns. Commonly used methods include logistic regression, decision trees, and discriminant analysis. They pinpoint important variables linked to default risk, like credit history, income, debt, consumption habits, and demographic factors.

These models might struggle with capturing complex, nonlinear relationships. Their effectiveness also hinges on data quality and updateability. Deep Learning (DL) excels here by processing data sophisticatedly, managing large datasets, and recognizing intricate patterns that traditional models might miss. Machine learning models vary greatly in terms of their degree of interpretability [1], which is a crucial aspect of this work. The purpose of the methodology developed is to provide greater interpretability to the data while also allowing for the exploration of the best characteristics of deep learning models.

Trustworthy AI involves creating and using AI systems that focus on ethics and societal values. It is based on core principles: fairness, robustness, privacy, explainability, and transparency, as shown in Figure 1. Fairness avoids bias, robustness ensures reliability, privacy protects data, explainability clarifies AI decisions, and transparency communicates system operations. These elements make AI effective, accountable, and ethically aligned. This work fits into the context of improving the explainability of Artificial Intelligence models used in the Finance area.

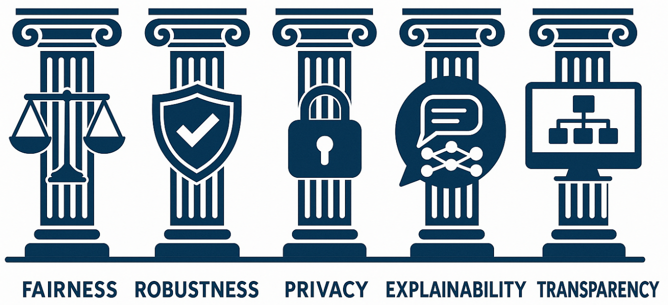

Fig. 1. The Five Pillars of Trustworthy AI.

The methodology introduced in this study draws parallels to the approach utilized in PINN (Physics Informed Neural Networks). As detailed in [2], PINN represent a novel class of neural networks that incorporate principles of physics directly into their structure and training processes. This integration facilitates a more effective modeling and simulation of complex physical phenomena. Unlike traditional models that rely exclusively on empirical data, PINN leverage governing differential equations as a fundamental component of their learning algorithms, enhancing their predictive capabilities and accuracy. In the case of this work, the use of an explainable model with coefficients derived from a Deep Learning model is analogous to an application similar to PINN, although it has a slightly different nature.

Although the topic discussed in this work is quite common in the literature, the innovation lies in the use of a hybrid, partially explainable architecture, where the final prediction is provided by the explainable model, albeit with coefficients defined by an unexplainable Deep Learning model. We believe that a methodology of this type is a good combination of the best features of both approaches: the explainability of a model such as Logistic Regression and the context-capturing ability of a Neural Network model.

II. Fundamentals

A. Associating Deep Learning and Explainable Models

The lack of explainability in Deep Learning models poses one of the primary challenges to their adoption across various fields, particularly in critical sectors such as healthcare and finance, as discussed in [3]. Although these models demonstrate high efficacy in prediction and classification tasks, their decision-making processes are often not easily comprehensible, leading to distrust and reluctance among users, who may be hesitant to incorporate the model's outcomes into their decision-making processes. Additionally, the lack of transparency can hinder adherence to regulations and ethical standards, which necessitate that automated decisions be auditable and understandable.

CFA Institute Code and Standards, [4] requires that asset analysis and management professionals, in addition to using reasonable basis, must keep a record of the intermediate steps that justify the investment decision or recommendation. Although using a model such as the one proposed in this paper does not exempt the professional from keeping the models he used, it makes their recording easier and more direct, as the coefficients can be recorded for a given set of feature values.

FinAI is a specialized area of Explainable AI (XAI) focused on financial applications, [3]. It involves methods to enhance the clarity and transparency of AI models in finance. FinAI highlights the importance of understanding model operations and outcomes, addressing accountability, transparency, and ethics. This field offers techniques to elucidate model decisions and interactions with financial data, essential for practitioners and regulators seeking transparency in AI decision-making.

Deep Learning models greatly excel in grasping data contexts, recognizing patterns, and extracting insights. Yet, the numerous parameters, intricate interactions, and non-linear functions hinder straightforward interpretation. Explainable AI enhances deep learning model explainability with techniques to render their processes and outputs more interpretable for users. Key methods by which XAI achieves this include:

Post-Hoc Interpretability: XAI typically utilizes post-hoc interpretability methods to explain the results of black-box models post-training. Techniques like LIME (Local Interpretable Model-agnostic Explanations), [5], and SHAP (SHapley Additive exPlanations), [6], offer local approximations of model predictions and feature importance.

Feature Relevance Techniques: XAI employs feature relevance techniques to determine which input features most impact deep learning model predictions. These methods calculate relevance scores, helping users grasp the factors influencing a model's decisions.

Visual Explanations: XAI employs visual tools like heatmaps and attention maps to show how input features influence model predictions, enabling users to easily understand feature relationships and their significance.

Counterfactual Explanations: These elucidate predictions by illustrating how variations in input data result in different outcomes. Counterfactual explanations assist users in grasping the conditions that could alter a model's decision.

Model-Specific Approaches: Certain XAI methods are designed for specific deep learning architectures, such as CNNs. For instance, Grad-CAM outlines areas of an image that significantly influence the model's classification.

Hybrid Models: Integrating inherently interpretable models with complex deep learning models strikes a balance between performance and explainability. These hybrids strive to preserve accuracy and enhance decision transparency.

An implementation of Hybrid models can involve employing a Deep Learning model to determine the parameters of an explainable model, which may be in the form of a specific function, a differential equation, or a stochastic process. This approach integrates the advantages of both methodologies.. Users can still manage key variables influencing outcomes, even when parameters are dictated by the Deep Learning model. This paper presents an example of this idea applied to a binary classification problem.

The methodology proposed in this work differs from LIME and SHAP in that explainability is intrinsic to the model, with the model already trained to adjust the parameters of the explainable model. LIME, [5], works by locally approximating a complex model with a simpler, more interpretable model that has good quality near a specific prediction. This allows users to understand the contribution of each feature to that particular prediction. It generates local explanations, which can vary significantly across different instances. On the other hand, SHAP, [6], provides a unified measure of feature importance based on cooperative game theory, specifically Shapley values. It attributes a model's prediction to its features, considering the contribution of each feature in all possible combinations, thus providing consistent and global insights into the model's behavior.

B. Classification Methods

In this section we briefly review typical Machine Learning (ML) classification algorithms to set the stage for the comparisons made in the Results section, emphasizing their mechanisms and suitability. These algorithms are well-established in the literature, which is why we will only briefly summarize each of them, with further details available in [7].

Logistic Regression predicts binary outcomes using predictors and transforms linear combinations into probabilities via the logistic function, excelling in binary classification with large datasets. *K-Nearest Neighbors (KNN)* classifies data based on the majority class of its K nearest neighbors using a distance metric. It's simple and effective for small to medium datasets but becomes costly for larger ones.

Decision Trees divide data hierarchically based on feature values, offering intuitive visualization but may overfit complex data. *Random Forest* combines multiple decision trees to enhance accuracy and control overfitting, making it robust against noise. *Support Vector Machine (SVM)* identifies the best hyperplane to separate classes in high-dimensional space, excelling with a clear class margin and using kernel functions for non-linear relationships.

Gradient Boosting builds models sequentially, correcting past errors and effectively capturing complex patterns, known for accuracy but requiring hyperparameter tuning. *Adaptive Boosting (AdaBoost)* combines weak classifiers into a strong one, focusing on misclassified instances for improved performance. *Extreme Gradient Boosting (XGBoost)* enhances gradient boosting with parallel computing and regularization for speed and efficiency.

LightGBM employs a histogram-based method for faster training and lower memory usage, handling large datasets efficiently and supporting categorical features directly, often outperforming traditional methods.

Multi-Layer Perceptron (MLP) is a neural network with multiple layers that captures complex data relationships but requires careful tuning and larger datasets for effectiveness, particularly for non-linear patterns.

III. LITERATURE REVIEW

The evolution of credit risk assessment is discussed in [8], highlighting the shift from traditional statistical methods and manual auditing to modern machine learning-driven models. The authors performed a systematic review of 76 research papers, focusing on the application of statistical, machine learning, and deep learning techniques in credit risk analysis. They identify challenges such as data imbalance, dataset inconsistency, model transparency, and the underutilization of deep learning models. The review findings indicate that deep learning models generally outperform traditional algorithms in credit risk estimation, and ensemble methods achieve greater accuracy than individual models.

A comparison of various credit default classification methods using the same dataset as this study is presented in [9]. Their focus is primarily on evaluating the performance of different architectures rather than proposing a different classifier. The results reported by the authors regarding the models discussed in this work are consistent with the findings presented here for the same database..

The need for robust models in the banking sector is underscored by the rising credit card usage and default risks, as examined in [10], which focuses on predicting credit card defaults using Machine Learning and Deep Learning methods. The study evaluates the accuracy, precision, and recall of algorithms such as Decision Trees, AdaBoost, and Artificial Neural Networks (ANN), revealing that Decision Trees and AdaBoost outperform ANN, although the performance differences are minor.

Another relevant topic addressed is the unbalanced credit card default data, which adversely affects prediction results. In [11], a new prediction model is proposed that combines the k-means SMOTE (Synthetic Minority Over-sampling Technique) algorithm with a Backpropagation (BP) neural network. The k-means SMOTE algorithm adjusts the data distribution, while feature importance is determined using a random forest, which sets the initial weights of the BP neural network for enhanced predictions. Experimental results indicate that this proposed model significantly improves prediction performance, with the neural network demonstrating superior overall predictive effectiveness compared to other models when feature importance is used as initial weights.

In the context of credit risk management, [12] employs Machine Learning techniques, particularly focusing on default prediction for small Italian companies. The research compares various ML classifiers against traditional logistic regression, revealing that ML techniques, especially neural networks and random forests, slightly outperform logistic regression in accuracy, although the improvement is not substantial enough to forgo classical methods. Consequently, while ML classifiers show promising results, logistic regression remains a viable option due to its interpretability and lower computational burden.

A novel model named CT-XGBoost is proposed in [13] for credit default prediction in the energy industry, specifically addressing the class imbalance problem often found in credit datasets. This model integrates the XGBoost algorithm with a cost-sensitive strategy and a threshold method to enhance its classification capabilities for default and non-default cases. The authors highlight that traditional models frequently overlook default instances due to the predominance of non-default samples.

Furthermore, [14] explores various ML models and ensemble techniques for predicting credit default risk, emphasizing the importance of addressing class imbalance through methods like SMOTE and ADASYN to boost model performance. The research utilizes hyperparameter tuning via GridSearchCV and evaluates models based on metrics such as accuracy, precision, recall, and AUC. Findings suggest that ensemble methods, particularly stacking, outperform individual models, showcasing their effectiveness in improving credit risk predictions.

The use of a Recurrent Neural Network (RNN) feature extractor with Gated Recurrent Unit (GRU) is proposed in [15] to analyze credit card payment history, capturing time-dependent features often overlooked by traditional models. This approach diverges from the one proposed in this work, as it investigates the temporal evolution of variables, while our study focuses on a cross-sectional analysis.

Lastly, [16] presents a framework for predicting company credit ratings through multimodal deep learning models that integrate structured and unstructured data. The findings reveal that a CNN-based model with hybrid fusion significantly outperforms others, emphasizing the vital role of textual data in enhancing predictive accuracy. The present work aims to augment the proposed method with a layer of explainability, making it more applicable in practical scenarios of corporate credit granting, particularly in adherence to strict corporate governance requirements that necessitate an explainable reasonable basis.

IV. Methodology

Figure 2 presents a schematic representation of the proposed architecture. From the available features, those that exhibit the highest explainability in relation to the target are chosen as factors. These selected variables will serve as inputs to the parametric model. The parameters of this model are determined by the Contextual Learning model, which incorporates all features, including those identified as factors. The primary goal of Contextual Learning is to calibrate the parameters of the parametric model based on the context conveyed by the features.

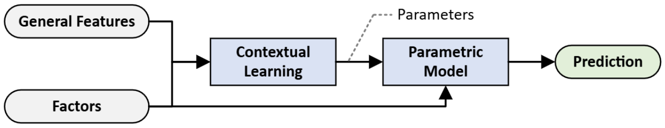

Fig. 2. Generic architecture of the partially explainable model consisting of the contextual learning model (Neural Network) and parametric model (Logistic).

The architecture depicted in Figure 2 is generic, and this study advocates for the implementation of a Deep Learning model within the Contextual Learning framework, capitalizing on its inherent ability to learn and identify patterns in the data. For this research, the selected parametric model is the logistic function, a widely used approach for binary classification tasks. Equation 1 describes the logistic function, where $\boldsymbol{x^{(f)}}$ is the factors vector and β_k corresponds to the loading of factor $x_k^{(f)}$ and $\boldsymbol{x^{(f)}} = (x_1^{(f)}, x_2^{(f)}, ..., x_N^{(f)}) \in \mathbb{R}^N$ is the vector of factors.

$$f(\boldsymbol{x^{(f)}}) = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{k=1}^{N} \beta_k x_k^{(f)}\right)} \tag{1}$$

The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_N) \in \mathbb{R}^{N+1}$ is the output of the Contextual Learning model, which receives as input all the features, general and factors, $\boldsymbol{x} = (\boldsymbol{x^{(g)}}, \boldsymbol{x^{(f)}})$, $\boldsymbol{\beta} = \Pi(\boldsymbol{x})$, where Π represents the mapping learned during the training of the Contextual Learning model.

A. Implementation Details

Figure 3 shows the architecture adopted to implement the context learning in association with logistic function. The features used as input for the neural network model consist of all the available features in the database, including those utilized in the construction of the explainable model, highlighted in the green region of the input representations in Figure 3.

The proposed architecture has two hidden layers with 100 and 50 neurons, respectively ($N_{H1} = 100$ and $N_{H2} = 50$), with total number of trainable parameters of 7,802 and L_2 regularization. All neurons adopted ReLu activation function, except the output layer witch adopts linear activation function to provide flexibility for positive and negative β's. The training process consists in minimizing the binary crossentropy loss function, Equation 2, over the training dataset and evaluation over the test dataset.

$$\mathcal{L}(\hat{y}_i) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \tag{2}$$

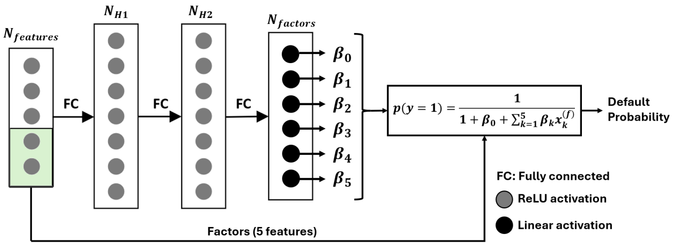

Fig. 3. Detailed architecture of the proposed architecture associating a Feed Forward Neural Network to Logistic function.

V. Database Description

The experimental database is public and available at https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients, containing 30,000 entries. Table I provides a detailed description of the dataset used in the experiments, which consists of 23 features and 1 target variable. This dataset is utilized for analyses related to credit and payment behavior. The target variable, labeled ”DEFAULT” indicates whether a default occurred, with ”0” representing no default and ”1” indicating that a default took place.

TABLE I
Dataset description consisting of 23 features and 1 target

Symbol	Type	Description
DEFAULT	Target	No dafault = 0; Defaut = 1
LIMIT_BAL	Feature	Amount of the given credit (NT dollar)
SEX	Feature	Gender (1: male; 2: female)
EDUCATION	Feature	Education level
MARRIAGE	Feature	Marital status
AGE	Feature	Age (year)
PAY_0	Feature	Repayment status for the month of September 2005
PAY_2	Feature	Repayment status for the month of August 2005
PAY_3	Feature	Repayment status for the month of July 2005
PAY_4	Feature	Repayment status for the month of June 2005
PAY_5	Feature	Repayment status for the month of May 2005
PAY_6	Feature	Repayment status for the month of April 2005
BILL_AMT1	Feature	amount of bill statement for September 2005
BILL_AMT2	Feature	amount of bill statement for August 2005
BILL_AMT3	Feature	amount of bill statement for July 2005
BILL_AMT4	Feature	amount of bill statement for June 2005
BILL_AMT5	Feature	amount of bill statement for May 2005
BILL_AMT6	Feature	amount of bill statement for April 2005
PAY_AMT1	Feature	amount paid in September 2005
PAY_AMT2	Feature	amount paid in August 2005
PAY_AMT3	Feature	amount paid in July 2005
PAY_AMT4	Feature	amount paid in June 2005
PAY_AMT5	Feature	amount paid in May 2005
PAY_AMT6	Feature	amount paid in April 2005

A. Factors Selection

Figure 4 presents an analysis of the correlation coefficients between the features. There is a group of variables with high correlation between them, representing the short-term history of bill states. Table II presents the Point Biserial correlation between the target and the continuous features. As can be seen in the second column, the correlation coefficient is very low and some of them are equal to zero considering 95% confidence.

Since it is expected that individuals with higher bills to have a higher LIMIT_BILL, it is logical to handle the normalized features by dividing each feature value by **LIMIT_BAL**. Consequently, we create normalized features, and their correlation values and p-values are detailed in the final two columns of the table. Figure 5 illustrates the correlation coefficients between these normalized features.

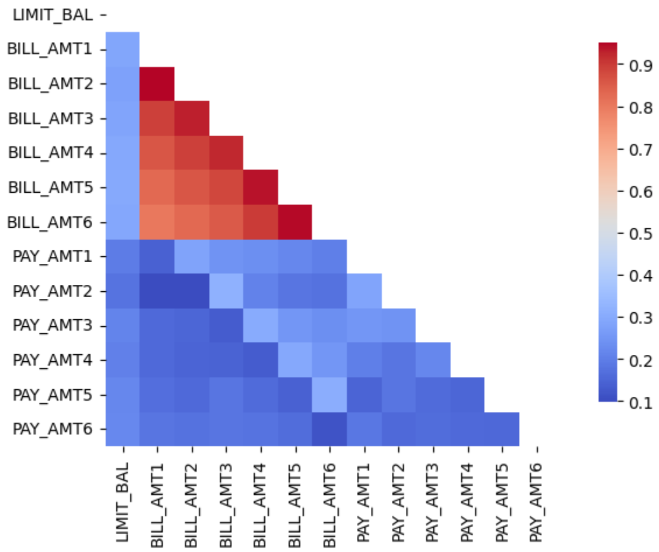

Fig. 4. Pearson Correlation Coefficient for continuous features.

Overall, the variables chosen as factors should be those with the highest explanatory power concerning the target. This is because users can derive insights on which factors are more or less significant for each new record.

TABLE II
POINT BISERIAL CORRELATION COEFFICIENT OF CONTINUOUS VARIABLES AND THE TARGET

Symbol	Feature	P-Value	Normazied Feature	P-Value
LIMIT_BAL	-0.154	0.0%	NA	NA
BILL_AMT1	-0.02	0.1%	0.086	0.0%
BILL_AMT2	-0.014	1.4%	0.099	0.0%
BILL_AMT3	-0.014	1.5%	0.104	0.0%
BILL_AMT4	-0.01	7.9%	0.116	0.0%
BILL_AMT5	-0.007	24.2%	0.119	0.0%
BILL_AMT6	-0.005	35.2%	0.123	0.0%
PAY_AMT1	-0.073	0.0%	-0.024	0.0%
PAY_AMT2	-0.059	0.0%	-0.032	0.0%
PAY_AMT3	-0.056	0.0%	-0.017	0.4%
PAY_AMT4	-0.057	0.0%	-0.016	0.4%
PAY_AMT5	-0.055	0.0%	-0.017	0.4%
PAY_AMT6	-0.053	0.0%	-0.012	3.2%

B. Class Imbalance

Class imbalance poses a significant challenge in machine learning, often causing algorithms to favor the majority class while neglecting the minority. In the domain of Credit Risk, however, the precise identification of infrequent occurrences like fraud and default is of utmost importance. To tackle this challenge, methods like oversampling and undersampling are utilized, alongside more sophisticated approaches such as SMOTE, which creates synthetic instances to even out the dataset. Alternative methods, like the generative models proposed in [17], can produce realistic synthetic samples to balance datasets and enhance rare fraud detection.

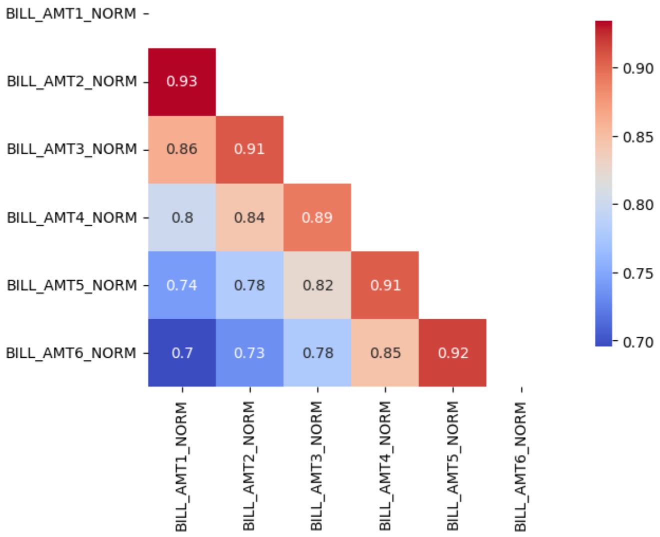

Fig. 5. Pearson correlation coefficient between pairs of the features normalized by LIMIT_BAL.

The dataset exhibits a significant imbalance, a common characteristic of default datasets, necessitating the need to rectify this imbalance prior to model training. To achieve this, the widely adopted SMOTE is employed in machine learning challenges to balance uneven datasets. This technique generates additional synthetic examples of the minority class by crafting new instances from the existing data, instead of merely duplicating them. The method involves choosing a data point from the minority class and generating new points by interpolating between it and its closest neighbors. By doing so, SMOTE assists in enhancing predictive model performance by mitigating the bias towards the majority class, thus ensuring a better balanced class representation during training.

Different numbers of neighbors were evaluated for applying SMOTE to the database, analyzing the performance of the logistic regression. It can be seen that 5 neighbors present a better result for most indicators. Figure 6 presents the results obtained.

VI. RESULTS AND DISCUSSIONS

Figure 7 shows the evolution of the loss function over the test training bases using the base after balancing. The training was performed by dividing the database into batches of 1000 records.

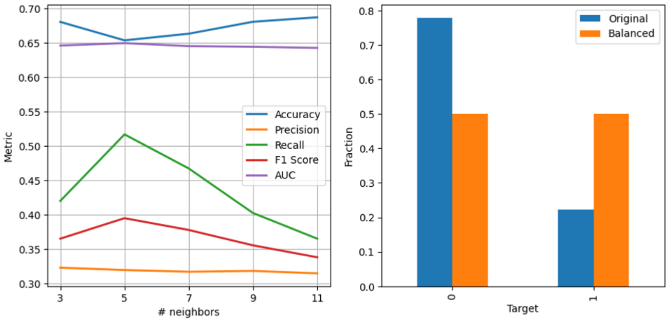

Fig. 6. Performance of Logistic regression over the test database for different number of neighbors choice for SMOTE balancing.

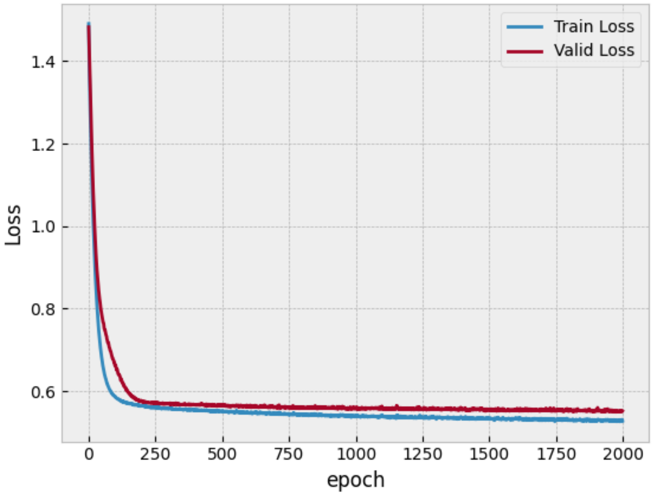

Fig. 7. Evolution of training and validation losses for de proposed architecture.

A. Metrics

Equation 3 shows the definition of Precision (P), Recall (R), F1-score ($F1$) and Accuracy (Acc) using the values count for True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

$$\begin{cases} P = \frac{TP}{TP+FP} & R = \frac{TP}{TP+FN} \\ F1 = \frac{2 \cdot P \cdot R}{P+R} & Acc = \frac{TP+TN}{TP+TN+FP+FN} \end{cases} \tag{3}$$

ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) are key metrics for assessing classification models. The ROC curve graphically shows the balance between the true positive rate (sensitivity) and the false positive rate (1-specificity) at various thresholds, illustrating the model's class separation capability. AUC measures the model's overall discriminative power, ranging from 0 to 1, with 0.5 indicating no ability (random guessing) and 1.0 denoting perfect discrimination. Hence, ROC and AUC are crucial for evaluating and comparing classification algorithms' performance.

B. Performance of the Models

Table III presents the performance results of the models used for comparison with the proposed methodology adjusted using the features directly, without any dimensionality reduction. It can be seen that the two models with the best performance are Gradient Boosting (bold) and LightGBM (italic), ranking by AUC. However, it is possible to notice that many of the models have a very similar performance. Another highlight is that the proposed model (Deep Logistic) has a considerably better performance than Logistic Regression in terms of F1 score and AUC.

TABLE III
PERFORMANCE METRICS OF ALL TESTED CLASSIFIERS FOR ALL FEATURES

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Deep Logistic	0.69 (*)	0.75	0.57 (**)	0.65 (*)	0.76 (*)
MLP	0.66	0.79 (*)	0.43	0.56	0.75 (**)
Gradient Boosting	0.68 (**)	0.74	0.55	0.63	0.74
Random Forest	0.66	0.78 (**)	0.44	0.57	0.74
LightGBM	0.66	0.75	0.48	0.59	0.74
XGBoost	0.64	0.76	0.43	0.55	0.72
AdaBoost	0.67	0.7	0.6 (*)	0.65 (**)	0.72
Logistic Regression	0.65	0.67	0.57	0.62	0.69
SVM	0.65	0.68	0.56	0.61	0.69
KNN	0.63	0.66	0.53	0.59	0.68
Decision Tree	0.6	0.64	0.44	0.52	0.6

(*) Best performing model. (**) Second best performing model

Figure 8 shows the ROC curve for Light GBM, which has the highest AUC value, as well as for traditional Logistic Regression and Deep Logistic Regression. It is possible to notice that the results are very close between LightGBM and Deep Logistic and both are significantly better than Logistic Regression.

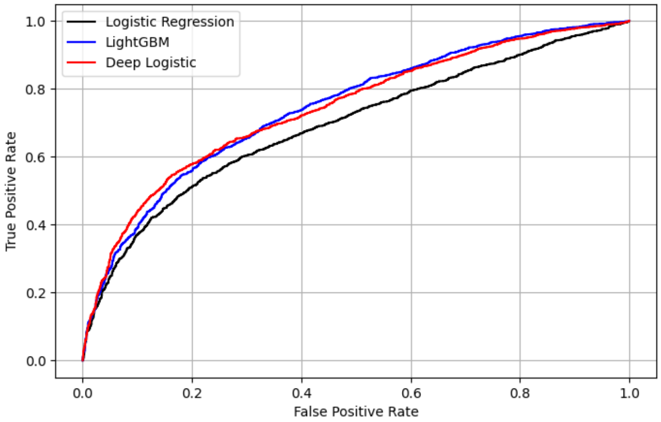

Fig. 8. ROC curve for Logistic Regression, Light GBM and Deep Logistic Regression

C. Dimensionality Reduction

Performing Principal Component Analysis for the features, we have the explained variance graph presented in Figure 9, which indicates that 6 principal components are enough to explain 90% of the data variance. For the purpose of comparing the performance of the methods applying PCA a priori, we will assume 6 principal components to describe the data.

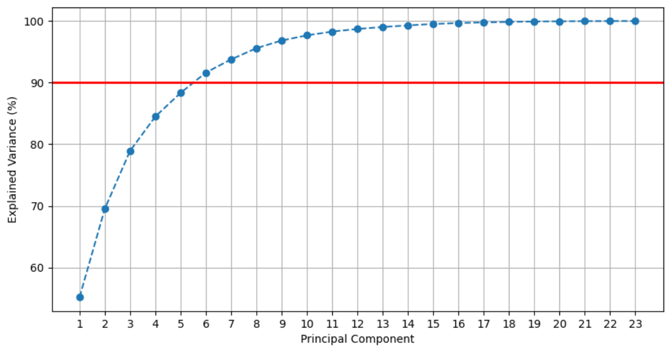

Fig. 9. Explained variance vs number of principal components. The selected number of principal components it the one capable of explaining at least 90% of the variance

It can be seen that the inclusion of PCA a priori did not result in an improvement in the performance of the tested models, as shown in Table IV. Since the objective of the proposed methodology is to allow the use of a Deep Learning model generating parameters of an explainable model, dimensionality reduction using PCA would not make much sense, since the features and factors would be replaced by principal components, compromising the interpretability of the coefficients resulting from the application of the Deep Learning model.

TABLE IV
PERFORMANCE METRICS OF ALL TESTED CLASSIFIERS USING 6 PRINCIPAL COMPONENTS AS FEATURES

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Gradient Boosting (PCA)	0.61	0.59	0.71	0.64 (*)	0.67 (**)
LightGBM (PCA)	0.6	0.59	0.68	0.63	0.66
Random Forest (PCA)	0.6	0.59	0.65	0.62	0.65
KNN (PCA)	0.61 (**)	0.63 (**)	0.53	0.58	0.65
AdaBoost (PCA)	0.57	0.55	0.76 (*)	0.64	0.65
XGBoost (PCA)	0.6	0.6	0.61	0.6	0.64
Logistic Regression (PCA)	0.58	0.58	0.55	0.56	0.62
SVM (PCA)	0.57	0.58	0.54	0.56	0.61
Decision Tree (PCA)	0.53	0.53	0.56	0.54	0.53

(*) Best performing model. (**) Second best performing model

Table V presents a comparison between traditional logistic regression using only the selected factors and using all the features. It can be seen that, in general, the performance is significantly better when all the features are used instead of using only the factors in logistic regression. The second comparison also presented in Table V is between logistic regression with all the features and the proposed methodology (Deep Logistic), in which the incremental gain in performance attributed to the fact that the Deep Learning model is responsible for capturing the context of the features is evaluated, although the final prediction is given by a logistic function. Naturally, it is necessary to recognize that the three models do not have the same level of explainability. While the first two are simpler models, with estimated coefficients that are valid for all samples, the second comes from a Deep Learning model.

TABLE V
COMPARISON BETWEEN LOGISTIC REGRESSION AND DEEP LOGISTIC

Classifier	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression (factors)	0.54	0.53	0.72	0.61	0.53
Logistic Regression (all)	0.65	0.67	0.57	0.62	0.69
Δ (all - factors)	0.11	0.14	-0.15	0.01	0.16
Deep Logistic	0.69	0.75	0.57	0.65	0.76
Δ (all - DeepLog)	0.04	0.08	0.00	0.03	0.07

D. Loadings Sensitivity

Figure 10 shows the histograms of the loadings for each of the factors across the train and test datasets. It can be observed that there is relative stability in the loadings, demonstrating that the context captured by the Deep Learning model enhances the loadings and results in better performance compared to logistic regression. Adjustments that lead to a large dispersion of β values are not appropriate, since interpretability is compromised by such erratic values. On the other hand, if the dispersion is very small, the model would be essentially equivalent to a simple logistic regression. The results obtained indicate a stability of the β defined by the Neural Network, large enough so that the model is not equivalent to a simple logistic regression (since the performance is better) and not erratic enough to make it useless.

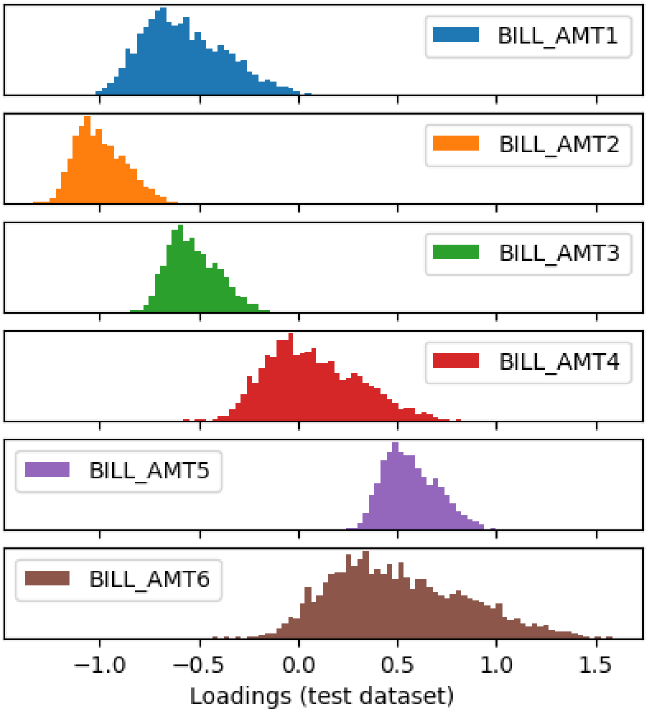

Fig. 10. Histogram of factor loadings in logistic regression for test datasets.

VII. CONCLUSIONS

This study presents a methodology that integrates a Deep Learning architecture with the logistic function as an alternative approach for classifying credit card defaults. The primary advantage of this architecture lies in its ability to leverage the strengths of Deep Learning models to identify patterns within the data while maintaining the explainability offered by the logistic function. From the available features, we selected those continuous variables with the highest correlation to the target variable, which will serve as inputs for the logistic function. The parameters of the logistic function are determined by the Deep Learning model.

The experiments conducted revealed that the proposed model outperformed Logistic Regression and performed comparably to other, while retaining the benefit of explainability. When comparing the performance of the explainable model with MLP, it is found that the inclusion of the final explainable layer does not harm the performance, on the contrary, the results indicate a better performance of the proposed model.

From a conceptual point of view, this paper presented an example of a hybrid model, a category that has gained traction for applications in which explainability is critical, such as in Finance and Healthcare. Specifically in the area of Finance, governance and legislation typically require that analyst recommendations be based on a clear rationale, such as the requirement presented in the CFA Code and Standards.

An important observation point for improving the work and for future research is to recognize that, although the results are encouraging, conducting experiments with datasets different from the one used to generate the experimental results presented here can be very useful for testing the robustness of the method in providing better results than other available classical models and non-explainable deep learning architectures, while also offering the advantage of being a partially explainable model.

References

[1] S. Consoli, D. R. Recupero, and M. Saisana, *Data Science for Economics and Finance: Methodologies and Applications*. Publisher Address: Springer Cham, 2021.

[2] S. Cuomo, V. Di Cola, F. Giampaolo, and et al., “Scientific machine learning through physics–informed neural networks: Where we are and what’s next,” *Journal of Scientific Computing*, vol. 92, p. 88, 2022.

[3] W. Yeo, W. Van Der Heever, R. Mao, and et al., “A comprehensive review on financial explainable ai,” *Artificial Intelligence Review*, vol. 58, p. 189, 2025.

[4] C. Institute, “Cfa institute code of ethics and standards of professional conduct,” 2024, accessed: 2025-06-01. [Online]. Available: https://www.cfainstitute.org/standards/professionals/code-ethics-standards

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should i trust you? explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[6] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.

[8] S. Shi, R. Tse, W. Luo *et al.*, “Machine learning-driven credit risk: a systemic review,” *Neural Computing and Applications*, vol. 34, pp. 14 327–14 339, 2022.

[9] R. Bhandary and B. K. Ghosh, “Credit card default prediction: An empirical analysis on predictive performance using statistical and machine learning methods,” *Journal of Risk and Financial Management*, vol. 18, no. 1, 2025. [Online]. Available: https://www.mdpi.com/1911-8074/18/1/23

[10] F. Wahab, I. Khan, and S. Sabada, “Credit card default prediction using ml and dl techniques,” *Internet of Things and Cyber-Physical Systems*, vol. 4, pp. 293–306, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2667345224000087

[11] Y. Chen and R. Zhang, “Research on credit card default prediction based on k-means smote and bp neural network,” *Complexity*, vol. 2021, no. 1, p. 6618841, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/6618841

[12] F. Bazzana, M. Bee, and A. Hussin Adam Khatir, “Machine learning techniques for default prediction: an application to small italian companies,” *Risk Management*, vol. 26, no. 1, 2024. [Online]. Available: https://doi.org/10.1057/s41283-023-00132-2

[13] K. Wang, J. Wan, G. Li, and H. Sun, “A hybrid algorithm-level ensemble model for imbalanced credit default prediction in the energy industry,” *Energies*, vol. 15, no. 14, 2022. [Online]. Available: https://www.mdpi.com/1996-1073/15/14/5206

[14] A. Akinjole, O. Shobayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, “Ensemble-based machine learning algorithm for loan default risk prediction,” *Mathematics*, vol. 12, no. 21, 2024. [Online]. Available: https://www.mdpi.com/2227-7390/12/21/3423

[15] T.-C. Hsu, S.-T. Liou, Y.-P. Wang, Y.-S. Huang, and Che-Lin, “Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1572–1576.

[16] M. Tavakoli, R. Chandra, F. Tian, and C. Bravo, “Multi-modal deep learning for credit rating prediction using text and numerical data streams,” *Applied Soft Computing*, vol. 171, p. 112771, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494625000821

[17] M. Tayebi and S. El Kafhali, “Generative modeling for imbalanced credit card fraud transaction detection,” *Journal of Cybersecurity and Privacy*, vol. 5, no. 1, 2025.