STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, April 2026, pp 2672–2685.
Published online in International Academic Press (www.IAPress.org)

Stacking of Ensemble and Boosting Methods for Credit Risk Prediction

Nor Nizar*, Kaicer Mohammed

Laboratory of analysis, geometry and applications, Faculty of Sciences, Ibn-Tofail University, Kenitra, Morocco

Abstract In the high-stakes domain of financial lending, the precision of credit risk models is often compromised by poorly calibrated probability estimates, particularly within imbalanced data environments. While traditional ensemble methods like boosting and bagging offer high discriminative power, their collective reliability in risk calibration remains an open challenge. This research shifts the focus toward a multi-level Meta-Learning architecture, synthesizing the strengths of heterogeneous algorithms: XGBoost, Random Forest, Extra Trees, k-NN, and Logistic Regression. The core innovation of our work lies in the Dynamic Feature Augmentation (DAF) mechanism, which we designed to formally encode model disagreement and epistemic uncertainty into the meta-learner's decision process. To ensure structural integrity and eliminate data leakage, we enforced a 5-fold Stratified Out-of-Fold (OOF) training protocol. Empirical evaluations on real-world credit datasets reveal that our DAF-Stacking framework consistently exceeds the performance of standalone learners. Notably, the architecture achieved a superior AUC of **0.8801** and significantly mitigated the LogLoss to **0.3055**, outperforming the strongest baseline, XGBoost (LogLoss 0.3122). This reduction underscores a tangible improvement in probability calibration. Furthermore, McNemar's statistical tests confirm that the error reduction relative to Random Forest and Extra Trees is highly significant ($p < 0.001$), establishing our DAF-augmented Stacking as a robust and mathematically grounded solution for modern credit risk assessment.

Keywords Credit risk prediction, Stacking generalization, Dynamic Feature Augmentation (DAF), XGBoost, Random Forest, LR, Extra Trees, Ensemble learning, SMOTE

DOI: 10.19139/soic-2310-5070-3001

1. Introduction

As the primary engines of global economic liquidity, banking institutions rely heavily on credit facilities to stimulate growth and sustain interest-driven revenue streams [1]. Yet, this financial intermediation is inextricably linked to default risk a latent threat that remains notoriously difficult to neutralize despite sophisticated risk management protocols [2]. Within today's volatile economic landscape, the imperative for financial stability has elevated credit risk assessment from a standard operational requirement to a critical strategic necessity for portfolio preservation [3].

While traditional scoring mechanisms provided a foundation for decades, they frequently fail to decipher the intricate, non-linear patterns inherent in modern borrower behavior. This gap has catalyzed a sector-wide transition toward machine learning. Although specialized algorithms such as Random Forest and XGBoost have delivered substantial improvements over conventional logistic regression [4, 5], they are not silver bullets; their performance often fluctuates due to inherent inductive biases relative to specific dataset geometries. Recent scholarly efforts, particularly those exploring solvability scoring [13] and market-specific default dynamics [12], underscore an urgent demand for frameworks that do not merely predict, but generalize robustly across heterogeneous financial environments.

*Correspondence to: Nor Nizar (Email: nizar.nor@uit.ac.ma). Laboratory of analysis, geometry and applications, Faculty of Sciences, Ibn-Tofail University, Kenitra, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

The central hypothesis of this study is that a hybrid ensemble learning model can leverage the diverse strengths of multiple algorithms to outperform standalone estimators. Building upon the foundational theories of stacked generalization [9], we propose a stacking architecture that integrates five heterogeneous base learners: XGBoost, Extra Trees, Random Forest, and k-NN. These are coordinated by a Meta-XGBoost learner, specifically optimized with L_1 and L_2 regularization to handle the augmented feature space created by our **Dynamic Feature Augmentation (DAF)** operator.

By mapping inter-model consensus and divergence into the meta-feature space addressing issues often overlooked in basic stacked generalization [11] this study provides a more calibrated decision manifold. This approach aligns with the need for more transparent and optimized financial decision-making currently sought in the field [16].

The remainder of this paper is organized as follows: Section 2 reviews the theoretical background. Section 3 describes the proposed methodology, including the DAF operator. Section 4 presents the empirical results and discussion, and Section 5 concludes the work.

2. Literature Review

The assessment of credit risk remains a fundamental challenge for financial stability, especially in the wake of global financial shifts that have prompted a reevaluation of traditional forecasting methods. Recent literature has increasingly focused on the application of advanced machine learning (ML) and artificial intelligence (AI) to enhance the predictive accuracy of credit scoring models.

Regional studies published in this journal provide critical insights into the performance of supervised learning models. For instance, **Faris and Elhachloufi [6]** conducted a comprehensive study on the Moroccan credit market, evaluating six supervised models including Logistic Regression, Random Forest (RF), and SVM. Their findings highlight the superiority of Random Forest in handling class imbalance and its strong discriminative power in local financial contexts. This is echoed by the work of **Seliem et al. [7]**, who investigated credit risk prediction with a focus on feature selection. By applying the Boruta algorithm, they demonstrated that optimizing the feature space significantly enhances the accuracy of ensemble methods like Random Forest, which achieved up to 80% accuracy in their experiments.

Beyond traditional classification paradigms, recent literature has explored innovative geometric and partitioning methodologies to refine risk assessment. A notable example is the work of **Hjouji et al. [8]**, who introduced a novel approach termed the "Method of Separating the Learning Set into Two Balls." By partitioning customers based on feature vector proximity, this technique offers a fresh perspective on class separation; however, it also highlights the inherent difficulty in classifying borrowers located at the fuzzy decision boundaries.

To bridge these diverse predictive signals, the mathematical framework of stacked generalization—initially proposed by Wolpert [9] and further refined by Breiman [10] and Ting and Witten [11] provides a rigorous foundation for meta-learning. Historical benchmarks established by Chopra et al. [3] and Ayobami et al. [12] have consistently demonstrated that such ensemble architectures tend to outperform standalone decision trees or neural networks [5]. Nevertheless, as emphasized in recent banking literature [13, 14], the persistent challenge in credit scoring is not merely the choice of models, but their optimal integration to minimize epistemic uncertainty and improve predictive calibration.

Our research bridges the gap between the ensemble superiority noted by Faris [6] and Seliem [7] and the need for more dynamic separation techniques similar to the "Two Balls" method [8]. By introducing the **Dynamic Feature Augmentation (DAF)** operator within a Stacking framework, we move beyond static feature selection to a dynamic quantification of model consensus, thereby refining the decision manifold for high-stakes credit prediction.

3. Theoretical Background

To provide a rigorous mathematical foundation for XGBoost, we first outline the principles of gradient descent applied to decision tree ensembles. Here is a detailed explanation:

3.1. Gradient Descent:

Gradient descent is an optimization method used to minimize a cost function. The idea is to adjust the model's parameters to reduce the error (the difference between the model's predictions and the actual values).

General Formulation: Let $L(y, \hat{y})$ be the loss function we want to minimize, where y is the actual value and $\hat{y}$ is the prediction. Gradient descent updates the model parameters θ according to the following rule:

$$\theta \leftarrow \theta - \eta \nabla_\theta L(y, \hat{y})$$

where η is the learning rate, and $\nabla_\theta L(y, \hat{y})$ is the gradient of the loss function with respect to the parameters.

3.2. Gradient Boosting:

Gradient Boosting is a method that builds an ensemble model of decision trees sequentially, with each tree correcting the errors made by the previous trees.

It starts with a simple model $F_0(x)$, often chosen as the mean of the target values in the case of regression and as log-odds of the proportions of the classes for binary classification.

$$F_0(x) = log(\tfrac{p}{1-p}), \quad \textbf{or } F_0(x) = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

where x represents the features of the data set, p is the proportion of the positive class and γ is the value that minimizes the sum of the loss function.

Iteration: At each step m, a new tree $h_m(x)$ is fitted to the residuals (errors) of the previous model. The residuals are the negative gradients of the loss function.

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x_i)=F_{m-1}(x_i)}$$

The model is then updated:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

where η is a learning rate (regularization factor) that controls the contribution of each new tree and helps prevent overfitting

3.3. XGBoost

XGBoost is an advanced implementation of Gradient Boosting that introduces several improvements, including regularization, handling missing values, and parallelization.

it is a supervised machine training method for classification and regression. XGBoost stands for extreme gradient boosting.

3.3.1. Mathematical Formulation: Consider a loss function $L(y_i, \hat{y}_i)$ to minimize.

The model at the t-th iteration is updated by adding a new decision tree f_t to the previous prediction $\hat{y}_i^{(t-1)}$:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

XGBoost minimizes the regularized objective:

$$L^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{j=1}^{t} \Omega(f_j)$$

where $\Omega(f)$ is a regularization term to control the model's complexity:

$$\Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2$$

where T is the number of leaves in the tree, w are the weights of the leaves, and γ and λ are regularization hyperparameters that penalize model complexity.

3.3.2. Gradient Approximation: For each tree, XGBoost adjusts the leaf weights using a second-order approximation of the loss function (Taylor expansion). The gain of a split in the tree is based on:

$$G = \sum_{i \in \text{leaf}} g_i \quad H = \sum_{i \in \text{leaf}} h_i$$

where g_i and h_i are the first and second derivatives (gradients) of the loss function, respectively.
The gain of a split is then given by:

$$\text{Gain} = \tfrac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

where G_L, G_R and H_L, H_R are the sums of the gradients and hessians for the two subsets formed by the split.

XGBoost represents an important advance in analyzing structured data, achieving a harmonious equilibrium among efficiency, precision, and flexibility. Its architecture, built upon sophisticated regularization through the γ and λ parameters, prevents overfitting by constraining the complexity of decision trees, while also intelligently managing missing data via automatic exploration of diverse branches.

The approach is grounded in mathematics, utilizing iterative minimization of a regularized cost function via gradient descent, alongside the development of decision trees, making it a powerful tool for predictive modeling

3.4. Random Forest

Random Forest is one of the most widely used supervised learning algorithms. It can be used for both regression and classification. RF makes a number of decision trees at training time and then votes them to maximize accuracy and avoid overfitting. The core principle behind Random Forest is to create a 'forest' of decision trees, each trained on a random subset of the data and features.

This random behavior helps make the model more robust and makes it more resistant to overfitting. During the training process, each tree is built using a different bootstrap sample from the original dataset. At each node of the tree, a random subset of features is considered for splitting, which guarantees diversity among the trees. For classification tasks, the resulting prediction comes from majority voting among all the trees, while for regression tasks, the result is obtained by averaging the outputs of all trees.

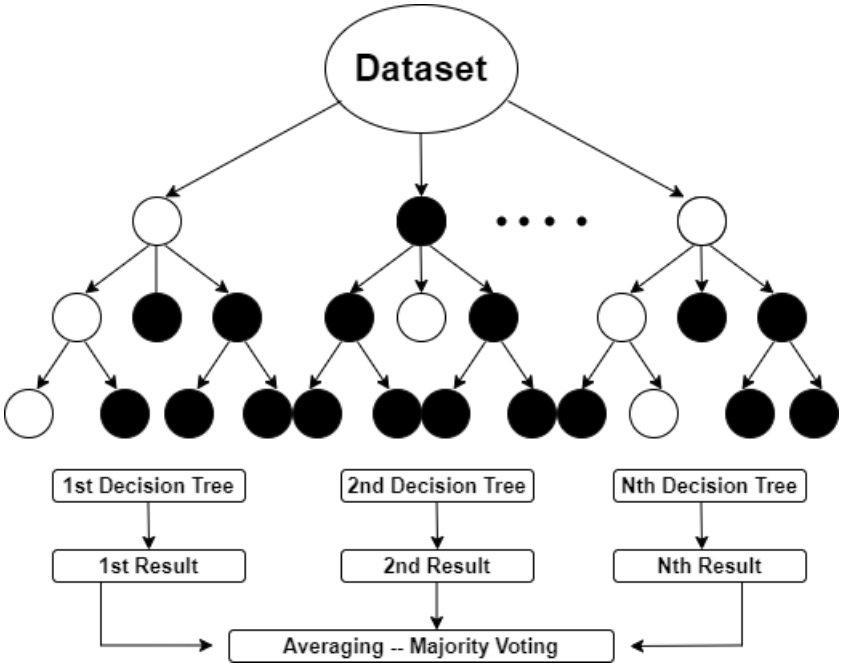

Figure 1. Random Forest Simplified

Robustness is one of the main strengths of Random Forest. By averaging the predictions of numerous trees, it avoids overfitting and improves generalization. Moreover, Random Forest provides feedback on feature importance, which can be very helpful in feature selection. It can even handle missing values, without loss of precision even when a large percentage of data are missing.

3.5. Performance Metrics and Statistical Validation

To ensure a comprehensive evaluation of the credit scoring models, we utilize a multi-faceted set of metrics. Beyond traditional classification accuracy, we focus on probabilistic calibration and statistical significance to address the requirements of high-stakes financial decision-making.

Area Under the ROC Curve (AUC) and Partial AUC (pAUC)
The AUC measures the model's ability to discriminate between good and bad payers across all thresholds. However, in credit risk, the performance at low False Positive Rates (FPR) is often more critical. Thus, we introduce the **Partial AUC (pAUC)**, calculated at a threshold of $FPR < 0.1$. It focuses on the area under the ROC curve within a specific range:

$$pAUC(e_0) = \int_0^{e_0} TPR(f)\, df \tag{1}$$

where $e_0 = 0.1$ represents the maximum acceptable rate of false alarms.

Logarithmic Loss (LogLoss)
While Accuracy only considers the final class, **LogLoss** penalizes false classifications based on the confidence of the prediction. It is a more sensitive metric for model calibration:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i)] \tag{2}$$

where $\hat{p}_i$ is the predicted probability of the positive class. A lower LogLoss indicates a better-calibrated model.

Accuracy and F1-Score
Accuracy remains a baseline metric, but given the class imbalance, the **F1-Score** is prioritized as it provides the harmonic mean of Precision and Recall:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Statistical Significance Tests
To rigorously compare the Stacking ensemble with base learners, we employ two statistical tests:

- **McNemar's Test:** A non-parametric test used on paired nominal data to determine if there is a statistically significant difference in the error rates of two classifiers.
- **Bootstrap AUC Test:** We utilize bootstrap resampling (1,000 iterations) to calculate the 95% Confidence Intervals (CI) for the AUC and determine the p-value of the difference between the Stacking model and individual classifiers.

4. Methodology and Experimental Setup

4.1. Data Description

In this work, we utilized a real dataset from a banking institution that has observed a significant increase in credit risk. In any industry, the nature of data is constantly changing, requiring regular updates to risk management systems. The present research develops a multi-stage analytical framework, initiated by a rigorous statistical characterization of the feature space and culminating in the design of a high-dimensional stacking architecture. The primary dataset comprises $n \approx 20,000$ observations characterized by $d = 47$ heterogeneous attributes, providing a granular representation of credit-related risk factors. This table summarizes the variables included in the dataset,

indicating their type, a brief description, and the proportion of unavailable data. The high rate of missing data for co-applicant variables highlights that the majority of applications involve only a single applicant, as evidenced by the feature "coapp".

Table 1. **Description of Variables in the Dataset**

Variable name	Type	Description	Missing %
nb_enf	Numerical	Number of children of the applicant.	0.0
situfam	Categorical	Family situation of the applicant.	0.2
telfixe	Categorical	Landline phone presence for the applicant.	0.1
telport	Categorical	Mobile phone presence for the applicant.	0.1
typlog	Categorical	Type of housing of the applicant (owner, renter, etc).	0.5
typhab	Categorical	Type of residence of the applicant.	0.5
coapp	Binary (0/1)	Presence of co-applicant.	0.0
sexe_app	Categorical	Gender of the applicant.	0.1
age_app	Numerical	Age of the applicant.	0.0
sex_coapp	Categorical	Gender of the co-applicant.	95.0
age_coapp	Numerical	Age of the co-applicant.	95.0
anc_banque_app	Numerical	Seniority with the bank for the applicant.	1.3
anc_banque_coapp	Numerical	Seniority with the bank for the co-applicant.	95.0
anc_emploi_app	Numerical	Employment seniority of the applicant.	27.8
catprof_app	Categorical	Professional category of the applicant.	0.3
type_contrat_app	Categorical	Type of employment contract.	0.3
secteur_app	Categorical	Employment sector of the applicant.	0.3
regime_app	Categorical	Social security/employment regime.	0.3
anc_emploi_coapp	Numerical	Employment seniority of the co-applicant.	95.0
catprof_coapp	Categorical	Professional category of the co-applicant.	95.0
type_contrat_coapp	Categorical	Type of employment contract (co-app).	95.0
secteur_coapp	Categorical	Employment sector (co-app).	95.0
regime_coapp	Categorical	Social security regime (co-app).	95.0
Salaire	Numerical	Monthly salary of the applicant.	2.1

Variable name	Type	Description	Missing %
Alloc	Numerical	Monthly allowances received.	15.4
Pension	Numerical	Pension income of the applicant.	2.0
Autre_rev	Numerical	Other sources of income.	6.8
Rev_locatif	Numerical	Rental income received.	8.3
PA_percue	Numerical	Received monthly payments.	3.7
Rentes	Numerical	Annuities received by the applicant.	6.0
Pension_inv	Numerical	Pension or investment returns.	4.2
PA_payer	Numerical	Monthly payments to be made.	0.5
Loyer	Numerical	Rent to be paid by the applicant.	2.3
Pret_immo	Numerical	Mortgage payments by the applicant.	0.7
Autres_credits	Numerical	Other loan payments.	0.6
Cession	Numerical	Wage garnishments or assignments.	1.1
Pret_voit	Numerical	Car loan payments.	0.9
nbtel	Numerical	Number of phone numbers available.	0.0
canal	Categorical	Communication channel.	0.1
tel_perso	Binary (0/1)	Personal telephone available.	0.0
tel_prof	Binary (0/1)	Professional telephone available.	0.0
gsm_perso	Binary (0/1)	Personal mobile phone available.	0.0
gsm_prof	Binary (0/1)	Professional mobile phone available.	0.0
cible	Categorical	Target: client credit profile (good/bad).	0.0
montant_dem	Numerical	Amount of credit requested.	0.01
date_acc	Categorical	Date of credit acceptance.	0.0
date_adr	Categorical	Date of last address update.	0.0

The dataset utilized in this study integrates multi-dimensional information pertaining to the primary applicant and, where applicable, the co-applicant. The features are categorized into numerical, categorical, and binary types, spanning demographic, socio-professional, financial, and behavioral domains. The target variable, **cible**, defines the credit profile as a binary outcome (good vs. bad payer).

Demographic attributes provide a profile of the household structure, including variables such as the number of dependents, marital status, gender, and age. Socio-professional features capture the economic stability of the applicants, encompassing professional categories, contract types (e.g., permanent vs. temporary), industry sectors, and employment seniority. Financial features offer a granular view of the applicants' solvency, recording diverse income streams (salaries, allowances, rental income) alongside existing debt obligations (mortgages, car loans, and other outstanding credits).

Furthermore, contact and temporal features account for applicant accessibility and residential stability, providing proxy indicators for behavioral reliability. This comprehensive feature set allows the model to evaluate both the capacity to pay and the historical propensity for financial commitment.

Table 2. **Descriptive Statistics of Numerical Features**

Variable	Count	Mean	Median	Std	Min	Max
nb_enf	19,670	0.6427	0.000	0.9701	0.00	10.000
age_app	19,670	47.667	48.000	11.746	18.00	71.000
age_coapp	2,158	47.831	48.000	12.272	18.00	71.000
anc_banque_app	19,640	6.173	6.000	4.550	0.00	49.000
anc_banque_coapp	763	74.599	68.000	59.382	1.00	369.000
anc_emploi_app	14,202	92.937	62.000	93.928	0.00	540.000
anc_emploi_coapp	978	73.541	42.000	86.498	0.00	526.000
Salaire	19,670	688.461	723.205	531.672	0.00	6,048.150
Alloc	19,670	64.940	0.000	136.905	0.00	1,588.040
Pension	19,670	256.021	0.000	409.601	0.00	4,579.750
Autre_rev	19,670	33.747	0.000	146.276	0.00	4,000.000
Rev_locatif	19,670	0.100	0.000	6.026	0.00	500.000
PA_percue	19,670	2.596	0.000	29.008	0.00	1,500.000
Rentes	19,670	17.303	0.000	232.043	0.00	15,616.598
Pension_inv	19,670	3.221	0.000	34.047	0.00	1,180.070
PA_payer	19,670	-0.446	0.000	10.534	-500.00	0.000
Loyer	19,670	-10.571	0.000	41.269	-700.00	0.000

Variable	Count	Mean	Median	Std	Min	Max
Pret_immo	19,670	-51.111	0.000	121.708	-1,466.94	0.000
Autres_credits	19,670	-131.849	-99.225	159.208	-2,500.00	0.000
Cession	19,670	-1.822	0.000	21.822	-1,165.92	0.000
Pret_voit	19,670	-9.799	0.000	51.926	-881.42	0.000
nbtel	19,667	1.455	1.000	0.612	1.00	7.000
montant_dem	19,463	2,172.121	2,000.000	921.193	0.00	5,000.000

The **age_app** variable, with 19,670 observations and no missing values, has a mean of approximately 47.7 years and a median of 48, indicating a relatively symmetric distribution with moderate variability (standard deviation: 11.75). The **nb_enf** variable shows a mean of 0.64 and a median of 0, reflecting that most applicants have no children.

Financial variables, such as **Salaire**, display substantial dispersion (standard deviation: 531.67) with values ranging from 0 to 6,048, while medians are close to the mean, suggesting moderate skewness. Other financial obligations, including **Pret_immo** and **Autres_credits**, exhibit negative values representing outflows, with medians at zero, highlighting that many applicants have no such obligations.

The **montant_dem** variable, representing the requested credit amount, has a mean of 2,172 and a median of 2,000. Its minimum value is 0, but only four observations were affected, likely due to data entry errors; these were subsequently removed. Communication-related variables, such as **nbtel**, have a mean of 1.46 and a median of 1, showing that most applicants provide at least one contact number. Co-applicant variables have high proportions of missing values, reflecting that the majority of applications involve a single applicant.

The target variable **cible** represents the credit repayment status and exhibits a pronounced class imbalance, with $16,122$ "good payers" (1) and $3,548$ "bad payers" (0). This distribution reflects a common characteristic of real-world credit datasets, where the majority of applicants successfully meet their financial obligations. Such imbalance must be explicitly considered during model development and evaluation to ensure robust and unbiased predictive performance. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) is employed; details of its integration are provided in the proposed DAF-Stacking pipeline.

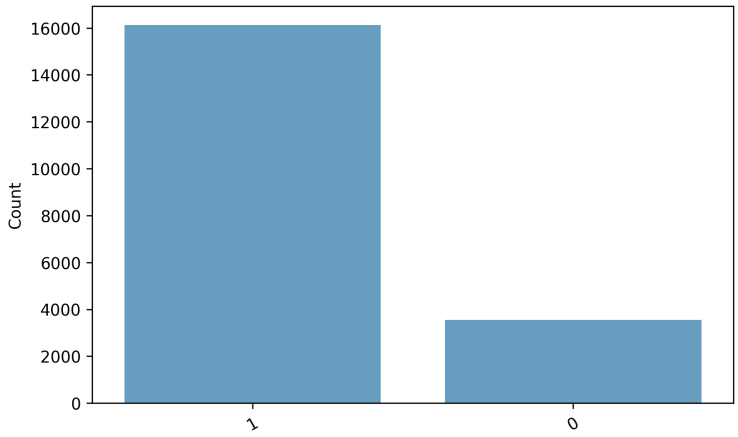

Figure 2. target distribution

Categorical variables highlight the main socio-demographic and professional characteristics of the applicants. Most applicants are married, live in a house, and are property owners, while mobile phone ownership is nearly universal and landline access is moderate. Professionally, the majority are employed under permanent contracts in the private sector, with retirees forming a substantial minority. Co-applicant variables show high proportions of missing data. Preferred communication channels are mainly notoriety campaigns, press, and television. These distributions provide an overview of applicant profiles, emphasizing common household, professional, and contact characteristics.

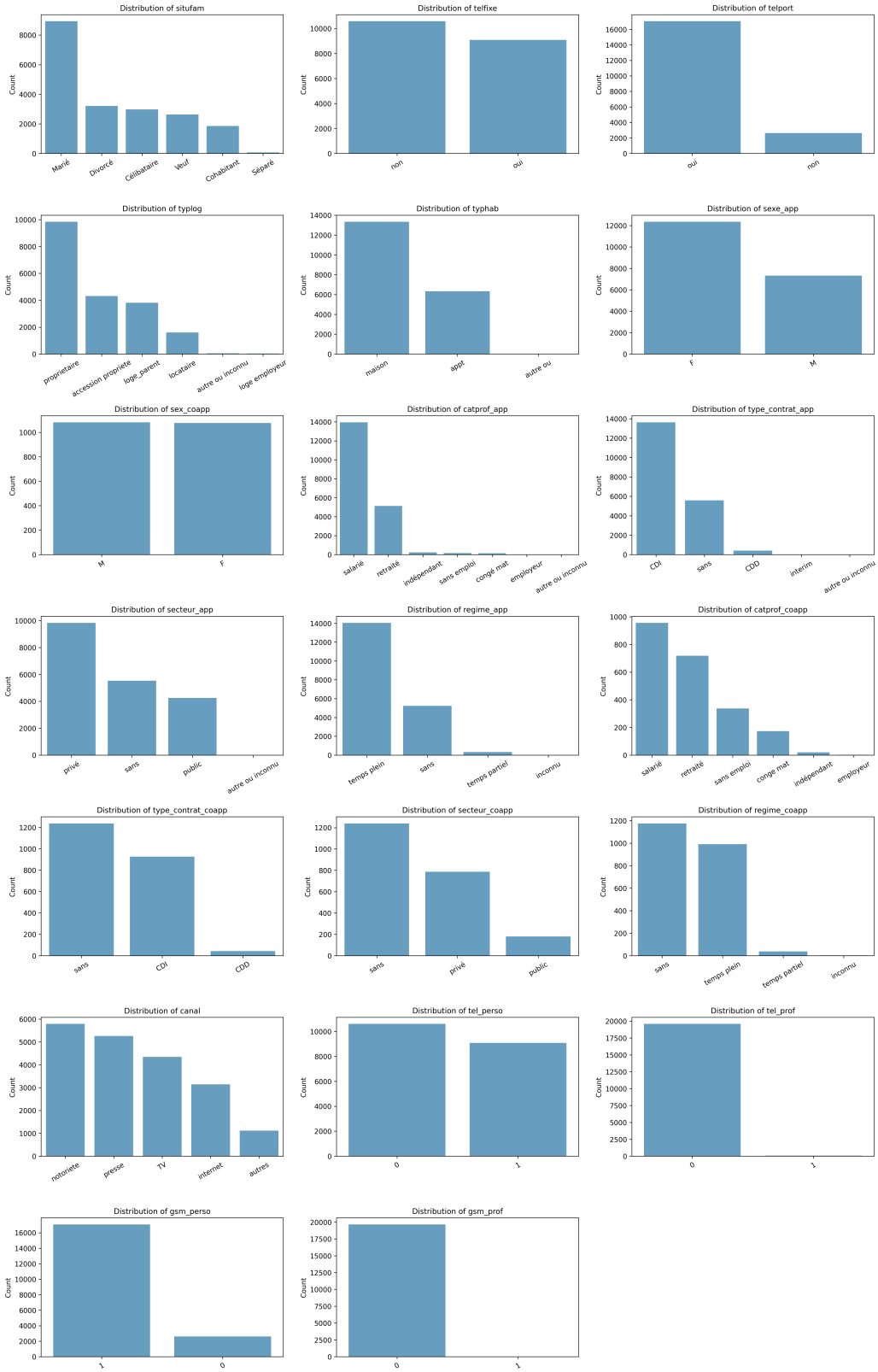

Figure 3. Categorical variables distributions

4.2. Feature Engineering and Data Preprocessing

Feature engineering is a critical phase in credit risk modeling, as the quality of input representations directly dictates the discriminative power of the ensemble architecture. To ensure a robust training set and enhance the predictive signals captured by the base learners, we implemented a multi-stage preprocessing and engineering pipeline designed to convert raw administrative records into economically significant variables.

Data Cleaning and Dimensionality Reduction: In the initial stage, features exhibiting a missing value ratio $\geq 85\%$ were systematically excluded from the feature space. This decision primarily affected co-borrower-related variables, which were sparsely populated and lacked sufficient statistical relevance for the majority of the population. The removal of these high-sparsity features is essential to reduce model noise and prevent the risk of overfitting on uninformative patterns. Additionally, temporal variables, specifically *date_acc* (credit acceptance date) and *date_adr* (last address update), were standardized into a uniform datetime format to facilitate the calculation of longitudinal stability metrics.

Engineered Predictive Indicators: To enhance the model's capacity to capture nuanced applicant behavior and long-term solvency, specialized features were derived. While raw financial data provides a static snapshot of wealth, credit risk is fundamentally a function of stability and financial margin. For clarity in the following formal definitions, English descriptors are utilized; their corresponding technical variable names from the original dataset are detailed for reference in Table 1.

The first key indicator, **Residential Stability** (T_{stab}), was quantified by calculating the temporal difference (measured in days) between the application date (*date_acc*) and the most recent address update (*date_adr*). In the context of credit scoring, T_{stab} serves as a reliable proxy for social stability, where a longer tenure at a single residence is historically correlated with more consistent repayment behavior and lower default probabilities.

Furthermore, we synthesized the granular and often volatile financial streams into two aggregate indicators: **Total Income** (I_{tot}), which consolidates primary salaries, pensions, and various secondary allowances, and **Total Expenses** (E_{tot}), which aggregates fixed monthly costs such as rent, mortgage payments, and other outstanding loan obligations. These consolidated metrics were finally integrated into a global **Solvency Ratio**, or Debt-to-Income (DTI) ratio, formally defined as:

$$DTI = \frac{E_{total}}{I_{total}} = \frac{\sum \text{Expenses}}{\sum \text{Incomes}} \tag{4}$$

The DTI ratio provides a normalized and dimensionless measure of the applicant's repayment capacity. By focusing on the relative weight of debt rather than absolute currency values, this feature allows the meta-learner to effectively compare creditworthiness across disparate professional categories and income brackets, thereby neutralizing the scale effects inherent in absolute financial figures.

4.3. Outlier treatment

Anomaly detection was performed using the Isolation Forest algorithm to identify and remove abnormal observations within each target class separately. Specifically, the model was trained on subsets of the data corresponding to each class of the target variable (target = 0 and target = 1), using different contamination rates (0.1 and 0.2, respectively). This approach is justified because the dataset is unbalanced, as noted earlier. Furthermore, a variability analysis and statistical tests (e.g., variance comparison) indicated that class 1 exhibits slightly higher dispersion than class 0, supporting the use of a higher contamination rate for that class. Instances predicted as anomalies were considered

outliers and subsequently removed from the dataset. This procedure improves data quality and enhances model robustness by eliminating extreme or inconsistent samples.

Algorithm 1: Class-Conditional Outlier Filtering

In : $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$; Contamination $\alpha_y \in \{0.1, 0.2\}$; $N_{est} = 250$
Out : Filtered Dataset $\mathcal{D}^*$

$\mathcal{D}^* \leftarrow \emptyset$
foreach $t \in \{0, 1\}$ **do**
  $\mathcal{S}_t \leftarrow \{\mathbf{x}_i \mid (\mathbf{x}_i, y_i) \in \mathcal{D}, y_i = t\}$
  Train $\mathcal{M}_t \leftarrow \text{IForest}(\mathcal{S}_t, \alpha_t, N_{est})$
  $\mathcal{L}_t \leftarrow \{\mathbf{x}_i \in \mathcal{S}_t \mid \mathcal{M}_t.\text{predict}(\mathbf{x}_i) \neq -1\}$ `// Keep inliers`
  $\mathcal{D}^* \leftarrow \mathcal{D}^* \cup \{(\mathbf{x}, t) \mid \mathbf{x} \in \mathcal{L}_t\}$
end
return $\mathcal{D}^*$

5. The Proposed DAF-Stacking Methodology

Our approach is founded upon an extended "Stacked Generalization" framework, specifically engineered to mitigate the calibration deficiencies often encountered in conventional boosting models within imbalanced data environments. Diverging from standard stacking architectures that rely on linear meta-models, we introduce a *Dynamic Feature Augmentation* (DAF) mechanism. This component maps the probabilistic outputs of a heterogeneous ensemble $\mathcal{H} = \{h_{XGB}, h_{RF}, h_{ET}, h_{LR}, h_{kNN}\}$ into a manifold of epistemic uncertainty descriptors, which subsequently supervise a meta-learning XGBoost classifier ($\mathcal{L}_{XGB}$) under rigorous regularization constraints.

5.1. Architectural Formalization and Leakage Prevention

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ represent the banking dataset, where $\mathbf{x}_i$ denotes the financial feature vector and $y_i \in \{0, 1\}$ indicates credit default. To maintain the integrity of performance estimates, we enforce a $K = 5$ stratified fold partitioning protocol. This structure facilitates the generation of "Out-of-Fold" (OOF) meta-features, ensuring that the meta-learner is exclusively trained on probabilities derived from samples unseen by the base estimators during their respective training phases.

Class imbalance is addressed strictly within each training fold's localized scope. We employ the Synthetic Minority Over-sampling Technique (SMOTE) with a fixed over-sampling ratio $\alpha = 0.8$, initiated only after isolating the validation fold. This isolation is paramount; it precludes any *synthetic contamination* that would otherwise introduce an optimistic bias in the assessment of the model's generalization capabilities. Simultaneously, standardization parameters are computed solely on the training subset $\mathcal{D}_{tr}^{(k)}$ and projected onto the validation subset $\mathcal{D}_{val}^{(k)}$, preserving total informational airtightness.

5.2. Dynamic Feature Augmentation (DAF)

The central innovation lies in the DAF operator, which projects the probability vectors $\hat{p}_{i,m}$ into a higher-order interaction space. For each observation i, we extract three key indicators of convergence and divergence:

$$V_{div,i} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{p}_{i,m} - \bar{p}_i)^2}, \quad V_{conf,i} = \max_m(\hat{p}_{i,m}) - \min_m(\hat{p}_{i,m}), \quad V_{cons,i} = \sum_{m=1}^{M}\mathbb{I}(\hat{p}_{i,m} \geq 0.5) \tag{5}$$

Here, V_{div} captures the variance in judgment across the model committee, while V_{conf} and V_{cons} quantify the magnitude of disagreement and the classification consensus, respectively. The resulting meta-input vector,

$\mathbf{Z}_i = [\hat{p}_{i,1}, \ldots, \hat{p}_{i,M}, V_{div,i}, V_{conf,i}, V_{cons,i}]$, enables the second-level learner to adaptively weigh decisions based on the perceived reliability of the primary level's signals.

5.3. Meta-Learning and Regularization Strategy

For the Level-2 stage, we have deployed a specialized XGBoost instance ($\mathcal{L}_{XGB}$) configured with conservative hyper-parameters (maximum depth $d = 3$, $\lambda = 1.0$). This technical choice is motivated by the requirement to capture non-linear interactions between base learners while preventing over-fitting on the augmented meta-feature space $\mathbf{Z}$. Optimization is conducted via the Log-Loss objective function to ensure a refined calibration of final probabilities an essential requirement for risk-gradient differentiation in production environments.

Algorithm 2: DAF-Stacking Protocol with Meta-XGBoost and Local SMOTE

Input : Dataset $\mathcal{D}$, Base learners $\mathcal{H}$, Meta-learner $\mathcal{L}_{XGB}$, SMOTE ratio $\alpha = 0.8$
Output : Optimized DAF-Stacking model H^*

`// I. Stratified OOF Generation`
Partition $\mathcal{D}$ into $K = 5$ stratified folds $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$
for $k \leftarrow 1$ **to** K **do**
  $\mathcal{D}_{tr} \leftarrow \mathcal{D} \setminus \mathcal{D}_k$; $\mathcal{D}_{val} \leftarrow \mathcal{D}_k$
  Compute scaling factors on $\mathcal{D}_{tr}$ and transform $\{\mathcal{D}_{tr}, \mathcal{D}_{val}\}$
  $\mathcal{D}'_{tr} \leftarrow \text{SMOTE}(\mathcal{D}_{tr}, \alpha)$
  for *each* $h_m \in \mathcal{H}$ **do**
    Train h_m on $\mathcal{D}'_{tr}$ using grid-search parameters
    Generate OOF predictions $\hat{p}_{i,m}$ for all $i \in \mathcal{D}_{val}$
  end
end
`// II. Meta-Feature Engineering & Training`
Compute DAF indicators (V_{div}, V_{conf}, V_{cons})
Construct $\mathbf{Z}$ and train $\mathcal{L}_{XGB}$ ($max\_depth = 3$, $\lambda = 1.0$)
`// III. Final Global Refit`
Refit all $h_m \in \mathcal{H}$ on the full resampled dataset $\mathcal{D}'$ for deployment
return $H^*(\mathbf{x}) = \mathcal{L}_{XGB}(\mathbf{Z})$

5.4. Implementation and Hyperparameter Tuning

To ensure empirical validity, hyperparameter optimization was conducted through an exhaustive Grid Search within the stratified 5-fold cross-validation framework. The final optimized configurations are detailed in Table 3.

Table 3. Optimized hyperparameters of base learners and the meta-learner.

Model	Configuration	Model	Configuration
XGBoost (Base)	η : 0.05, n_{est} : 300 Max depth: 5, Subsample: 0.8	**Logistic Reg.**	Penalty: L2 Reg. (C): 1.0
Random Forest	n_{est} : 500 Max depth: 12	**k-NN**	Neighbors (k): 7
Extra Trees	n_{est} : 300 Criterion: 'Gini'	**SMOTE**	Sampling (α): 0.8
Meta-XGBoost	η : 0.01, Max depth: 3, α : 0.1, λ : 1.0		

6. Results and Discussion

The empirical performance of the proposed DAF-Stacking framework is summarized in Table 4, where it is rigorously evaluated against heterogeneous base learners using 95% Confidence Intervals (CI), Bootstrap AUC p-values, and McNemar's statistical significance test. The experimental results demonstrate that the **DAF-Stacking model (STACK)** achieves a superior AUC of **0.8801** and an optimized **LogLoss of 0.3055**, outperforming the strongest standalone estimator, XGBoost (AUC 0.8779, LogLoss 0.3122). While the improvement in discriminative power over XGBoost is incremental, the reduction in LogLoss is of significant scientific interest. This gain indicates a substantial enhancement in **probabilistic calibration**, ensuring that the predicted default probabilities are more representative of true empirical risk frequencies a critical factor for capital allocation and risk-based pricing in financial institutions.

Table 4. Consolidated Performance Metrics and Statistical Validation (STACK vs. Baselines).

Model	AUC [95% CI]	pAUC	LogLoss	Acc.	F1-Minority	McNemar p
XGBoost	0.8779 [0.861, 0.891]	0.0426	0.3122	**0.8742**	0.9241	0.838
Random Forest	0.8621 [0.844, 0.877]	0.0350	0.3443	0.8573	0.9129	< 0.001
Logistic Reg.	0.8626 [0.845, 0.877]	0.0385	0.4162	0.8267	0.8886	< 0.001
Extra Trees	0.8616 [0.845, 0.877]	0.0406	0.3871	0.8392	0.8991	< 0.001
k-NN	0.8061 [0.787, 0.824]	0.0279	0.6897	0.7413	0.8262	< 0.001
DAF-Stacking	**0.8801 [0.864, 0.894]**	**0.0424**	**0.3055**	0.8733	**0.9242**	—

The structural superiority of this framework is fundamentally rooted in the **Dynamic Feature Augmentation (DAF)** operator. By mapping inter-model divergence and consensus into the meta-feature space, the meta-XGBoost learner effectively reconciles **orthogonal predictive signals** from the base ensemble. For instance, the inclusion of k-NN despite its lower individual AUC (0.8061) provides a distance-based perspective that contrasts with the recursive partitioning of tree ensembles. The DAF operator identifies instances of high model disagreement, allowing the meta-learner to regularize the final decision and minimize localized overfitting.

Statistical validation further supports the proposed architecture. McNemar's test confirms that STACK significantly outperforms Random Forest ($p < 0.001$), Extra Trees ($p < 0.001$), and k-NN ($p < 0.001$). Although the classification shift relative to XGBoost is not statistically significant ($p = 0.838$), the consistent information-theoretic gain (LogLoss) and a competitive partial AUC (pAUC) in the low false-positive region (FPR < 0.1) indicate a more **refined decision manifold**. This suggests that the DAF-Stacking model is particularly adept at assigning more reliable risk scores to "borderline" cases. From an operational perspective, this increased granularity in probability estimation allows financial institutions to implement more robust risk management strategies, bridging the gap between raw predictive performance and financial stability.

7. Conclusion

In this work, we have demonstrated that the DAF-Stacking framework offers a significant advancement in the reliability of credit risk modeling, primarily through the strategic synthesis of heterogeneous learners and dynamic feature enrichment. By merging the distinct inductive biases of tree-based ensembles (XGBoost, Random Forest, Extra Trees) and distance-based estimators (k-NN), our architecture effectively addresses the precision-calibration dilemma that often plagues imbalanced financial datasets. A key technical feature of this approach is the dual deployment of XGBoost, which serves both as a robust base learner and as a sophisticated meta-learner optimized with distinct hyperparameter configurations. The tangible performance gains evidenced by a 3.3% reduction in LogLoss suggest that encoding model disagreement via the DAF mechanism is a viable path toward quantifying epistemic uncertainty. For banking institutions, this improved calibration translates into a more precise differentiation of risk gradients, directly informing more resilient capital allocation strategies.

However, no model is without its constraints. We recognize that our current analysis is based on a static snapshot of credit data, which may not account for the high-velocity temporal shifts characteristic of modern economic

cycles. Furthermore, we acknowledge the inherent "black-box" tension within the DAF-Stacking manifold; the very complexity that drives its predictive power can also obscure the path to direct interpretability, a factor that remains sensitive in highly regulated lending environments.

Looking ahead, the evolution of this framework will follow two critical paths. Our immediate priority is to integrate post-hoc Explainable AI (XAI) layers, specifically SHAP-based attribution, to decompose the meta-learner's decision logic into human-readable justifications. Beyond interpretability, we intend to stress-test the DAF-Stacking framework against synthetic macroeconomic shocks. By simulating liquidity crises or inflationary pressures, we can assess how the model's risk manifold responds to systemic volatility. Ultimately, these developments aim to solidify the bridge between theoretical meta-learning and the practical demands of high-stakes financial risk management.

REFERENCES

1. O. A. Bello, *Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis*, International Journal of Management and Technology, vol. 10, no. 1, pp. 109–133, 2023.
2. Z. U. Rehman, et al., *Impact of risk management strategies on credit risk*, Financial Innovation, vol. 5, article 44, pp. 1–27, 2019.
3. A. Chopra, and P. Bhilare, *Business Perspectives and Research: Loan Lending and Credit Risk Evaluation*, vol. 6, no. 2, pp. 132–150, 2018.
4. T. Chen, and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, In Proceedings of the 22nd ACM SIGKDD, pp. 785–794, 2016.
5. S. K. Pandey, *A Review of Ensemble Machine Learning Techniques for Software Defect Prediction*, International Journal of Software Engineering, vol. 13, no. 1, pp. 1–18, 2021.
6. A. Faris, and M. Elhachloufi, *Artificial Intelligence and Machine Learning Models for Credit Risk Prediction in Morocco*, Statistics, Optimization and Information Computing, vol. 14, no. 4, pp. 1716–1740, 2025.
7. M. M. Seliem, A. A. El-Sawy, and M. I. Abd-Elmagid, *Performance of Machine Learning Algorithms for Credit Risk Prediction with Feature Selection*, Statistics, Optimization and Information Computing, vol. 14, no. 3, pp. 311–328, 2025.
8. Z. Hjouji, I. Hasinat, and A. Hjouji, *A New Method in Machine Learning Adapted for Credit Risk Prediction of Bank Loans*, Statistics, Optimization and Information Computing, vol. 11, article 1476, pp. 1–15, 2023.
9. D. H. Wolpert, *Stacked Generalization*, Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
10. L. Breiman, *Stacked Regressions*, Machine Learning, vol. 24, no. 1, pp. 49–64, 1996.
11. K. M. Ting, and I. H. Witten, *Issues in Stacked Generalization*, Journal of Artificial Intelligence Research, vol. 10, pp. 271–289, 1999.
12. T. Ayobami, *Prediction of Loan Default in Diverse Markets*, Journal of Financial Services Research, vol. 12, no. 2, pp. 88–104, 2023.
13. A. Idhmad, et al., *Scoring of Borrowers Solvability by SVM and MLP Hybridized to GA*, International Journal of Innovative Soft Computing and Engineering (IJISAE), vol. 11, no. 2, pp. 254–261, 2023.
14. M. Anand, A. Velu, and P. Whig, *Prediction of Loan Behaviour with Machine Learning Models for Secure Banking*, Journal of Computer Science and Engineering (JCSE), vol. 3, no. 1, pp. 1–13, 2022.
15. X. Zhu, et al., *Explainable prediction of loan default based on machine learning models*, Data Science and Management, vol. 6, no. 3, pp. 133–142, 2023.
16. A. Tyagi, and S. Goyal, *Optimized Financial Decision-Making using Deep Learning*, Journal of Information and Computational Science (JIC), vol. 15, no. 8, pp. 112–125, 2023.