Received: 19 June 2024 | Revised: 9 December 2024 | Accepted: 14 February 2025 | Published online: 3 March 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 3(3) 284–294
DOI: 10.47852/bonviewAIA52023655

BON VIEW PUBLISHING

Credit Default Prediction Using Time Series-Based Machine Learning Models

Yujuan Qiu[1,*] and Jianxiong Wang[2]

[1]*School of Engineering and Applied Science, The George Washington University, USA*

[2]*McIntire School of Commerce, University of Virginia, USA*

Abstract: Credit card defaults are among the most significant risks in the financial world, with the potential to negatively impact the overall financial health of the entire economy. Enhancing the accuracy of predicting and identifying credit defaults is essential in mitigating credit losses and minimizing financial risks in credit risk management. This research specifically focuses on the prediction of credit card defaults by comparing various traditional machine learning models. More importantly, it proposes a novel hybrid framework that integrates convolutional neural networks, long short-term memory, and attention mechanisms. By incorporating time-series components into our hybrid model, we achieved a notable improvement in predictive accuracy, outperforming the best traditional model by 16%. This study highlights the significant benefits of integrating temporal sequences into credit risk models, as it can greatly enhance the precision, reliability, and overall performance of credit card default predictions, offering important advantages for improving long-term financial stability and reducing associated risks.

Keywords: credit default prediction, time-series model, machine learning, convolutional neural network, long and short-term memory

1. Introduction

The significant impact of credit card defaults on financial institutions and the overall economy is apparent. Forecasting and identifying potential defaulters in credit cards is a key, almost gatekeeping, step in the process of credit risk management. If we allow a forecast and/or a potential defaulter identification to fail, it will directly translate into a credit loss event to occur. Credit card lending has been part and parcel of modern life for almost a century. The extensive credit defaults that were seen during the 1997 Asian Financial Crisis demonstrate the crisis's far-reaching worldwide economic effects [1–3]. This case highlights the worldwide economic instability and individual lenders that are the fallout from a lending failure.

In the past, projections for credit defaults relied on traditional scoring systems like FICO, which brought together a number of indicators that reflected a person's creditworthiness and credit history. In recent years, however, we've seen the emergence of big data and machine learning (ML), which have taken this field to new heights, allowing us to identify more complex, non-linear patterns within our transaction data. We have seen new and improved ways to use these patterns to deploy ML models from the basic Logistic Regression all the way up to advanced techniques like Neural Networks and ensemble methods such as Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM). Studies by researchers such as Chou and Lo [4] as well as Kim et al. [5] have demonstrated the effectiveness of these models in improving prediction accuracy. Recent breakthroughs in hybrid ML techniques for credit default forecasts have shown substantial improvement in accuracy and interpretability. The model known as the Hybrid Algorithms Multi-Stage combines unsupervised and supervised learning, resulting in an impressive prediction of loss given default. The model achieves this improved result across a diverse set of datasets [6]. Additionally, hybrid quantum neural networks have emerged, leveraging quantum computing to achieve a remarkable prediction accuracy [7]. Furthermore, Bayesian networks, which facilitates casual analysis and what-if scenarios, have been used to make more informed decision and enhance interpretability in credit risk predictions [8]. Credit default prediction is evolving, and it is doing so not just with more standard methods like Random Forest (RF) and Artificial Neural Networks but also with hybrid approaches that combine different techniques to get better results and insights [9, 10].

However, a critical gap remains in the thorough assessment of model performance across different models, especially when it comes to the non-linearity and data imbalance problems that are common in credit default datasets. This is a difficult problem to solve, and it is a problem we must solve if we are to create improved predictive models and tools to forecasting credit defaults [11–14]. Moreover, time-series analysis combined with ML models has been underexplored to consider account statements as part of interconnected sequences, which represents an opportunity to innovate in the field of credit default prediction [15].

This research addressed this gap by presenting two results. The first is to compare model performance across conventional ML models with an in-depth evaluation of hyperparameter tweaking. The second is to present a novel hybrid LSTM-CNN-Attention architecture and compare it with the best performer from

*Corresponding author: Yujuan Qiu, School of Engineering and Applied Science, The George Washington University, USA. Email: yqiu59@gwu.edu

traditional models. We aim to overcome the constraints of past research and present new insights by providing a comprehensive evaluation of different models, including our novel hybrid LSTM-CNN-Attention framework.

The remainder of the paper is organized as follows: Section 2 presents the ML approach developed in this research. Section 3 introduces the datasets to evaluate the performance of the approach. Section 4 presents and discusses the experimental results. Section 5 summarizes the new findings of this research.

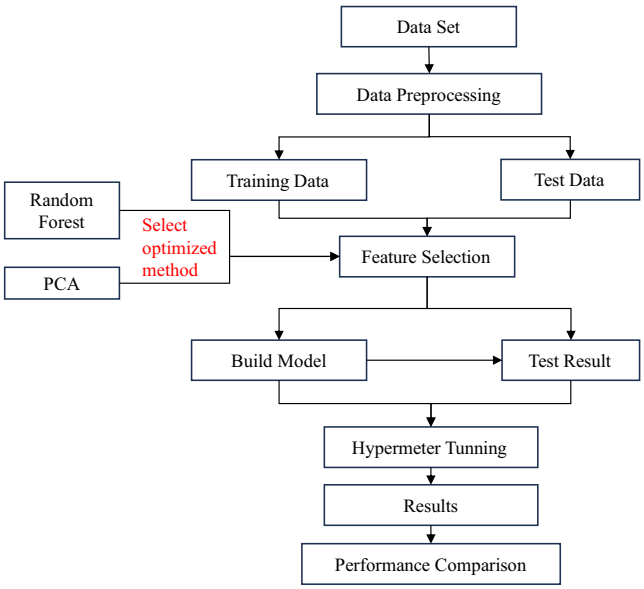
2. Methodology

2.1. Overview

In our comprehensive analytical framework designed for credit card default forecasting, our analytical approach follows a standardized framework developed to assess and compare the predictive performance of diverse ML models in the context of credit card default forecasting, as shown in Figure 1. Initial preprocessing of the dataset is conducted to ensure data relevance and quality. The dataset is then partitioned into training and testing subsets to facilitate model development and objective evaluation. Our feature selection process also diverges from conventional methods, employing a comparative analysis between two distinct techniques to identify the most predictive features: principle component analysis (PCA) [16] and RF [17]. This precision in feature selection is complemented by hyperparameter tuning, a critical step in refining model performance, particularly vital for addressing the challenges posed by imbalanced datasets prevalent in credit default scenarios.

Building on this basis, we provide a new hybrid ML framework: the Hybrid CNN-LSTM-Attention Framework. This novel approach combines the strengths of convolutional neural networks (CNNs) [18] for feature extraction, long short-term memory networks (LSTMs) [19] for capturing temporal dependencies, and an Attention mechanism [20] that concentrates especially on important features of the data. The Attention layer dynamically prioritizes the most important features, the CNN layers find noteworthy patterns in the data, and the LSTM components evaluate these patterns in the context of credit behavior over time. This ensures that the predictions are accurate and informed [15]. This hybrid model adds multi-dimensional data to predicting credit defaults while broadening our research. By contrasting traditional models with our proposed Hybrid CNN-LSTM-Attention Framework, we assess method efficacy and show that the predictive accuracy of the models has improved [4]. In addition, we compared the fundamental methodology of the proposed hybrid model with a broad spectrum of traditional ML models, including XGBoost [21], LightGBM [22], Neural Networks [18], and Logistic Regression [23]. This comprehensive method highlights the benefits of utilizing the benefit of deep learning to recognize patterns in the intricate pattern of financial data, creating a new standard for accuracy and insight in credit risk evaluation.

Figure 1
Research architecture workflow for traditional machine learning models



2.2. Feature selection

Feature selection is crucial for refining accuracy and interpretability of ML models, especially related to credit default predictors [15]. This study includes a comparative analysis of PCA [16] and a RF-based approach [17], demonstrating their advantages in handling linear and non-linear data complexities [3]. The principle behind PCA is a linear transformation. PCA identifies orthogonal axes in the feature space to maximize variance and then projects the original data onto these axes. Alternately, the RF algorithm uses ensemble learning to evaluate features across many decision trees. This algorithm can handle both linear and non-linear relationships well and gives a feature importance ranking as output [4].

2.3. Proposed hybrid ML framework

To provide a thorough study of credit default risk, the suggested novel hybrid ML technique combines the advantages of CNNs, LSTMs, and Attention processes. The structural and conceptual framework of this hybrid model is substantially different from that of traditional methods. By using time-series analysis to fully use the sequential character of financial data, our Hybrid CNN-LSTM-Attention Framework also fills a significant research gap in the field of credit default prediction. This novel method sees a customer's financial history as a linked sequence, reflecting the temporal evolution of credit activity, in contrast to existing models that examine individual data points in isolation [14].

Traditional ML models, such as XGBoost, treat each customer data—corresponding to individual statements—as isolated, ignoring the temporal sequence of credit behavior. In contrast, our hybrid model integrates time-series data to acknowledge the sequential relationships inherent in financial activities, treating each customer history as a continuous narrative rather than a disjointed collection of events. This methodology not only aligns with the inherent structure of financial datasets but also enhances predictive accuracy by recognizing the importance of historical context in forecasting defaults.

The hybrid model provides a thorough examination of credit risk by combining CNNs for the extraction of spatial characteristics, LSTMs for the analysis of temporal sequences, and a unique Attention mechanism for concentrating on the most important information [12]. Because of this fusion, the model can identify subtle patterns and connections in the data that traditional models can miss due to their lack of temporal attention. Because it carefully highlights important qualities and situations that have a substantial influence on a customer's likelihood of default, the Attention layer plays a particularly important role in making a targeted and accurate prediction [14].

This hybrid model is unique because it implements a comprehensive approach that brings together temporal sequence analysis and spatial pattern recognition—pattern recognition in space—augmented by the Attention mechanism's ability to provide strategic focus. This type of framework for analysis is new in the financial modeling for credit risk. Thus, the Hybrid CNN-LSTM-Attention Framework sets a new benchmark in this area by providing a novel, more accurate, and interpretable approach to assessing the risk of credit default that could change the way risk is managed in the financial services sector.

2.3.1. CNN layer

The CNN layer employs convolution operations to extract high-level features from the input data [18]. For a given input X, a convolution operation applies a filter Wof size k, producing a feature map F:

$$F(i) = \mathrm{ReLU}\left(\sum_{j=1}^{k} X(i+j)\right).W(j) \tag{1}$$

where ReLU is the Rectified Linear Unit activation function, enhancing non-linearity. In the context of our model, the CNN layer analyzes patterns across features within each timestep, identifying key indicators that might signify credit default risk.

CNNs therefore are adept at extracting hierarchical patterns within data through convolutional layers, making them ideal for identifying salient features within our complex financial datasets.

2.3.2. LSTM layer

A recurrent neural network, LSTM, is good at capturing temporal dependencies and sequences. This is crucial when we are attempting to understand the credit behavior of people over time —to understand the "little things" that happen in chronological order that add up to a certain type of credit behavior [19]. LSTMs take in data one piece at a time and remember the right details for the right amount of time. The key to the success of LSTM is its specialized structure, which has several gates to regulate the information flow. An LSTM unit updates its cell state C_t and hidden state h_t at each timestep t using the following formula, where σ is the sigmoid function, and tanh is the hyperbolic tangent function, providing non-linearity:

1) Forget Gate: Determines which information from the cell state should be discarded or retained, allowing the model to forget irrelevant details from the past.

$$f_t = \sigma\left(w_f .[h_{t-1}, x_t] + b_f\right) \tag{2}$$

2) Input Gate: Decides which added information from the current input should be added to the cell state, enabling the model to update its memory with relevant new observations.

$$i_t = \sigma\left(w_i .[h_{t-1}, x_t] + b_i\right) \tag{3}$$

$$\tilde{C}_t = \tanh(w_c \cdot [h_{t-1}, x_t] + b_c) \tag{4}$$

3) Cell State Update: Combines the old cell state and added information to form the current cell state.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

4) Output Gate: Determines what part of the cell state should be output at the current timestep, influencing the final prediction.

$$O_t = \sigma\left(w_o .[h_{t-1}, x_t] + b_o\right) \tag{6}$$

$$h_t = O_t * \tanh(c_t) \tag{7}$$

Here, x_t is the input at timestep t, h_{t-1} is the previous hidden state, and c_{t-1} is the previous cell state. The weights W and biases b are re parameters learned during training, and $*$ denote element-wise multiplication. This architecture allows LSTMs to mitigate the vanishing gradient problem common in traditional RNNs, making them capable of learning dependencies over many timesteps. In the context of credit default prediction, the ability of the LSTM to maintain and update the cell state c_t over time allows it to remember significant information and forget the irrelevant, capturing the temporal dependencies in a customer's credit history [5]. For instance, patterns of late payments or sudden increases in spending can be crucial indicators of default risk, which the LSTM layer is adept at identifying and learning from leveraging its sophisticated gating mechanisms. This capability makes the LSTM layer a core component of our hybrid model, enabling a dynamic and temporally aware analysis of credit behavior that traditional models, which treat each data point independently, cannot provide.

2.3.3. Attention layer

By concentrating on the most significant characteristics and time steps, the Attention mechanism improves the interpretability of this hybrid model [20]. It is an essential element that improves the CNN-LSTM architecture's interpretative power. The Attention mechanism concentrates on particular sequential data segments that are more imperative to credit default prediction by assessing the significance of various timesteps in the LSTM output. The model may devote more computing resources to examining the critical periods of a customer's financial history for precise default prediction because of this selective focus.

Our model utilizes a custom Attention mechanism, which computes a context vector that captures relevant information across all timesteps. The unique needs and subtleties of credit default prediction are what motivate our hybrid model to incorporate a proprietary Attention mechanism [11]. By creating a custom Attention layer, we ensure optimal alignment with the given goal by customizing the focus mechanism to the distinct temporal and feature-related complexity of the underlying credit default data. This customization improves the model's overall efficacy and interpretability in the context of credit risk

assessment by making it easier to important events and trends in a customer's financial history that generic attention models might miss.

The attention scores are computed as follows:

$$e_t = \tanh(W * h_t + b) \tag{8}$$

$$a_t = \frac{\exp(e_t)}{\sum_{t=1}^{T}(a_{t^*} h_t)} \tag{9}$$

$$c = \sum_{t=1}^{T} a_t * h_t \tag{10}$$

where $\mathbb{W}$ and b are learnable parameters of the attention layer, h_t is the hidden state from the LSTM at timestep t, e_t is the energy associated with each timestep, and c is the resulting context vector, focusing the model attention on the most informative parts of the sequence. This mechanism allows the model to prioritize specific periods in a customer's financial history that are more indicative of potential default, enhancing predictive accuracy.

2.4. Hypermeter tuning

Hyperparameter tuning is a critical step in refining traditional ML models for credit card default prediction, a domain where ensemble models like RFs and Gradient Boosting Machines (e.g., XGBoost) excel due to their robust predictive capabilities [24]. This process involves adjusting non-learnable model settings, such as learning rates and tree depths, to optimize performance [11]. Unlike the application of hyperparameter tuning in our hybrid CNN-LSTM-Attention Framework, where the focus is on integrating time-series data to capture sequential relationships, traditional models primarily benefit from this tuning in their standalone application.

Since hyperparameter adjustment is essential for improving model accuracy and generalizability, we focus on it in our research for both traditional and hybrid models. In financial environments, where decision-making may be greatly impacted by the accuracy of default projections, this topic is especially important. Our technique, leveraging on the grid search, investigates a wide variety of hyperparameters, differentiating our work from research that could depend on default settings or scant parameter investigation.

The different combinations of hyperparameters are examined by a grid search, which extends to the full picture of the optimization landscape. The method guarantees that each model is tuned to a finely wrought setting, which in turn guarantees better predictive power. An in-depth tuning of standard models is a crucial part of our approach, and it prepares the way for a comparison with our ensemble hybrid model. The demonstration of the power of hyperparameter tuning with standard credit risk models is an original contribution of our research, which tends to get overlooked in this field [3].

By fine-tuning hyperparameters to make models work better, we ensure that their forecast performance is evaluated fairly. This gives us a solid base for comparing these models with our new hybrid framework. This optimization shows how important it is to choose the right hyperparameters to get a good model result, especially in situations where misled estimates could have big effects.

2.5. Evaluation metrics

The models are evaluated using a set of metrics. These metrics provide different and complementary views of model performance. The principal metric we use is Accuracy. In addition, we compute and consider the following metrics: Precision, Recall, $F1$-score, and the area under the precision-recall (PR) curve, and receiver operating characteristic (ROC) curve [25]. The ROC curve is a plot of true positive rate (Recall) against the false positive rate at various threshold settings. A ROC curve offers insight into the trade-offs between true positives and false positives [26]. PR curves plot Precision versus Recall and are especially useful for evaluating models on imbalanced datasets. Unlike the ROC curve, where we look for a single point corresponding to an optimal trade-off between true positive and false positive rates, the PR curve gives us two dimensions along which we can characterize a model [27]. AUC (area under the curve) summarizes the ranks, indicating a model's ability to identify the positive class in the first instance and quantifying retrieval and ranking prowess [28].

$$AP = \int_0^1 p(r)dr \tag{19}$$

Ensemble models, especially those with high average precision (AP) values, represent a more balanced trade-off between PR—something that is especially crucial for imbalanced dataset. When we apply this framework to our credit card default prediction problem with the goal of identifying the "best" model, we're not only enriching the risk management strategies of our financial institution but also offering a more sophisticated means of assessing model effectiveness. Based on our results, ensemble models seem to have the clearest handle on minimizing false positives and maximizing the true detection of credit card defaults.

3. Data Sources

3.1. Overview

This study leverages a dataset that spans 18 months of anonymized and standardized credit card account information. The dataset originates from the American Express Default Prediction challenge on Kaggle [1]. The dataset includes 41,989 unique customers. Each customer has between 2 and 13 statements, so an array of credit behavior and statement is our basis for data. Our records yield over 1 million data points tied to 191 different features. Each feature of a customer's credit activity for a given statement period is at the level of individual entries, providing a granularity of credit behavior.

3.2. Data exploration

The dataset is divided into five main areas that are directly connected to how people use credit; this makes the dataset particularly well suited for predicting credit defaults. The dataset scale, combined with the array of 191 features and 41,989 data points, offers a profound, multifaceted view of the consumer credit dynamics.

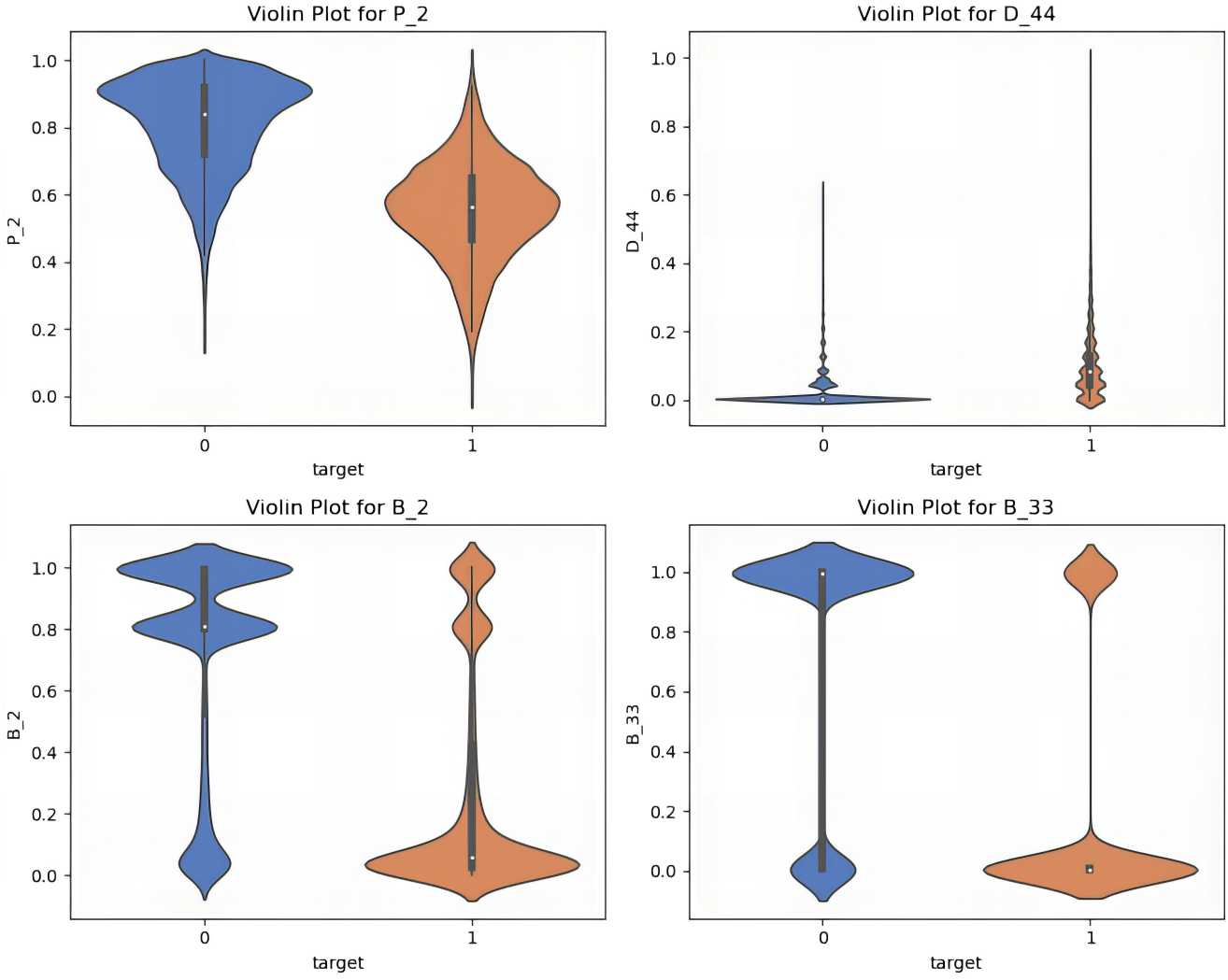
1) Delinquency Variables: Reflect late or missed payments, crucial for assessing the likelihood of future defaults.
2) Spend Variables: Capture spending patterns, signaling financial behavior that may precede default events.
3) Payment Variables: Indicate payment habits, which evaluates a customer ability to manage debt.

4) Balance Variables: Represent indebtedness, serving as direct indicators of financial health.
5) Risk Variables: Encompass broader risk factors, including economic conditions that could impact creditworthiness.

In analyzing our dataset, thorough data cleaning and preparation were crucial to ensure data quality and usability. Out of 191 features, two were categorical and required transformation into numerical values to comply with the needs of our ML algorithms. We tackled missing values by applying mode imputation for categorical features and median imputation for numerical ones, thus achieving a consistent dataset without missing entries. Feature selection was also a pivotal part of our study, aimed at boosting the model effectiveness and clarity. We explored two methods: PCA and RF. The PCA, a method for linear dimensionality reduction, operates by identifying new orthogonal axes that optimize data variance. Conversely, RF, an ensemble learning technique, integrates information from several decision trees to evaluate feature significance [5]. The RF method has proven useful in identifying variables for credit default prediction through the utilization of significance scores [1]. These scores quantify a feature contribution to the accuracy of the model by measuring its utilization across the ensemble of decision trees. Analyzing the distribution of feature data is crucial for obtaining insights into our dataset. A violin plot serves as an effective tool for this purpose, which offers a composite view that merges the attributes of a box plot with a kernel density plot. It not only showcases the median and interquartile ranges of the data but also the probability density at different values through its characteristic, mirrored density plots. Such detailed visualization facilitates a comparative analysis of data distributions across various categories, providing an informative snapshot of underlying patterns that may influence credit default outcomes.

As shown in Figure 2, P_2, D_44, B_2, and B_33 were identified as the top four features with the highest importance. Their violin plots for features P_2, D_44, B_2, and B_33 demonstrate distinct distribution patterns that have implications for credit default data. The broader spread of P_2 in category 0 and its concentration in category 1, along with the pronounced bimodality in B_2, indicate non-linear relationships with credit default, suggesting that simple linear models may not adequately

Figure 2
Violin plot for top features



capture these dynamics. D_44 and B_33, with their tight distributions, especially in category 1, point to potential data imbalance, where defaults may be less frequent than non-defaults. Therefore, the key takeaway is that while the varied feature distributions provide valuable differentiation between our target categories, predictive modeling must account for non-linearity and class imbalance. Employing advanced feature engineering, non-linear models, and strategies to balance the data will be crucial in developing a robust model for credit default prediction.

4. Result

4.1. Result overview

In our study's results, we carefully looked at how well standard ML models and our new hybrid model worked at three different levels:

1) Feature Selection Results. First, we stress how useful ensemble methods like RF are for choosing features for large, complicated financial datasets.
2) Hyperparameter Impact on Traditional Models: We look at how tweaking hyperparameters affect models like XGBoost, LightGBM, Neural Networks, and Logistic Regression. Finding out how hyperparameters affects models like XGBoost and LightGBM is very important because it helps improve model performance and gives us ideas on how to make predictions more accurate in financial datasets.
3) Performance Comparison Among Traditional Models: After we adjusted the hyperparameters, we can see which of the standard models shows best performance according to our performance metrics. Among these traditional models, XGBoost has presented to be the best performer. Its performance scores are higher than other models. This finding aligns with the fundamental algorithm of XGBoost and shows that it is relatively more sophisticated model to deal with complicated, high-dimensional datasets that are common in credit default prediction.

Hybrid Model Versus Traditional Model Performance: We aim to assess the additional benefit of including time-series analysis via CNN and LSTM layers, together with the specific emphasis offered by the Attention mechanism, by comparing the best-performing conventional model (XGBoost) with our Hybrid CNN-LSTM-Attention Framework. This comparison will clarify the effectiveness of the hybrid model in capturing temporal connections and prioritizing variables, providing improved forecast accuracy and insights into default risk factors.

4.2. Feature selection results

Our examination reveals that the RF method exhibited superior accuracy than PCA for credit card default prediction for most results, as detailed in Table 1. Interestingly, when applying RF for feature selection, Neural Network models and our advanced Hybrid CNN-LSTM-Attention framework did not exhibit the same level of enhanced performance. This discrepancy suggests that while ensemble methods like RF excel in capturing and evaluating intricate data patterns, Neural Networks, and hybrid models may require more nuanced or alternative feature selection approaches to fully capitalize on their architecture, especially given their capacity to model complex, high-dimensional data and temporal sequences [13]. Despite this, we continue to utilize RF as our feature selection method due to its overall superior results and its effectiveness in providing a clear ranking of feature importance, which simplifies the model development process by enabling a focused approach on the most predictive variables.

Table 1
PCA vs. random forest comparison of performance metrics across machine learning models

Model	Evaluation metrics	PCA	RF	Comparison
XGBoost	Accuracy	0.871	0.879	1.0%
	Precision	0.847	0.855	0.9%
	Recall	0.904	0.914	1.0%
	*F*1-score	0.875	0.883	1.0%
	AP Score	0.932	0.945	1.4%
	AUC	0.941	0.950	1.0%
LightGBM	Accuracy	0.858	0.866	0.9%
	Precision	0.832	0.838	0.8%
	Recall	0.898	0.907	1.0%
	*F*1-score	0.864	0.871	0.9%
	AP Score	0.917	0.929	1.3%
	AUC	0.929	0.938	1.0%
Neutral Network	Accuracy	0.870	0.868	−0.2%
	Precision	0.856	0.859	0.3%
	Recall	0.888	0.881	−0.8%
	*F*1-score	0.872	0.870	−0.3%
	AP Score	0.929	0.931	0.3%
	AUC	0.938	0.940	0.2%
Logistic Regression	Accuracy	0.850	0.854	0.4%
	Precision	0.840	0.840	0.0%
	Recall	0.865	0.874	1.1%
	*F*1-score	0.852	0.857	0.5%
	AP Score	0.910	0.914	0.5%
	AUC	0.923	0.927	0.4%
Proposed Hybrid Model	Accuracy	0.997	0.996	−0.1%
	Precision	0.994	0.998	0.4%
	Recall	0.992	0.993	0.1%
	*F*1-score	0.993	0.997	0.4%
	AP Score	0.992	0.994	0.2%
	AUC	0.994	0.995	0.1%

Our research underscores the use of ensemble methods like RF in feature selection for complex financial datasets. In our research, by meticulously selecting the top 47 features from an initial set of 191 based on important scores, we highlight the superiority of the ensemble method in enhancing model performance and interpretability over linear techniques like PCA. This approach reinforces the significance of methodical feature selection and model evaluation in developing sophisticated credit risk assessment tools.

4.3. Hypermeter tuning impact

Our analysis reveals that hyperparameter tuning is a significant factor in improving the predictive performance of conventional ML models in the area of credit default prediction. The models XGBoost and LightGBM demonstrate improvement in the performance metrics we discussed earlier—accuracy, precision, recall, and *F*1-score—after we performed a hyperparameter tuning of them, shown in Table 2. This result shows the benefit of hyperparameter

Table 2
Before vs. after hyperparameter tuning comparison of performance metrics across machine learning models

Model	Metrics	Before	After	Comparison
XGBoost	Accuracy	0.879	0.882	0.3%
	Precision	0.855	0.856	0.1%
	Recall	0.914	0.919	0.6%
	*F*1-score	0.883	0.887	0.4%
	AP Score	0.945	0.949	0.4%
	AUC	0.950	0.953	0.3%
LightGBM	Accuracy	0.866	0.868	0.2%
	Precision	0.838	0.841	0.3%
	Recall	0.907	0.907	0.1%
	*F*1-score	0.871	0.873	0.2%
	AP Score	0.929	0.933	0.4%
	AUC	0.938	0.940	0.2%
Neutral Network	Accuracy	0.868	0.866	−0.2%
	Precision	0.859	0.845	−1.7%
	Recall	0.881	0.898	2.0%
	*F*1-score	0.870	0.871	0.1%
	AP Score	0.931	0.937	0.7%
	AUC	0.940	0.927	−1.4%
Logistic Regression	Accuracy	0.854	0.854	0.0%
	Precision	0.840	0.840	0.0%
	Recall	0.874	0.874	0.0%
	*F*1-score	0.857	0.857	0.0%
	AP Score	0.914	0.914	0.0%
	AUC	0.927	0.927	0.0%
Hybrid Model	Accuracy	0.996	0.998	0.2%
	Precision	0.998	0.997	−0.1%
	Recall	0.993	0.999	0.6%
	*F*1-score	0.997	0.998	0.1%
	AP Score	0.994	0.998	0.4%
	AUC	0.995	0.998	0.3%

optimization in enhancing the performance of ensemble models with complex datasets where imbalanced patterns might skew prediction [5].

However, the effects of hyperparameter optimization on Neural Networks show a more complex outcome. The effectiveness of tuning seems to be contingent on the exact architecture of the network, its complexity, and the nature of the credit default data. This uncovers the delicate balance that exists between model and data, suggesting that a one-size-fits-all approach to hyperparameter tuning is probably ineffective, particularly for models like Neural Networks. By contrast, Logistic Regression, due to its simple structure, shows modest improvement from hyperparameter tuning. We expect this outcome because the model provides few options for configuring adjustments. This highlights that not all models are equally responsive to performance improvements through hyperparameter tuning.

These findings are key to selecting and fine-tuning models for predicting credit defaults. They underscore the necessity of hyperparameter tuning with a cautious use depending on the specific model and the nature of the dataset. We therefore will systematically optimize traditional models and build a framework for their comparison with the advanced hybrid framework we propose in this work.

4.4. Traditional model performance comparison

Based on our assessment, we find that traditional ML models for predicting credit defaults work better when they're ensembles like XGBoost or LightGBM. These two models outperform their counterparts and appear to be the best methods to use when dealing with an ensemble of knowledge-based, decision-tree models. XGBoost, LightGBM, and similar approaches are models that make use of a multitude of "decision trees" and arrive at predictions by averaging, or consolidating, the predictions made by all the trees. The reason these methods work better—apart from the fact that they use many trees and do not overfit when they could—is that they employ sharpening procedures to bring down the bias and pump up the forecast variability.

Figure 3 displays four ROC curves, each representing the performance of a different ML model: XGBoost, LightGBM, Neural Network, and Logistic Regression. ROC curves are graphical plots that illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied. They plot the true positive rate against the false positive rate at various threshold settings. The AUC is a measure of the ability of the classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. An AUC of 0.5 suggests no discrimination, while an AUC of 1.0 indicates perfect discrimination. In our case, XGBoost has the highest AUC, indicating it has the best performance among the four in terms of distinguishing between the positive and negative classes. Logistic Regression has the lowest AUC, which suggests it is less capable than the other models, yet still provides a good classification ability.

The superior performance of XGBoost and LightGBM is also evident in their higher AP at 0.948 and 0.933 individually shown in Table 3. This metric is crucial in the context of credit default prediction, where the cost of false positives (incorrectly predicting a default) can be significant. The higher AP score indicates that these models are more effective in ranking predictions by probability, ensuring that the highest-ranked predictions are more likely to be true defaults. This capability is particularly beneficial in imbalanced datasets common in credit default scenarios, where the number of non-default cases significantly outweighs the default cases.

The reasons why XGBoost and LightGBM work are due to a few main factors. These include their ability to deal with missing values, their use of gradient boosting to cut down on forecast mistakes, and their adaptability in model tuning, which allows different hyperparameters be optimized to make the model optimized. These models also use advanced regularization methods to keep them from overfitting, which makes sure that they work well with data they haven't seen before. The ensemble models, on the other hand, did much better than Neural Networks and Logistic Regression in our study, though they were still good. This difference might be because credit default data is naturally complicated and doesn't have simple, straight relationships. Ensemble models are better at capturing and modeling these relationships accurately.

Figure 3
ROC curve comparison

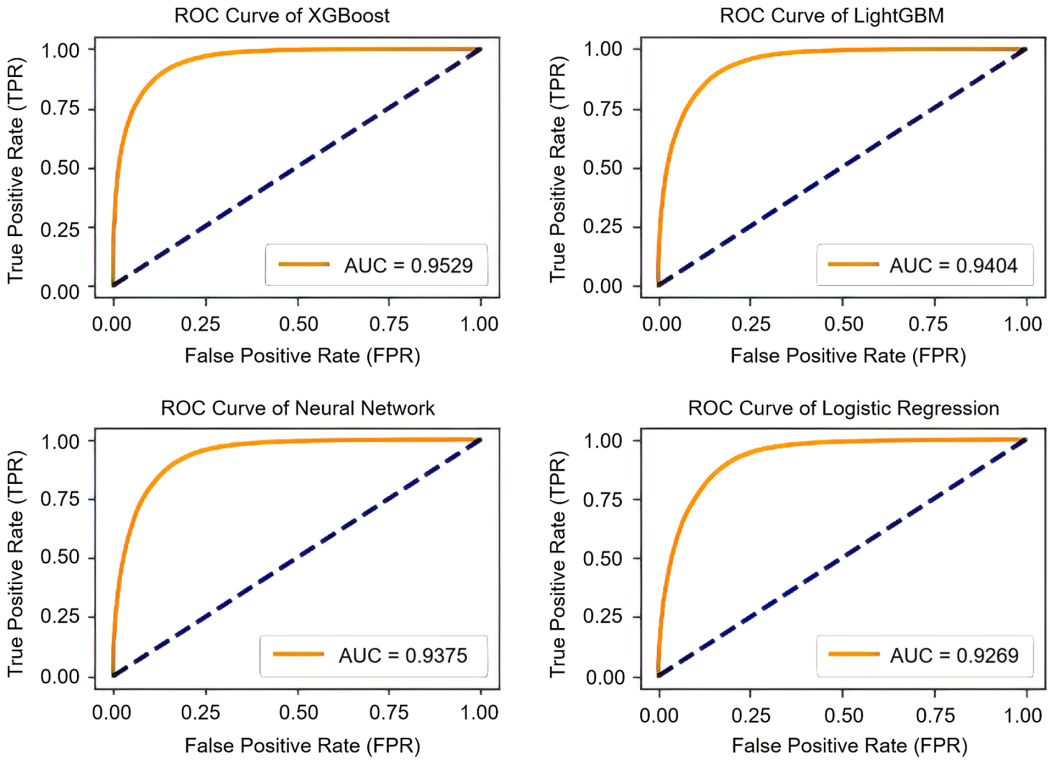


Table 3
Performance matrix comparison for traditional models

Performance metrics	XGBoost	LightGBM	Neural network	Logistic regression
Accuracy	0.8824	0.8676	0.8665	0.8538
Precision	0.8562	0.8407	0.8446	0.8402
Recall	0.9192	0.9071	0.8983	0.8739
*F*1-Score	0.8866	0.8727	0.8706	0.8567
AP Score	0.9489	0.9331	0.9375	0.9144
AUC	0.9529	0.9404	0.9269	0.9269

4.5. Hybrid model versus traditional model performance

After identifying the XGBoost model as the top performer among traditional ensemble models, our research uses this as the baseline and compares it with our Hybrid CNN-LSTM-Attention Framework. The implementation of the hybrid model demonstrated outperformed results in both accuracy and precision scores with a significant increase of 13–16% over the XGBoost benchmark, as shown in Table 4. This improvement not only highlights the hybrid model's efficacy but also underscores the marginal benefit of incorporating sequential data analysis into predictive modeling.

Hybrid model makes good use of the fact that financial actions happen over time by using CNNs to extract spatial features, LSTMs to capture temporal dependencies, and an Attention method to focus on important information. Credit risk is more accurately shown when the financial history of each customer is learned by model, rather than as a bunch of separate data pieces. The model's outperformance shows how important temporal analysis is for predicting credit failure, which adds to the novel edge of our hybrid approach for working with large, complicated financial datasets.

Table 4
Performance matrix comparison

Performance metrics	XGBoost	Hybrid model	Comparison
Accuracy	0.8824	0.9981	13%
Precision	0.8562	0.9974	16%
Recall	0.9192	0.9987	9%
*F*1-Score	0.8866	0.9982	13%
AP Score	0.9489	0.9984	5%
AUC	0.9529	0.9976	5%

This hybrid CNN-LSTM-Attention Framework is a big step forward in predicting credit failure. By looking at financial records as linked cycles, it not only improves the accuracy of predictions but also fills in an important research gap. This point of view is important for giving a more complex picture of credit risk by leveraging on the creative combination of CNNs, LSTMs, and Attention mechanisms.

However, the practical implementation of this complex model needs to consider its computing requirements. The intricacy of the hybrid model and the extensive sequential data processing requires significant computer resources, which may not be available. The selection of a model for practical implementation must balance predicted accuracy with computing limitations [15].

4.6. Limitations and future research directions

The hybrid framework of CNN-LSTM-attention proposed here has shown to have superior predictive power, defeating best performers of traditional models like XGBoost and LightGBM. Still, it has its drawbacks. Combining CNNs, LSTM networks, and attention processes, this hybrid model is computationally complex and calls for considerable power and memory. It is hard to come by in resource-limited settings, especially for organizations that can't access high-performance computing. On top of that, the model may not handle big data as well as with financial transactions in real time. Scalability issues like processing times, throughput, and resource memory leaks could become potential problems to limit the hybrid's payback potential.

In the future, researchers could concentrate on discovering methods to increase the efficiency of the hybrid model and maintain its accuracy. Filtering out unnecessary parameters is a job for model pruning. Knowledge distillation takes information from a large, difficult model and puts it into a smaller, simpler one. Both of these tasks should reduce the amount of computing that needs to be done.

Another area for future study could be how well the proposed hybrid structure works with various types of sequential financial data. The model could be modified to examine loan payback records, changes in stock prices, or data from insurance claims—sequential data that are all connected over time. Investigating the model's performance across these different situations could yield insights into its predictive power for various use cases.

4.7. Implications for research, practice, and society

The proposed hybrid CNN-LSTM-Attention framework enhances credit risk management by increasing prediction accuracy compared to traditional ML models. This has many positive implications for research, practice, and society.

From the research perspective, the hybrid model is a new way to use sequential data for financial risk forecasts. Its combination of space and time features makes it a good jumping-off point for research into more advanced ML models that could be used in other business scenarios. This study also provides insights into how to make high-performance, low-complexity hybrid models work better, which could lead to the development of new model architectures that maintain high predictive power without straining computational resources.

Applying this model will yield beneficial effects on our society too. The model's enhanced capacity to predict the risk of credit default will inject new stability into our economy by reducing the incidence of default and, thereby, ensuring that more individuals enjoy a fair chance at obtaining the credit they need. When financial institutions issue loans these days, they do so with an optimistic bias and some suboptimal estimates of risk. Even though the hybrid model achieves impressive precision, its computational intensity can limit its immediate applicability in some situations. We propose the use of modern technologies such as cloud computing, which provides scalable resources for dealing with massive amounts of calculations, to remedy this shortcoming. In addition, settings with high data volume might consider using distributed computing systems for an effective real-time deployment of the model.

4.8. Real-world financial implications

The hybrid CNN-LSTM-Attention architecture substantially improves prediction accuracy, but the financial costs of prediction results should be considered, especially when it comes to false positives. Financial firms may miss out on opportunities when a false positive occurs—erroneously classifying a trustworthy consumer as a defaulter—in the context of predicting whether someone will default on their credit card. Such errors could result in a loss of credit for many who can and would pay. It puts the financial institution at risk of lowering satisfaction with the credit experience, of lowering customer loyalty, and of lowering the revenues associated with much of the financial operations that firms engage in. The hybrid model has boosted predictive accuracy, with a 16% increase over conventional benchmarks, which may reduce the concerns linked to false positives. Providing more precise classifications allows financial institutions to optimize their decision-making processes regarding risk and return. We should remember, nonetheless, that misclassifications can and do happen, even with more precision, and when they happen, they have effects. Future studies might encompass learning methods that account for the economic fallout of incorrect positives and negatives during model training. Moreover, modeling methods that ascribe to the appearance of profit, and which may favor business-specific aims over generic measures of accuracy, could show more usable results for the specific applications we are concerned with. By these means, we could link model performance to their practical uses, with the added assurance that whatever framework we suggest aligns with the operational goals and economic implication of the financial institutions.

5. Conclusion

The hybrid CNN-LSTM-Attention framework we developed is novel to the field of credit default prediction. It achieved an improved prediction accuracy of 0.9981, outperforming traditional models like XGBoost (0.8824). This improvement emphasized that the hybrid model incorporates temporal dynamics into the credit risk modeling process leading to a benefit in model performance. In this study, we also examined the impact of hyperparameter tuning on the conventional ML as part of the study and then did a thorough comparative study that highlighted the importance of tuning for achieving precision in model building.

By investigating the effect of hyperparameter tuning on traditional ML models, we demonstrate that it impacts different models in diverse ways, thus highlighting the potential risk of applying a one-size-fits-all method across different types of models. In our view, hyperparameter tuning would be most beneficial for ensemble models. The hybrid model utilizes time-series data to identify intricate patterns in credit behavior, leading to a notable 16% enhancement in predictive accuracy relative to

conventional models. The improved precision of the hybrid model might lead to more reliable credit risk evaluations, decreasing default rates and refining credit distribution system. Moreover, its capacity to evaluate sequential data could help real-time credit monitoring and adaptive risk management, which is aligned with the evolving requirements of contemporary financial systems. This study illustrates how hybrid models may transform financial decision-making processes by integrating advanced ML techniques with practical applications.

However, we acknowledge that the complexity of the hybrid CNN-LSTM-Attention framework poses potential challenges for scalability and real-world application. While current computing resources are often sufficient, practical adoption may still require more consideration of computational resources. To address these concerns, future work could explore strategies to simplify the hybrid model without compromising accuracy. We aim to ensure that this framework could deliver its full potential in real-world credit risk management applications by balancing model complexity with operational constraints. In the future, we will explore proprietary or industry-specific datasets from financial institutions to confirm the robustness and applicability of the proposed framework to a wide range of contexts.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/competitions/amex-default-prediction, reference number [1].

Author Contribution Statement

Yujuan Qiu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Jianxiong Wang:** Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization.

References

[1] Howard, A., AritraAmex, Xu, D., Vashani, H., inversion, Negin, & Dane, S. (2022). *American express: Default prediction*. Retrieved from: https://www.kaggle.com/competitions/amex-default-prediction

[2] Arora, S., Bindra, S., Singh, S., & Nassa, V. K. (2022). Prediction of credit card defaults through data analysis and machine learning techniques. *Materials Today: Proceedings*, *51*, 110–117. https://doi.org/10.1016/j.matpr.2021.04.588

[3] Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, *29*(4), 563–574. https://doi.org/10.1016/j.ijforecast.2013.04.003

[4] Chou, T., & Lo, M. (2018). Predicting credit card defaults with deep learning and other machine learning models. *International Journal of Computer Theory and Engineering*, *10*(4), 105–110. https://doi.org/10.7763/IJCTE.2018.V10.1208

[5] Kim, H., Cho, H., & Ryu, D. (2018). An empirical study on credit card loan delinquency. *Economic Systems*, *42*(3), 437–449. https://doi.org/10.1016/j.ecosys.2017.11.003

[6] Fan, M., Wu, T. H., & Zhao, Q. (2023). Assessing the loss given default of bank loans using the hybrid algorithms multi-stage model. *Systems*, *11*(10), 505. https://doi.org/10.3390/systems11100505

[7] Sengar, A., Vijayaraghavan, A., & Hwang, J. (2023). A comparative analysis of hybrid quantum neural networks in binary credit defaulting tasks. In *IEEE MIT Undergraduate Research Technology Conference*, 1–5. https://doi.org/10.1109/URTC60662.2023.10535032

[8] Liu, J., Zhang, X., & Xiong, H. (2024). Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation. *Journal of Forecasting*, *43*(5), 1625–1660. https://doi.org/10.1002/for.3080

[9] Uddin, M. S., & Rahman, M. A. (2024). A comparative study of machine learning algorithms for enhanced credit default prediction. In *Evolutionary Artificial Intelligence: Proceedings of ICEAI 2023*, 201–216. https://doi.org/10.1007/978-981-99-8438-1_15

[10] Lai, Y. (2023). Credit default analysis and prediction based on machine learning. *Highlights in Business, Economics and Management*, *21*, 782–790. https://doi.org/10.54097/hbem.v21i.14762

[11] Li, D. (2019). Credit card fraud identification based on unbalanced data set based on fusion model. In *IEEE 1st International Conference on Civil Aviation Safety and Information Technology*, 235–239. https://doi.org/10.1109/ICCASIT48058.2019.8973167

[12] Qiu, Y. (2019). *Estimation of tail risk measures in Finance: Approaches to extreme value mixture modeling*. Master's Thesis, Johns Hopkins University.

[13] Qiu, Y., & Wang, J. (2024). A machine learning approach to credit card customer segmentation for economic stability. In *Proceedings of the 4th International Conference on Economic Management and Big Data Applications*, 1–8. http://dx.doi.org/10.4108/eai.27-10-2023.2342007

[14] Sayjadah, Y., Hashem, I. A. T., Alotaibi, F., & Kasmiran, K. A. (2018). Credit card default prediction using machine learning techniques. In *Fourth International Conference on Advances in Computing, Communication & Automation*, 1–4. https://doi.org/10.1109/ICACCAF.2018.8776802

[15] Guo, K., Luo, S., Liang, M., Zhang, Z., Yang, H., Wang, Y., & Zhou, Y. (2023). Credit default prediction on time-series behavioral data using ensemble models. In *International Joint Conference on Neural Networks*, 1–9. https://doi.org/10.1109/IJCNN54540.2023.10191783

[16] Greenacre, M., Groenen, P. J., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*, 100. https://doi.org/10.1038/s43586-022-00184-w

[17] Antoniadis, A., Lambert-Lacroix, S., & Poggi, J. M. (2021). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, *206*, 107312. https://doi.org/10.1016/j.ress.2020.107312

[18] Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

[19] Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural*

Computation, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

[20] Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

[21] Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235

[22] Wang, D. N., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on LightGBM. *Information Sciences*, *602*, 259–268. https://doi.org/10.1016/j.ins.2022.04.058

[23] Itoo, F., Meenakshi, & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*(4), 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

[24] Cardoza, I., García-Vázquez, J. P., Díaz-Ramírez, A., & Quintero-Rosas, V. (2022). Convolutional neural networks hyperparameter tunning for classifying firearms on images. *Applied Artificial Intelligence*, *36*(1), 2058165. https://doi.org/10.1080/08839514.2022.2058165

[25] Miao, J., & Zhu, W. (2022). Precision–recall curve (PRC) classification trees. *Evolutionary Intelligence*, *15*(3), 1545–1569. https://doi.org/10.1007/s12065-021-00565-2

[26] Nahm, F. S. (2022). Receiver operating characteristic curve: Overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1), 25–36. https://doi.org/10.4097/kja.21209

[27] Agarwal, A., Sharma, P., Alshehri, M., Mohamed, A. A., & Alfarraj, O. (2021). Classification model for accuracy and intrusion detection using machine learning approach. *PeerJ Computer Science*, *7*, e437. https://doi.org/10.7717/peerj-cs.437

[28] Zhou, Q. M., Zhe, L., Brooke, R. J., Hudson, M. M., & Yuan, Y. (2021). A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. *Diagnostic and Prognostic Research*, *5*, 13. https://doi.org/10.1186/s41512-021-00102-w

How to Cite: Qiu, Y., & Wang, J. (2025). Credit Default Prediction Using Time Series-Based Machine Learning Models. *Artificial Intelligence and Applications*, *3*(3), 284–294. https://doi.org/10.47852/bonviewAIA52023655