



# OPEN Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach

Chioma Ngozi Nwafor<sup>1✉</sup>, Obumneme Nwafor<sup>2</sup> & Sanjukta Brahma<sup>1</sup>

This paper uses a generalised stacking method to introduce a novel hybrid model that combines a one-dimensional convolutional neural network 1DCNN with extreme gradient boosting XGBoost. We compared the predictive accuracies of the proposed hybrid architecture with three conventional algorithms-1DCNN, XGBoost and logistic regression (LR) using a dataset of over twenty thousand peer-to-peer (P2P) consumer credit observations. By leveraging the SHAP algorithm, the research provides a detailed analysis of feature importance, contributing to the model's predictions and offering insights into the overall and individual significance of different features. The findings demonstrate that the hybrid model outperforms the LR, XGBoost and 1DCNN models in terms of classification accuracy. Furthermore, the research addresses concern regarding fairness and bias by showing that removing potentially discriminatory features, such as age and gender, does not significantly impact the hybrid model's classification capabilities. This suggests that fair and unbiased credit scoring models can achieve high effectiveness levels without compromising accuracy. This paper makes significant contributions to academic research and practical applications in credit risk management by presenting a hybrid model that offers superior classification accuracy and promotes interpretability using the model agnostic SHAP framework.

**Keywords** Credit risk, One-dimensional convolutional neural network, Shapley Additive explanation, Explainable artificial intelligence, And Logistics regression

Credit risk assessment is one of the main activities within the financial services industry, influencing the dynamics of lending and the broader economic landscape. Machine learning (ML) models have shown significant promise in enhancing predictive accuracy in credit risk assessment models because of their ability to model complex datasets. However, many credit risk analysts and regulators are wary of machine learning (ML) models because of concerns about their so-called “black box” nature. The black-box phenomenon alludes to the complexities of ML models that make interpreting their results challenging. In peer-to-peer (P2P) lending platforms, using black-box machine learning (ML) models to assess creditworthiness presents specific challenges. While efficient, these models lack transparency, making it difficult for borrowers and lenders to understand the decision-making process. This lack of interpretability in predictive ML models undermines trust in those models, especially in credit risk modelling and management, where decisions have significant financial consequences. In addition, this opacity challenges the adherence to fairness and regulatory compliance and hampers efforts to identify and mitigate biases that may disadvantage certain groups. Ultimately, black-box ML models lack transparency and accountability, undermining trust in the P2P platform and affecting its usability and acceptance. Unlike traditional linear regression models, where analysts can refer to the model's estimated coefficients to interpret the model predictions, ML models are so complex that it is challenging to determine an established link between the model's inputs and the output. A plethora of research attempts to address the trade-off between higher predictive power and explainability, giving rise to the concept of eXplainable Artificial Intelligent (XAI)<sup>1,2</sup>. These explainable techniques focus on identifying the contributions of input features to the model's overall and individual prediction.

This paper uses one of the frequently used XAI frameworks (Shapley Additive explanation), a concept borrowed from cooperative game theory, to open the “black box” of a novel ML-powered credit risk model<sup>3</sup>.

<sup>1</sup>Glasgow School for Business and Society, Department of Finance, Accountancy and Risk, Glasgow Caledonia University, Glasgow, Scotland. <sup>2</sup>School of Computing, Engineering and Built Environment, Glasgow Caledonian University, Glasgow, Scotland. ✉email: Chioma.Nwafor@gcu.ac.uk

The Shapley framework is a powerful tool for pinpointing sources of bias and facilitating more equitable lending practices by attributing the contribution of each feature to the prediction outcome. This study has three objectives: First, we compared the predictive accuracies of three distinct modelling approaches – a one-dimensional convolutional neural network (*1DCNN*), *XGBoost*, a novel hybrid convolutional neural network and *XGBoost* (*CNNs – XGBoost*) architecture and logistic regression (*LR*) using a P2P consumer credit dataset with over twenty thousand observations. Second, we used the SHAP algorithm kernel-based explainers to explain the model's ranking of overall features' significance and provide interpretability<sup>1</sup> of the basis of the model's predictions. The comparative analysis allows us to assess the efficiency of these models and explore how advancements in XAI can bridge the gap between high predictive accuracy and the need for transparency in financial decision-making. Finally, we examined whether removing potentially discriminatory features, specifically gender and age, from the hybrid model would affect its ability to classify outcomes accurately.

In fulfilling our research objectives, this paper makes three original contributions to the existing body of knowledge and the credit risk management practice. First, we demonstrated a novel approach by enhancing the one-dimensional convolutional neural network (*1DCNN*) performance through integration with an ensemble *XGBoost* model using the generalised stacking framework. This contribution improves the literature by offering a hybrid model that outperforms traditional methods and provides a practical tool for credit risk analysts seeking more accurate risk assessments. Secondly, we addressed concerns regarding fairness and bias by demonstrating that eliminating potentially discriminatory features, such as age and gender, from the hybrid model does not significantly detract from its classification accuracy. This contribution is particularly relevant to the ongoing discourse on ethical AI, offering evidence that fair and unbiased credit scoring models do not have to compromise on effectiveness.

Thirdly, we showed that the *1DCNN* model better provided a complete view of the factors influencing credit risk than the *XGBoost* because of the limitations of *XGBoost* in providing local explainability. Collectively, these contributions enhance our understanding of how to improve the accuracy, transparency, and fairness of ML-powered credit scoring models. The rest of the paper is structured as follows: Section two presents a brief review of related literature. In section three, we discussed the SHAP framework and the research hypothesis. Section four is the study's methodology and description of the dataset, section five discusses the experiment's results, and section six is the conclusion and direction for future research.

## Related studies

Credit risk assessment methodologies have evolved from traditional statistical techniques to sophisticated machine learning models. The transition to ML scoring models is exacerbated by the volume, variety and velocity of data available to banks, which have put them in a position to exploit advanced analytics in managing credit risk. Many studies have applied deep learning models in discriminating between good and bad credit risk. For example<sup>4</sup>, employed a hybrid method combining a relief algorithm with CNN to score credit risk. Their experiments revealed that the hybrid relief-CNN model showed better predictive accuracy than the logistic regression (LR) and random forest (RF) models. In another study<sup>5</sup>, used deep learning with clustering and merging, and the result showed high prediction accuracies. Similarly<sup>6</sup>, proposed a hybrid model using a dimension reduction approach and a multi-layer ensemble classifier. The results showed that the hybrid model achieved better performance in credit scoring. While<sup>7</sup> compared ensemble methods and logistics regression in large dataset, and found that ensemble methods, particularly *XGBoost* significantly outperformed logistic regression.

As<sup>8</sup> highlighted, integrating Artificial Intelligence (AI) in credit scoring improves predictive accuracy but also raises concerns regarding transparency and interpretability. The opaque nature of these models limits the stakeholders' ability to understand the basis of the model's decision-making processes<sup>9–11</sup>. In response to these concerns, the field of Explainable AI (XAI) has emerged as a critical area of research, aiming to make AI systems more transparent and their decisions understandable to humans<sup>2</sup>. XAI is a set of methods that allows users to understand and trust the outputs created by machine learning methods. The most widely used XAI frameworks is the Shapley Additive exPlanation (SHAP) proposed by<sup>3</sup>. The SHAP is a model-agnostic XAI approach that can be applied to the predictive output, regardless of which model generated it<sup>12</sup>. It is based on the concept of cooperative game theory introduced by<sup>13</sup>. A unique feature of Shapley-based explainable method is its ability to explain single-point prediction (local explanations) and provide a global explanation- i.e. explaining the overall workings of the model (global explanations) in a fair and interpretable manner<sup>12</sup>.

Our study evaluates the contributions of loan customers' age and gender to predicting default using three different ML models and highlights the important need to evaluate models based on their fairness and ability to reduce bias. This shift in focus emphasises the need to examine how well a model predicts default and how discriminatory features affects classification accuracy. The works of<sup>14</sup> on one-dimensional Convolutional Neural Networks (1)<sup>15</sup>, on the *XGBoost* model for corporate failure prediction, and<sup>16</sup> on logistic regression provide a solid foundation for comparing the performance of these methods and the hybrid model in the context of credit risk assessment. Each study contributes to understanding the trade-offs between model complexity, performance, and interpretability.

## Shapley explainability framework and formulation of hypothesis

Societies exhibit a wide range of biases and inequalities, which are inherently present in the data that train machine learning systems. Without deliberate efforts to counteract these biases, algorithms developed from such data will inevitably lead to prejudiced results, exacerbating the disparities that marginalised groups face<sup>17</sup>.

<sup>1</sup> In this study will have used interpretability and explainability interchangeably.

Bias in machine learning models refers to systematic and unfair discrimination against certain groups based on data. Bias can occur due to various factors, including biased data collection, unrepresentative training datasets, or flawed algorithmic design<sup>18</sup>. For instance, ethnicity, age, gender, or socioeconomic biases can lead to unfair outcomes or decisions by automated systems<sup>19</sup>. Like all algorithmic decision-making systems, automated credit scoring systems are fed with historical loan data to predict future defaults. However, allowing discriminatory features such as ethnicity, gender, and age in the training dataset could result in a biased automated system. This potential bias can have significant consequences, affecting the fairness and accuracy of credit risk assessment. Consequently, removing these features from the credit scoring training dataset can reduce potential bias in such a system. However, the possible contributions of these features to the model's classification accuracy are yet to be explored. The present study bridges that gap by introducing age and gender in the novel hybrid model to understand their sensitivity to the model's classification accuracy. In addition, transparency in ML systems is one of the ways of reducing bias. Transparency in ML systems involves making the workings of ML systems understandable to humans, especially regarding how the model arrives at its decisions. The Shapley framework has been frequently used in previous studies to achieve transparency<sup>8,12,20</sup>.

**Shapley Value** The Shapley value, named after Lloyd Shapley, who introduced it in the 1950s, is a concept from cooperative game theory that offers a fair distribution of payoffs among players based on their contribution to the total payoff of the game<sup>3</sup>. The Shapley value is used in explainable artificial intelligence (XAI) to determine the importance of features in machine learning models. To understand the contribution of each feature to the model's prediction, we calculate the Shapley values for all features<sup>21</sup>. Therefore, the SHAP value for any feature is the average marginal contribution of the feature across all possible feature coalitions.

To understand the mathematical foundation of Shapley values, consider a cooperative game with a set of players  $N$ , where each player is an integer from 1 to  $n$ . A coalition  $S$  is any subset of  $N$ , including the empty set and  $N$  itself. The game is defined by a characteristic function  $v : 2^N \rightarrow \mathbb{R}$  that assigns a value to each coalition  $S$ , representing the total payoff that the members of  $S$  can obtain by cooperating. The value of the empty set is Zero ( $v(\emptyset) = 0$ ). The Shapley value aims to distribute the total payoff generated by coalition  $N$  among the players in a way that reflects their individual contributions. The contribution of a player  $i$  to a coalition  $S$  is the marginal value that  $i$  adds to  $S$ , defined as  $v(S \cup \{i\}) - v(S)$ . The Shapley value of player  $i$ , denoted by  $\varphi_i(v)$ , is calculated as the weighted average of  $i$ 's marginal contributions to all possible coalitions that  $i$  can join. Mathematically, it is given by:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \quad (1)$$

Where  $|S|$  is the cardinality of set  $S$ ,  $n$  is the total number of players, and the summation is over all subsets  $S$  of  $N$  that do not include player  $i$ . If we assume a linear model and we want to explain a prediction for a variable  $x^i$ . The prediction for  $x^i$  is:

$$\hat{\omega}(x^{(i)}) = \alpha + \beta_1 x_1^{(i)} + \dots + \beta_p x_d^{(i)} \quad (2)$$

Where  $x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}$  are feature values. Let  $\varnothing_j$  be a contribution of feature  $j$  on prediction  $\hat{\omega}(x^{(i)})$ . The contribution  $\varnothing_j$  is given as:

$$\begin{aligned} \varnothing_j(\hat{\omega}) &= \beta_j x_j^{(i)} - E(\beta_j \mathbf{X}_j) \\ &= \beta_j x_j^{(i)} - \beta_j E(\mathbf{X}_j), \end{aligned} \quad (3)$$

Where  $E(\mathbf{X}_j)$  is the expected value for feature  $j$ . If all feature contributions are summed up for one feature, the result is

$$\begin{aligned} \sum_{j=1}^d \varphi_j(\hat{\omega}) &= \sum_{j=1}^d (\beta_j x_j^{(i)} - E(\beta_j \mathbf{X}_j)) \\ &= \left( \alpha + \sum_{j=1}^d \beta_j x_j^{(i)} \right) - \left( \alpha + \sum_{j=1}^d E(\beta_j \mathbf{X}_j) \right) \\ &= \hat{\omega}(x^{(i)}) - E(\hat{\omega}(\mathbf{X})), \end{aligned} \quad (4)$$

Equation (4) is the difference between the predicted value for the feature  $x^{(i)}$  and the average predicted value for the dataset  $\mathbf{X} = \{\mathbf{X}_j\}_{j=1}^d$  where  $\mathbf{X}_j = \{x_{ij}\}_i^N = 1$ .

The Shapley framework has several desirable properties that justify its widespread use in XAI studies. First is its efficiency, which ensures that the total payoff generated by a cooperative game is fully distributed among the players. The second property is the symmetry property, which implies that if two features contribute equally to the prediction outcome across all possible subsets of features, they should be assigned the same importance or Shapley value.

The symmetry principle ensures that the Shapley value does not discriminate between players based on their identity but rather on their contribution to the game. This property underpins the fairness and equity of the

Shapley approach in attributing contributions in explainable AI, ensuring that identical contributions receive identical rewards. The third property is the dummy player property which plays a critical role in explainable AI by helping to identify and disregard non-contributing features in predictive models. Suppose a feature is determined to be a dummy (i.e., its inclusion or exclusion does not change the model's predictions), in that case, its Shapley value will be low or zero, indicating little to no importance. This enhances the interpretability and efficiency of models and ensures that explanations are focused on features that genuinely influence the model's predictions. The dummy player property respects the principle of fairness in the model explanation. It prevents overestimating the importance of features that, in reality, do not influence the model's decisions. The above-mentioned Shapley properties help in enhancing transparency in ML models.

In this study, we addressed the biases in ML credit scoring systems by proposing and evaluating a hybrid model through various hypotheses. First, we assessed the effectiveness of the hybrid model's performance in credit risk analysis. The null hypothesis ( $H_0$ ) is that the hybrid model does not significantly improve the accuracy of credit risk predictions compared to conventional credit scoring models. While the alternative hypothesis ( $H_1$ ) is that the hybrid model significantly improves the accuracy of credit risk predictions compared to traditional credit scoring models.

$$H_0 : \mu_{\Omega} = \mu_{LR}; \mu_{CNN}; \mu_{XGBoost} \quad (5)$$

$$H_1 : \mu_{\Omega} \neq \mu_{LR}; \mu_{CNN}; \mu_{XGBoost} \quad (6)$$

Where  $\mu_{\Omega}$  represents the mean predictive accuracy of the hybrid model.  $\mu_{LR}; \mu_{CNN}; \mu_{XGBoost}$  represents the mean predictive accuracies of the logistics regression, the one-dimensional neural network and the *XGBoost*. Second, we explored the explainability of decisions made by the one-dimensional *1DCNN* model compared to logistic regression and *XGBoost*. Here,  $H_0$  suggests that there is no significant difference in the explainability of credit risk decisions between the *1DCNN* model and traditional methods (*XGBoost* and *LR*). While the alternative hypothesis is that the explainability of the credit risk decisions made by the *1DCNN* model is significantly better than those made by traditional credit scoring models. Let's denote the explainability of the *1DCNN* model as  $E_{1D}$  and the explainability of traditional models as  $E_{trad}$ .

$$H_0 : E_{1D} - E_{trad} = 0 \quad (7)$$

$$H_1 : E_{1D} - E_{trad} > 0 \quad (8)$$

Where  $E_{1D}$  and  $E_{trad}$  are measured using the contributions of the inputs on the model's decision. Lastly, we investigated the effect of excluding discriminatory features on the hybrid model's accuracy. Let us denote the classification accuracy of the hybrid model with all features as  $A_{full}$  and the model's classification accuracy after removing discriminatory features as  $A_{reduced}$ . Mathematically, these hypotheses can be represented as:

$$H_0 : A_{full} - A_{reduced} = 0 \quad (9)$$

$$H_1 : A_{full} - A_{reduced} \neq 0 \quad (10)$$

In this formulation,  $A_{full}$  represents the accuracy of the model when it includes all features, and  $A_{reduced}$  represents the accuracy after the removal of discriminatory features. The null hypothesis  $H_0$  suggests that the accuracy remains unchanged after removing such features (age and gender). In contrast, the alternative hypothesis  $H_1$  suggests a significant difference in accuracy, indicating that the discriminatory features had an impact on the model's performance.

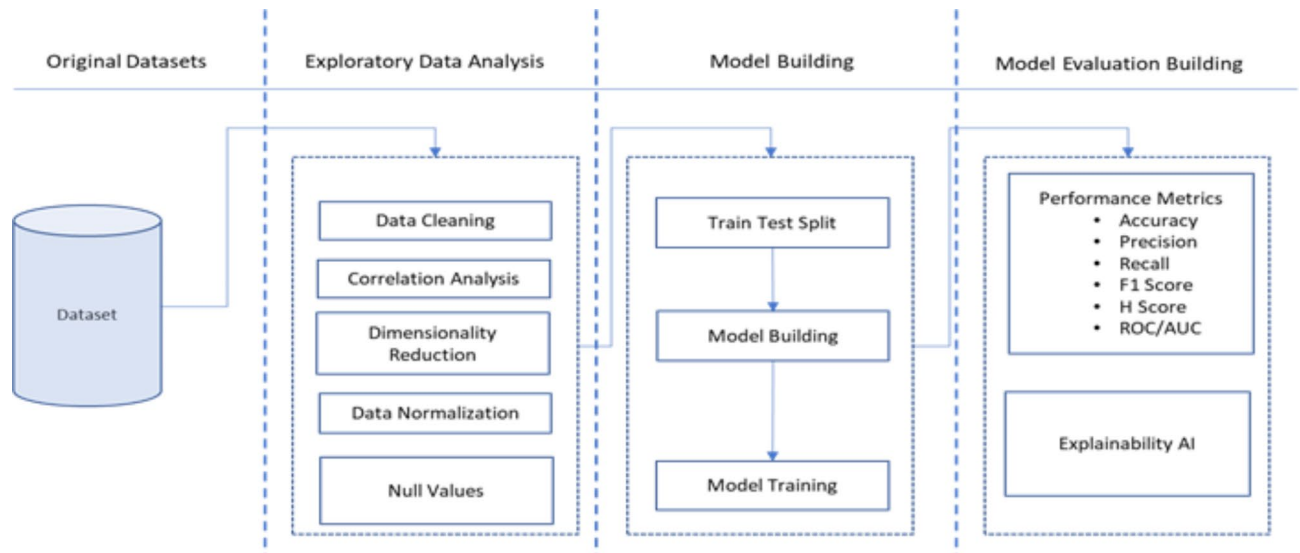
## Methodology and data

We employed a hybrid machine learning model that combines a *1DCNN* with *XGBoost* using a generalised stacking framework and the SHAP algorithm for feature importance analysis. The software tool used is Python Programming Language, which runs on Jupiter Notebook. The Python packages used include Keras, Sklearn, NumPy, Pandas, Shap, Matplotlib, and *XGBoost* Classifier, with additional reliance on Python-based libraries for data pre-processing and model evaluation. Figure 1 shows a schematic diagram of the experiment's process flow. This involves three iterative processes: exploratory data analysis, model building, and model evaluation.

The dataset used in this experiment was customer loans from Lending Club, a P2P lending company. The original dataset is publicly available on<sup>22</sup> and corresponds to all the loans issued by the firm between 2007 and 2018. The dataset has one million forty-eight thousand five hundred and seventy-six observations with one hundred and forty-five input features. During the exploratory data analysis, we removed features with more than 40% missing values and dropped highly correlated features. The reduced input feature space used for this experiment comprises twenty-five features and twenty-five thousand, five hundred and thirty-five observations. Next, we addressed labelling categorical variables, transforming them into a numerical format using label encoding. We used the Min-Max scaling technique to normalise the input features, which scales the features to a fixed range of 0 to 1. The formula used is:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (11)$$

For the training phase, the dataset was split into training (80%) and testing (20%) samples to provide a robust evaluation of the model's performance on unseen data. The 80/20 split is a commonly used rule in machine



**Fig. 1.** Process flow diagram of experiment methodology.

learning for dividing data into training and testing sets<sup>8,20</sup>. This ratio achieves a sufficient balance between having enough data to train the model effectively (80% for training) and enough data to test and evaluate the model's performance (20% for testing). With larger datasets, the 80% training portion would capture the variability and distribution of features within the data, which means the model training covers most scenarios expected in the application. We adopted the  $K$ -fold validation method to assess the effectiveness of the models, particularly to avoid overfitting. This approach randomly divides the dataset into  $K$  groups or approximately equal-sized folds<sup>23</sup>. The model is trained  $K$  times, using  $K - 1$  folds as the training set and the remaining single fold as the validation set. This process is repeated so that each of the  $K$  folds is used as the validation set exactly once.

This study used 5 groups (5-fold cross-validation). The first 4-subsets were used to develop the training sample for the model. The resulting model is assessed for accuracy on the remaining 1/5th of the sample (test sample). The results of the models from each fold are averaged to produce a single estimation. This averaging helps to reduce variability and provides a more accurate measure of model performance. Finally, we employed Grid Search as our optimisation strategy to select the optimal hyperparameters for each model. This method works through all possible combinations of parameters, which, while potentially computationally expensive, ensures that no potential configuration within the specified range is overlooked<sup>24,25</sup>. This can be particularly advantageous in scenarios where the optimal hyperparameter values are unknown, and a detailed search is necessary to ascertain the best-performing parameters.

### Logistic regression (LR)

(LR) is a statistical method that analyses the relationship between one or more independent variables and a binary dependent variable<sup>26</sup>. logistic regression is used for binary classification tasks, making it highly relevant in credit risk analysis, where the goal is to predict whether an event (loan default) will occur<sup>16</sup>. The core of logistic regression is the logistic function (Sigmoid function), which maps any real number into a value between 0 and 1, making it interpretable as a probability. The logistic function is defined as:

$$P(Y = 1) = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad (12)$$

Where,  $P(Y = 1)$  represents the probability of the dependent variable equals to 1 (loan default),  $e$  is the base of the natural logarithm,  $\beta_0$  is the intercept term, and  $\beta_1, \beta_2, \dots, \beta_n$  are the coefficients of the independent variables  $X_1, X_2, \dots, X_n$ , which influence the outcome.

$$Odds = \frac{P(Y = 1)}{1 - P(Y = 1)} \quad (13)$$

Equation (13) is the odds ratio which is the ratio of the probability of the event occurring to the probability of it not occurring.

$$\log \left( \frac{P(Y = 1)}{1 - P(Y = 1)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (14)$$

Equ (14) is the natural logarithm of the odds ratio which leads to the logit function. The equation linearly relates the independent variables to the log odds of the dependent event. While logistic regression provides



a straightforward, interpretable model for credit risk prediction, its linear nature and limitations in handling complex, non-linear interactions and unstructured data make it less effective in capturing the full spectrum of risk factors.

#### *XGBoost*

*XGBoost* builds on the gradient boosting framework, where new models are added sequentially to correct the errors made by existing models<sup>27</sup>. The main idea is to consecutively add trees to the model, where each new tree corrects the residual errors made by the existing ensemble of trees. The overall prediction model in *XGBoost* is an ensemble of decision trees formulated as:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (15)$$

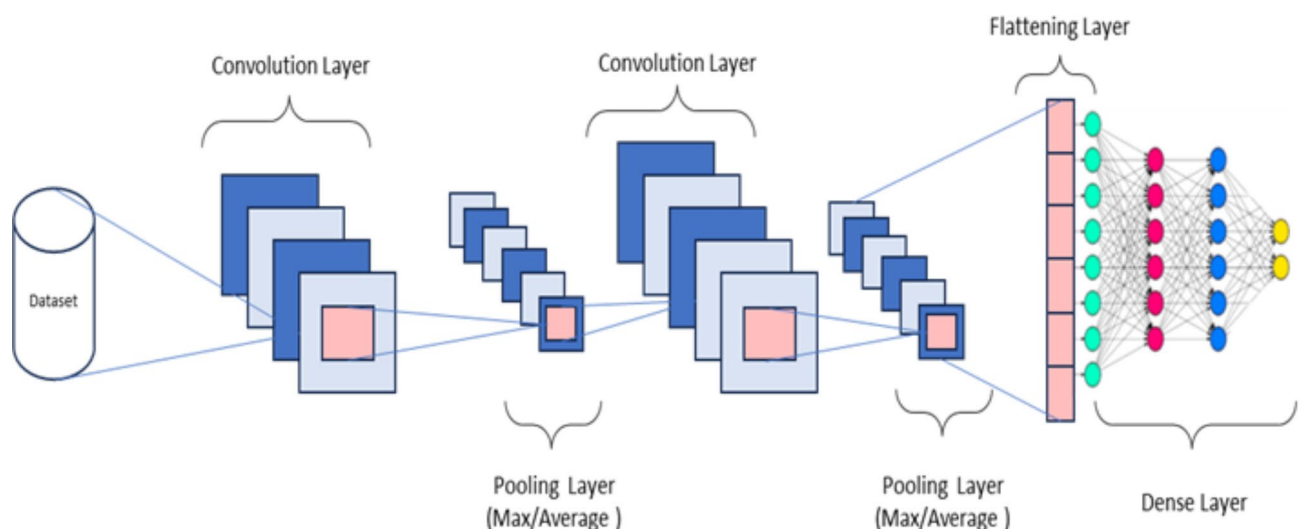
Where  $\hat{y}_i$  is the model's prediction for the  $i^{th}$  instance,  $f_k$  represents an individual decision tree,  $x_i$  is the feature vector for the  $i^{th}$  instance, and  $K$  is the number of trees. The objective function in *XGBoost* combines a loss function  $L$  which measures how well the model's prediction match the actual data. The regularisation term  $\Omega$  penalises the complexity of the model to prevent overfitting. It can be expressed as:

$$Obj = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (16)$$

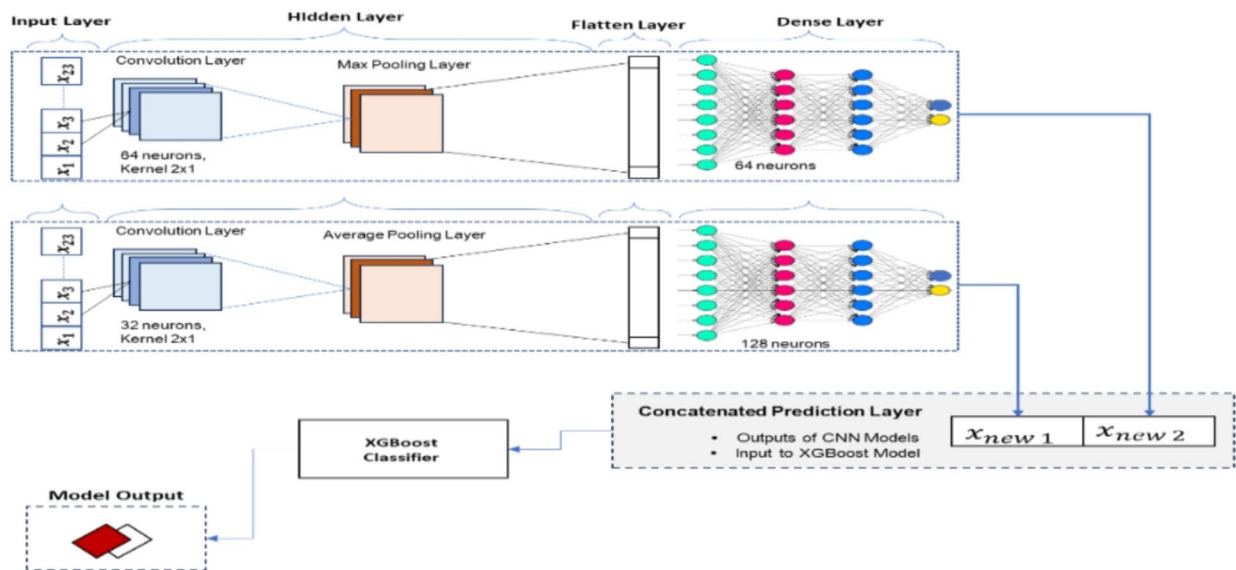
Where,  $y_i$  is the actual value for the  $i^{th}$  instance, and  $\Omega(f_k)$  represents the regularisation term for the  $k^{th}$  tree. *XGBoost* optimises this objective function using a gradient descent approach, where it calculates the gradient of the loss function with respect to the model's predictions. The model is updated in a direction that minimises the loss. The regularisation term helps in reducing overfitting by controlling the model's complexity. Its gradient boosting mechanism enables the model to learn complex, non-linear relationships between the features and the target variable, which is crucial for accurately assessing credit default risk. Although *XGBoost* offers several advantages for predicting credit risk, it has limitations that necessitate using our proposed hybrid model. A significant limitation of this model is its approach to sequential or temporal data. *XGBoost* is designed to handle static, structured data like most tree-based models. Sequential or temporal data contains dependencies where the significance of a current observation may depend on previous observations. For example, a series of missed payments or a steadily increasing debt-to-income ratio over time could indicate a growing risk of default that might not be as evident when considering these factors independently of their temporal context. *XGBoost's* standard approach does not explicitly model such temporal dependencies, which could lead to less accurate risk assessments for individuals with significant behavioural patterns over time.

#### *Convolutional neural network*

Convolutional neural network (CNN) offers an optimal architecture for learning complex features in large datasets heuristically. CNN has significant applications in computer vision, image classification/analysis<sup>28</sup>, nuclear and particle physics<sup>29</sup>. A key feature of CNN methods is their ability to achieve 'spatial invariance'-i.e. they can learn to recognise and extract image features anywhere in the image (Ibid). CNN uses a unique feature called 'convolution' to condense images to a form that is easier to process without losing features crucial for prediction accuracy. Figure 2 below is a schematic diagram of a CNN. It shows the flow from the input layer through convolutional layers, followed by pooling layers, to a flattening step. After flattening, the data passes through dense (fully connected) layers, culminating in an output layer with a sigmoid activation function for the binary classification. The convolutional layer applies a set of learnable filters to the input data. In the context of



**Fig. 2.** Schematic diagram of a convolutional neural network.



**Fig. 3.** Architecture of the hybrid *CNN – XGBoost* model.

Feature name	Mean	Standard deviation	Maximum	Minimum
<i>Emp_Length</i>	5 years	4 years	11 years	0 year
<i>Annual_Inc</i>	\$80,696.06	\$81,295.23	\$5,119,032.00	\$0.00
<i>Debt to income DTI ratio</i>	19.67	17.09	25.57	0
<i>Age</i>	49 years	18 years	80 years	33 years
<i>Loan_amt</i>	\$16,618.45	\$10366.70	\$40,000	\$1000
<i>Term</i>	44 months	11 months	60 months	36 months
<i>Interest rate</i>	14.54%	5.56%	30.99%	5.31%
<i>Total bankcard Credit limit</i>	\$25,476.43	\$25,594.13	\$4,174,00.00	\$0.00
<i>Revolving line utilization</i>	\$43,531.28	\$25,374.95	\$146,300	\$0.00
<i>Total credit balance excluding mortgage</i>	\$52,594.60	\$54,106.31	\$1,077617.00	\$0.00
<i>Revolving balance</i>	\$15,710.89	\$22,833.64	\$1,049,095.00	\$0.00
<i>Total payments</i>	\$1785.03	\$2522.43	\$40,856.679	\$0.00
<i>Remaining outstanding principal</i>	\$12,288.02	\$10,830.76	\$40,000.00	\$0.00
<i>Interest received to date</i>	\$629.27	\$763.65	\$8898.69	\$0.00
<i>Total current balance of all account</i>	\$137,766.00	\$163,096.10	\$216,3151.00	\$0.00

**Table 1.** Descriptive statistics of the features.

a *1DCNN*, these filters slide (convolve) over one dimension of the data, processing one-dimensional sequence data. This operation aims to extract high-level features from the input sequence. Each filter in the convolutional layer is designed to detect specific features, such as trends. The output of this layer is a set of feature maps representing the presence of detected features across the sequence.

A non-linear function such as the Rectified Linear Unit (ReLU) function is applied on feature maps. In a binary classification task, the last dense layer has a single neuron with a sigmoid activation function, which outputs a probability indicating the likelihood of the input belonging to one class or the other. The one-dimensional convolutional neural network (*1DCNN*) is capable of analysing temporal and sequential data, making it an ideal tool for examining financial behaviours over time. Unlike traditional neural networks and ensemble models, *1DCNN* applies convolutional operations to one-dimensional data, extracting significant features from sequences such as payment histories or credit utilisation trends. This capability allows the model to identify temporal patterns that signal potential financial distress or stability in borrowers, providing critical insights for predicting loan defaults. Although CNNs are very effective in local and hierarchical pattern detection in data, they might only partially capture complex, non-linear interactions between features across the whole dataset. XGBoost can complement this by explicitly modelling these interactions and dependencies. Hence, there is a need to develop a hybrid model that leverages the strengths of both models.

### The hybrid model

Our hybrid model employs a specialised technique known as generalised stacking, which layers multiple initial models to form a composite, hybrid architecture. The generalised stacking method is better suited for hybridisation because it allows the meta-model (XGBoost) to learn from the residuals or errors of the base models' (CNNs) initial predictions. This process refines the final output by addressing and correcting specific weaknesses in the base models' predictions. In addition, the generalised stacking offers flexibility in how the models are combined. For instance, the meta-model was trained to focus on aspects where the base models underperform, such as overfitting training data, leading to poor generalisation of unseen data. The meta-model (XGBoost) mitigates this by adjusting the final decision boundary based on a detailed understanding of errors made by the CNNs.

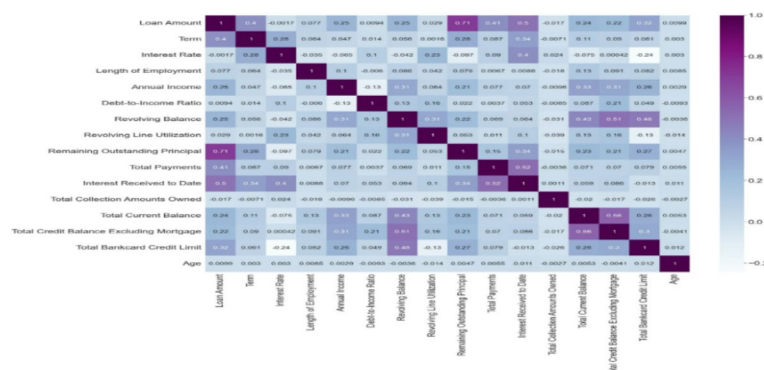
Our objective in introducing the hybrid model is to enhance features extraction and robustness by integrating diverse modelling approaches. The hybrid model capitalises on the complementary strengths of deep learning and ensemble techniques. Specifically, the one-dimensional CNNs can effectively capture local dependencies and patterns in the data, leveraging their strength in handling sequence or time-series data. On the other hand, XGBoost is very effective in handling structured data and can model the non-linear interactions between credit risk features. Overall, the hybrid model can better generalise to new data by taking advantage of the strengths of each model, reducing the risk of overfitting to the training data compared to using a single model.

Figure 3 is the schematic diagram of the hybrid architecture. Specifically, the dataset was trained using 1DCNNs. The predictions from these base models are used as input features for another model (meta model)—*XGBoos*. After the data pre-processing discussed in Sect. [Methodology and data](#), we designed the architectures for the two one-dimensional CNNs. We trained the data independently to capture different patterns in the data, with their predictions forming meta-features. These meta-features create a new dataset for training the meta-model. XGBoost is trained on this meta-dataset to learn from and correct errors made by the CNNs, optimising performance through hyperparameter tuning. The stacked model's predictions are validated using the performance evaluation metrics discussed in Sect. [Performance evaluation metrics](#). The hybridisation takes advantage of the CNN models for feature extraction capabilities and XGBoost for handling complex feature interactions.

### Data

Table 1 is the descriptive statistics of the numeric features. Using Pearson's correlational heat map on continuous variables and the *Chi-square* test for categorical variables, we identified and dropped highly-correlated variables. We experimented with the new credit risk dataset containing a reduced dimension in the input feature space. The features cover a range of information pertinent to assessing the risk and viability of the loans from both the borrower's financial and personal standpoint. Based on the distribution of the dataset, we created two demographic data, gender and age, to test our second hypothesis. Figure 4 shows the correlation heatmap. Some interesting insights from the correlation suggest a strong positive correlation of 71% between the loan amount and outstanding principal, indicating that higher loan amounts tend to have higher remaining principals.

There was expected as larger loans take longer to pay off. A moderate positive correlation of 28% between interest rate and loan term suggests that longer-term loans come with higher interest rates. We discovered a significant negative correlation between loan status and interest rate. This suggests that higher interest rates are associated with a specific loan status outcome. This could imply that loans with higher interest rates are less likely to be in good standing, perhaps due to the increased financial burden on the borrower. There was a negative correlation between loan status and annual income, indicating that the probability of a loan default decreases as the borrower's annual income increases. We consider these relationships within the broader context of loan management and borrower behaviour to glean insights into risk factors and opportunities for improving loan outcomes. Table 6 in the [appendix](#) presents the definitions and measurements of the variables employed in our study.



**Fig. 4.** Correlation matrix.



	Definition	Measurement
<b>Dependent Variable</b>		
<i>Loan status</i>	charged-offs, default, and late (31–120 days)	(1, if loan is in default and 0, if performing)
<b>Borrower Characteristics</b>		
<i>Emp_Length</i>	Length of current employment	Years
<i>Annual_Inc</i>	Annual Income	Thousand
<i>DTI</i>	Debt-to-Income ratio	Ratio
<i>Home_owns</i>	Home ownership status	Rent = 1, Own outright = 2, Own with Mortgage = 3, Other = 4
<i>Pub_Rec</i>	Public record bankruptcies	Categorical variable ....
<i>Disbursement method</i>	The method by which the borrower receives their loan.	Categorical variable
<i>Tax liens</i>	legal claim against the assets of a person who fails to pay taxes.	Thousands
<i>Age</i>	Measures the age of the loan customers	Years
<i>Gender</i>	Measures the gender of the loan customers	Binary (Male or Female)
<b>Loan Characteristics</b>		
<i>Loan_amt</i>	Loan Amount	Thousands (numeric)
<i>Term</i>	Loan Contract Term	Months
<i>Interest</i>	Interest rate on the loan	%
<i>Account balance</i>	Cash balance in borrow loan proposer's account	Thousands (numeric)
<i>Verification status</i>	Source of income verified	Dummy (1 if income was verified, 0 = otherwise)
<i>Application type</i>	Type of loan application	Individual or joint application
<i>Credit limit</i>	Credit limit on an account	Thousands (numeric)
<i>Loan purpose</i>	A category provided by the borrower for the loan request	Categorical variable
<b>Repayment Characteristics</b>		
<i>Revolving line utilization</i>	Amount of loan the borrower is using	Thousands (numeric)
<i>Total credit balance excluding mortgage</i>	Total credit balance without mortgage	Thousands (numeric)
<i>Revolving balance</i>	Total revolving credit	Thousands (numeric)
<i>Interest received to date</i>	Total interest payments received to date	Thousands (numeric)
<i>Total payments</i>	Total payments received to date for the funded amount	Thousands (numeric)
<i>Remaining outstanding principal</i>	Measures the loan principal still outstanding	Thousands (numeric)
<i>Principal received to date</i>	Measures the principal the firm has received to date	Thousands (numeric)
<i>Total current balance of all account</i>	The current balance of customer $i$ in all accounts	Thousands (numeric)
<i>Average current balance all account</i>	The average balance of customer $i$ in all account	Thousands (numeric)
<i>Total open to buy on revolving bankcards</i>	Total amount that investors can buy on revolving bankcards	Thousands (numeric)

**Table 6.** Variables definitions and measurements.*Performance evaluation metrics*

We used the following metrics to test the performance and predictive accuracies of the *1DCNN*, *XGBoost*, logistic regression and the hybrid models. Precision (P) is a metric that measures the number of correct (True) positives classifications. It is calculated as the ratio of correctly predicted positive (True positives) and all the total number of positives that were predicted.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)} \quad (17)$$

The Recall which measures the number of True positive classification made out of all positive predictions is another evaluation metric used in this paper. It measures the model's ability to correctly identify True Positives and is measured as

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \quad (18)$$

The F-1 score combines the precision and recall scores of the model to measure the model's accuracy. This evaluation metric is calculated as the harmonic mean of the precision and recall scores and it is calculated as:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (19)$$

The precision and recall for each class (default and non-default) were used to calculate the macro and weighted scores using the formulae

$$macro - average = \frac{\sum_k SCORE_k}{N} \quad (20)$$

$$weighted - average = \sum_k SCORE_k W_k \quad (21)$$

The  $SCORE_k$  in Eqs. 19 and 20 denote either the precision, recall and F1-score for each of the target variable class  $k = \{0 \text{ and } 1\}$  and  $W_k$  is the ratio of the number of examples in the training or testing dataset. We also calculated the area under the curve (AUC) and the  $H - measure$  for the hybrid. The AUC measures the predictive power of a model<sup>21</sup> and it is calculated as follows:

$$AUC = \frac{1}{2} \left( 1 + \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \right) \quad (22)$$

The higher the AUC, the better the predictive power of the model. To measure the cost of misclassification in our hybrid model, we calculated the  $H - measure$ , a statistical measure proposed by Hand in 2009 as an alternative to the ROC curve (AUC) used for evaluating the performance of binary classification models<sup>30</sup>. argues that the AUC does not consider that different misclassification errors can have different consequences. Consequently, the  $H - measure$  addresses this limitation by incorporating the concept of misclassification costs into its calculation. It acknowledges that the costs associated with false positives and negatives can vary greatly depending on the application. In the current case, the cost of mistakenly granting a loan to a customer who defaults (false positive) often differs from the cost of denying a loan to a customer who would not have defaulted (false negative). The  $H - measure$  is calculated as:

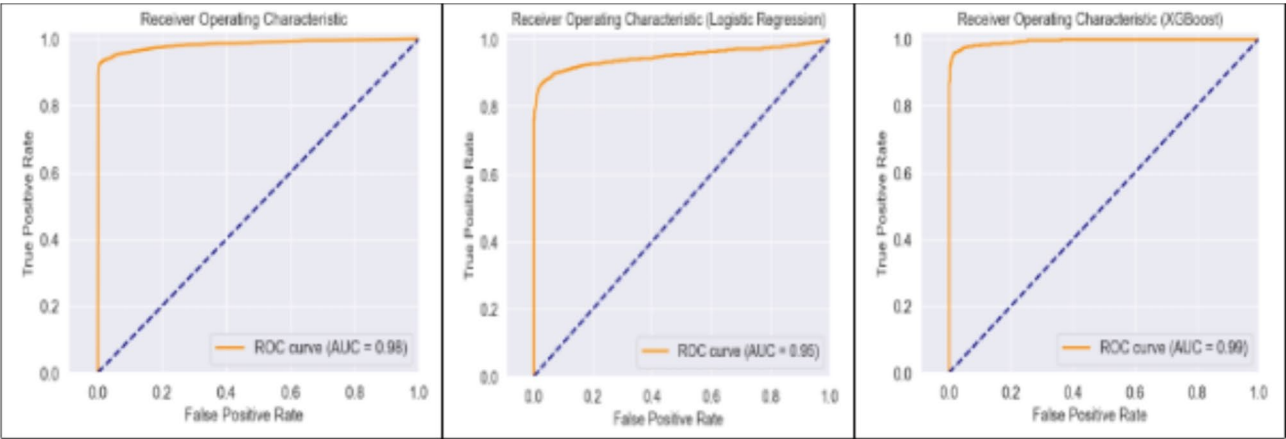
$$H = 1 - \frac{EC(hb_{classifier})}{EC(rd_{classifier})} \quad (23)$$

Where  $EC(hb_{classifier})$  is the expected cost of the hybrid classifier and  $EC(rd_{classifier})$  is the expected cost of a random classifier. The measure ranges for 0 to 1, where 1 indicates no misclassification cost and 0 indicates a performance equivalent to random guessing.

Hybrid IDCNNs-XGBoost						
	Precision	Recall	F1-Score	Accuracy	AUC	H - Score
0	0.94	0.98	0.96	0.96	0.98	0.95
1	0.98	0.94	0.95			
Macro Avg.	0.96	0.96	0.96			
Weighted Avg	0.96	0.96	0.96			
1DCNN						
0	0.95	0.99	0.97	0.94	0.98	0.92
1	0.98	0.94	0.96			
Macro Avg	0.96	0.96	0.94			
Weighted Avg	0.96	0.96	0.96			
Logistics Regression						
0	0.88	0.98	0.93	0.92	0.95	0.95
1	0.98	0.86	0.92			
Macro Avg	0.93	0.92	0.92			
Weighted Avg	0.93	0.92	0.92			
XGBoost						
0	0.92	0.98	0.95	0.94	0.99	0.95
1	0.98	0.91	0.94			
Macro Avg	0.95	0.94	0.94			
Weighted Avg	0.95	0.94	0.94			

**Table 2.** Performance evaluation metrics across the different ML models with all the variables.

$$H_0 : \mu_{\Omega} = \mu_{LR} : \mu_{CNN}$$



**Fig. 5.** AUCs for the hybrid, LR and *XGBoost* models.

	Accuracy	Precision
<i>XGBoost</i> classification	80%	76.46%
Support Vector classification	76%	85.19%
Random Forest Classification	83.2%	82.95%
Neural Network Classification	70.4%	65.18%
Logistic Regression	80.8%	78.04%
Hybrid Model	<b>90.0%</b>	<b>84%</b>

**Table 3.** Comparison of the hybrid model and other models using the German credit dataset. **Source:** Hoffman (1984) Statlog (German Credit Data). Baseline Performance and hybrid model performance.

Authors	Dataset	State of the act studies	AUC	Accuracy
Zhang et al., (2020) <sup>36</sup>	Lending club P2P U.S. A	Online integrated credit scoring model (OICSM). OICSM integrates gradient boosting decision tree (GBDT) and neural network (NN).	0.72	
Ahelegbey and Giudici (2023) <sup>37</sup>	SMEs P2P lending across Europe	Factor clustering-based approach segmenting a heterogeneous population into groups with more homogeneous characteristics.	0.82	
Lyócsa et al., (2022) <sup>38</sup>	Lending club P2P U.S.A	Neural Network Classification	N/R	0.72
An-Hsing et al., (2022) <sup>39</sup>	Lending Club P2P U.S. A	2-Layer Neural network	N/R	0.87
Hybrid model	Lending club P2P U.S. A	Hybrid (1DCNN- <i>XGBoost</i> )	<b>0.98</b>	<b>0.96</b>

**Table 4.** Comparative performance of ML models on P2P lending datasets. Note: Most of the included studies compared more than one ML model, and we reported the model with the highest AUC or accuracy scores.

Discussion of results from the experiment

Table 2 below presents the results for our first hypothesis, which tests the effectiveness of hybrid model, the 1DCNN, logistic regression and *XGBoost* models in predicting default. The null hypothesis is that the hybrid model does not significantly improve the accuracy of credit risk predictions compared to traditional credit scoring models. We used the macro and weighted average scores to summarise individual model performances across all classes. While the macro average assesses the model's ability to predict each class with equal importance to ensure the model performs well across all classes, including the minority ones. Conversely, the weighted average calculates the metric for each class, weighting it by the class's prevalence or support (the number of true instances for each class). Table 2 suggests that the two metrics are the same for the 1DCNN and the hybrid model – 0.96. However, the classification accuracy for the models indicates that the hybrid model slightly outperforms the conventional models, rejecting the null hypothesis of no improvement in classification accuracy with the hybrid approach. Specifically, the hybrid model achieved a classification accuracy of 96%, compared to 94% for the 1DCNN, 94% for the *XGBoost* and 92% for the logistic regression model.

This 2% enhancement in classification accuracy with the hybrid model is significant in the context of credit risk assessment. In such models, a 2% increase in accuracy can significantly impact the precision of risk evaluation, potentially leading to more reliable identification of defaulters versus non-defaulters. This improvement directly

translates to reduced financial risk and losses for lending institutions by enabling more informed lending decisions, highlighting the value of even seemingly modest gains in model performance.

Additionally, the improvements in performance metrics, detailed in Table 2, support the decision to combine CNN and XGBoost into a hybrid model. We saw improvements in classification accuracy (0.98 vs. 0.94), precision for the non-default class (0.94 vs. 0.92), recall for the default class (0.94 vs. 0.91), and F-score for both classes. This synergy improves the overall accuracy and enhances the model's ability to generalise across diverse datasets.

Figure 5 above shows the AUC scores for the three models. We found the AUC curve for the *XGBoost* model is slightly higher at (0.99), than the Hybrid (0.98) and the logistic regression (0.95). The *H - score*, which adjusts for the distribution of class probabilities and the cost of misclassifications, being 0.95 for all the models except the 1DCNN suggest that the models are highly effective at balancing the trade-offs between false positives and false negatives. This indicates that the models predict accurately and does so in a way that align with the practical cost considerations of misclassifying instances. Therefore, we can conclude that the hybrid model effectively differentiates between clients who will default and those who will not, with a strong emphasis on minimising costly errors (like incorrectly identifying a good client as a risk or vice versa).

To evaluate the effectiveness of the hybrid model, we applied it to the German credit dataset, originally used by Hofmann in 1994, allowing us to benchmark its classification accuracy against the results from other studies that also utilised this dataset. Table 3 shows the performance evaluation for the baseline models, as reported by<sup>31</sup>, and the performance of the hybrid model. We assessed model performance using accuracy and precision as metrics. The results indicated that the hybrid model performed better than the traditional models on the German Credit dataset.

Additionally, we assessed the performance of different ML models on the Peer-to-Peer (P2P) dataset and compared these results to the hybrid model's performance, providing a contextual performance evaluation. Table 4 is the performance data from a selection of these studies. The proposed model demonstrated superior performance, achieving an AUC of 0.98 and an accuracy of 0.96. However, variations in the configurations of the P2P datasets might influence these performance metrics. For example, two studies using the Lending Club dataset could report different results due to significant differences in the selection of input features and the number of observations included in the analysis. We also explored the models' explainability- the second hypothesis. Using Global explainability, we sought to understand how the 1DCNN, *XGBoost* and *LR* models prioritise the importance of different features when assessing credit risk. The findings from our experiments provided clear evidence to reject the null hypothesis that all the models would similarly explain credit risk decisions.

Figures 6, 7 and 8 below presents the Global explainability for the models. Global explainability in machine learning-based credit scoring models is critically important because of the high-stakes nature of financial decision-making and the increasing regulatory scrutiny on automated decision systems. All features are shown in the order of global feature importance, the first feature is the most important, and the last is the least important feature. Figures 6, 7 and 8 reveal a significant discrepancy in how these models rank the importance of various features. For example, the 1DCNN and *XGBoost* places the *interest rate* as the fifth most critical feature, whereas the logistic regression (LR) model ranks it much lower at the fourteenth position. This variation in feature importance ranking, especially concerning the *interest rate*, highlights a fundamental difference in how these models perceive its influence on default prediction. How models prioritise certain features over others can affect decision-making, especially in tailoring financial products and managing risk. A model that significantly underestimates or overestimates the importance of features like interest rates may lead to suboptimal lending decisions, affecting both the lender's risk exposure and the borrower's financial well-being. While the 1DCNN and *XGBoost* models appear to rank features similarly, *XGBoost* classified half of the features as dummy players—i.e., their inclusion or exclusion does not change the model's predictions. In these instances, their Shapley values are low or zero, indicating little to no importance.

In addition, we evaluated how much each feature has contributed to an individual targeted prediction (local explainability). In other words, a credit risk analyst who wishes to understand the contribution of each input feature on the output of the model for a single observation (customer) can use the SHAP framework to achieve this. Figures 9 and 10 below shows the Shapley plot for two customers. A negative Shapley value for a feature regarding a specific prediction means that the feature contributes to shifting the prediction away from the "default" class towards the "no default" class, compared to the average prediction of the model. The bar length shows the SHAP value for each feature for the customers (observation). The blue bar shows a negative effect, while the red bar shows a positive effect. The absolute SHAP value reveals the feature with the most significant effect in predicting whether this customer will default or not. The relevance of local explainability becomes evident when we compare the predictions for two customers in Figs. 9 and 10. For both customers, the model identified "*remaining outstanding principal*" as a crucial factor contributing to a lower probability of default. This indicates that as the remaining principal decreases, the likelihood of the customer defaulting on their loan diminishes, according to the model's learned patterns. This finding aligns with financial logic—customers with less debt remaining might be less likely to default, possibly due to a shorter remaining loan term or their proven ability to make consistent payments. However, the model's interpretation of "*interest received to date*" shows differences between the two customers. For customer 2, the amount of interest received so far is associated with a lower default probability than the model's average prediction. This could indicate that customer 2 has been consistent in their payments, resulting in a substantial interest payment, which the model interprets as a positive indicator of financial stability and reduces the predicted risk of default.

On the other hand, for customer 1, the same feature—"*interest received to date*"—increases the probability of default, contrary to the effect observed for customer 2. This suggests that, for customer 1, the accrued interest might signal a different underlying issue, such as the customer having paid interest without significantly reducing the principal, possibly indicating financial strain or reliance on minimal payments. Alternatively, this could



**Fig. 6.** Features ranking 1DCNN model.

reflect a different loan structure or payment history the model has learned to associate with higher risk. This contrast in how the same feature affects predictions for different customers highlights the importance of local explainability. It allows us to understand the specific context in which the model operates for each individual, providing insights that are not apparent from global model behaviour alone. For lenders, this means they can make more informed decisions, offering tailored advice or interventions based on the unique financial situation of each customer. Local explainability ensures transparency for customers, helping them understand why certain decisions were made, which can build trust in the lending platform and its assessment process.





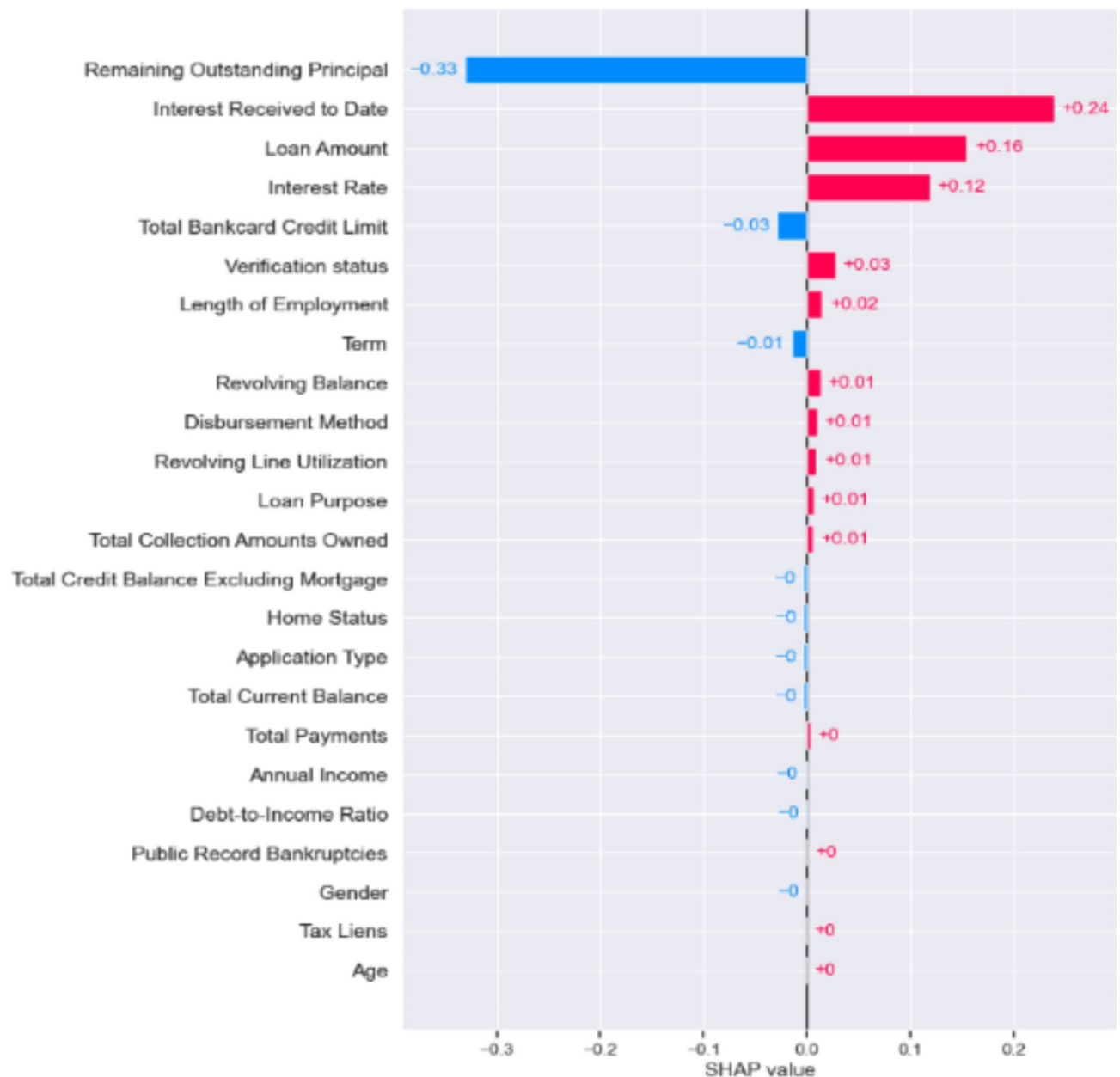
**Fig. 7.** Features ranking Logistic regression model.

Achieving local explainability is central in reducing bias in machine learning models<sup>32</sup> because it allows us to see the variables driving the model's decision. It plays a crucial role in reducing bias in lending, particularly regarding potentially discriminatory features like age and gender. Ensuring fairness in machine learning models, especially those used in sensitive applications such as credit risk assessment, is important. Biases in these models can lead to unfair outcomes, disproportionately affecting certain demographic groups and perpetuating systemic inequalities. The SHAP values in our model showed that for customer 1, both age and gender had minimal impact on the decision, with SHAP values of -0 and +0, respectively. For customer 2, age had a similarly low



**Fig. 8.** Features ranking *XGBoost* model

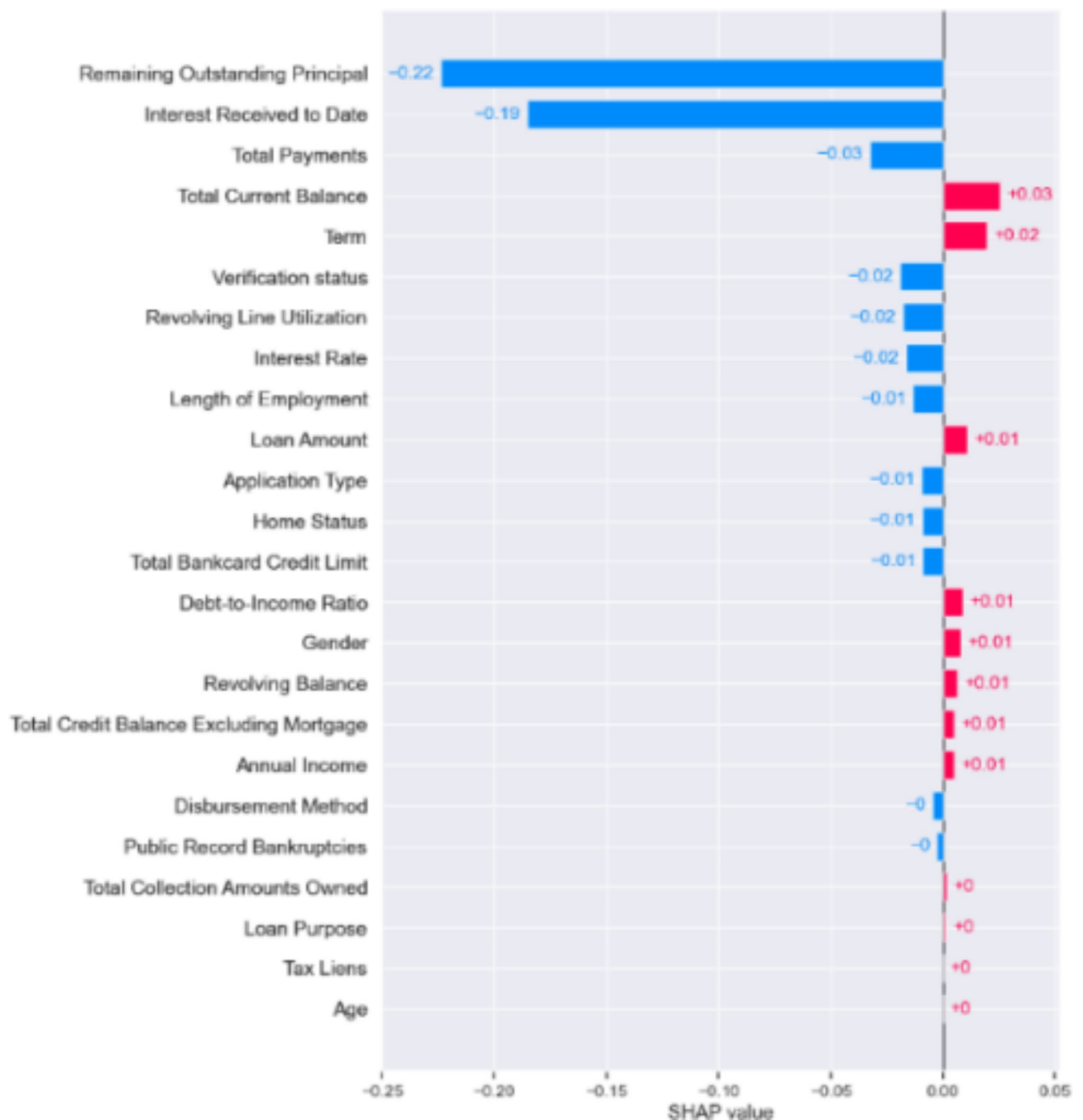
SHAP value of +0, and gender was slightly higher at +0.01 but still ranked very low in significance, placing tenth from the bottom in terms of influence. The low SHAP values for age and gender indicate that the model does not heavily rely on these features when predicting the likelihood of default for these customers. This is a positive indicator of fairness in the model, suggesting that discriminatory factors do not disproportionately influence it. Furthermore, by providing these granular details, local explainability enables model developers and auditors to take corrective actions if bias is detected. For instance, if age or gender were found to have a disproportionately



**Fig. 9.** 1DCNN local explainability (customer 1).

high impact on predictions for some customers, it would signal the need for further investigation and potential re-training of the model, removing these two features to ensure that it does not unfairly penalise certain groups.

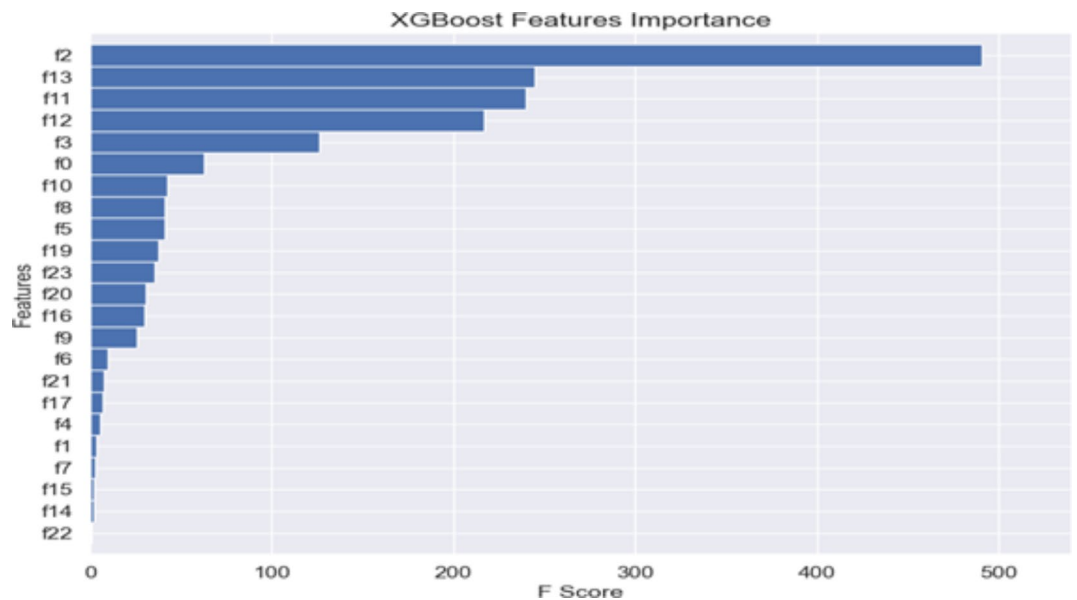
In our final analysis, we examined whether removing discriminatory features, specifically gender and age, from the hybrid model would affect its ability to classify outcomes accurately. This line of enquiry is essential in light of the U.S. Equal Credit Opportunity Act (ECOA). The ECOA is designed to regulate all entities that participate in credit decisions in their ordinary course of business, such as banks, retailers, credit unions and credit card issuers<sup>33</sup>. The Act, a Federal statute (15 U.S.C. § 1691), prohibits creditors from discriminating in any aspect of a credit transaction based on an applicant's race, colour, religion, national origin, sex, marital status or age (if the applicant is legally capable of entering a contract). An equivalent Act in the U.K., is the Equality Act 2010, which replaced the previous discrimination legislation in Great Britain. The Act prohibits discrimination relating to sex, race, sexual orientation, religion or belief and age in the public and wider society<sup>34</sup>. The findings presented in Table 5 below, indicate that the hybrid model's classification accuracy remains essentially unchanged (0.96) without these two variables. This suggests that gender and age do not play a significant role in the model's predictions—see Figs. 6, 7, 8, 9 and 10, akin to “dummy players” that do not contribute to the outcome. Removing features like gender and age helps ensure that the model's assessments are based on relevant financial behaviours rather than characteristics that could lead to unfair discrimination. This approach aligns with ethical guidelines and regulatory standards, promoting fairness and equality in credit decision making.



**Fig. 10.** 1DCNN local explainability (customer 2).

Hybrid 1DCNNs-XGBoost						
	Precision	Recall	F1-Score	Accuracy	AUC	H – Score
0	0.94	0.98	0.96	0.96	0.98	0.95
1	0.97	0.94	0.95			
Macro Avg.	0.96	0.96	0.96			
Weighted Avg	0.96	0.96	0.96			

**Table 5.** Performance evaluation metrics hybrid model without gender and age.



**Fig. 11.** *XGBoost* features importance ranking. **Notes:** F0 denotes the Loan Amount, F1 Loan Term, F2 Interest Rate, F3 Length of Employment, F4 Home Status, F5 Annual Income, F6 Verification Status, F7 Loan Purpose, F8 Debt-to-Income Ratio, F9 Revolving Balance, F10 Revolving Line Utilization, F11 Remaining Outstanding Principal, F12 Total Payments, F13 Interest Received to Date, F14 Application Type, F15 Total Collection Amounts Owned, F16 Total Current Balance, F17 Public Record Bankruptcies, F18 Tax Liens, F19 Total Credit Balance Excluding Mortgage, F20 Total Bankcard Credit Limit, F21 Disbursement Method, F22 Gender, F23 Age.

Again, to test the robustness of the Shapely algorithm in ranking features importance, we used *XGBoost*'s built-in feature importance functionality. We aimed to assess the contributions of age and gender in the classification of default and no default in our dataset further. Figure 11 shows that the built-in features ranking supports our earlier findings that age and gender do not contribute significantly to the classification outcomes. Specifically, gender, F22, had zero contributions, while age, F23, did not make it to the first ten features as in the IDCNN model.

While *XGBoost*'s built-in features importance provides a quick way to gauge the relative importance of features, it is crucial to remember that this method is biased towards numerical features and features with a larger number of categories<sup>35</sup>. Since it measures importance based only on the specific structure of trees in *XGBoost*, its insights might not fully capture feature interactions or dependencies that model-agnostic methods like SHAP could reveal. In addition, the feature importance metrics provided by *XGBoost*, such as weight, gain, and cover, are global measures. That is, these metrics are aggregated across all trees within the model and provide only overall features importance without showing how features affect individual predictions. SHAP, on the other hand assigns each feature an importance value for a particular prediction allowing us to understand how each feature in the input contributes to the final decision made by the model.

## Conclusion

This paper uses a novel architecture to address the “black box” nature of the ML-powered credit scoring model. Despite their predictive ability, they lack interpretability, which could hinder trust and regulatory compliance, especially in the financially sensitive area of credit risk analysis. To achieve our aim, we compared the classification accuracies and features extraction capabilities of a one-dimensional convolutional neural network (1DCNN), a novel hybrid model, *XGBoost* and logistic regression, on a peer-to-peer (P2P) consumer credit dataset. Our analysis demonstrated that the hybrid model outperformed the conventional models regarding classification accuracy. We achieved a classification accuracy of 96% for the hybrid model, surpassing the 1DCNN's 94%, *XGBoost*'s 94% and logistic regression's 92%. Additionally, using the Shapley framework enhanced the transparency and interpretability of our model, enabling us to pinpoint and mitigate likely biases effectively. The implications of our work extend beyond academic discourse, offering tangible benefits for the credit industry by providing credit analysts with a tool for assessing loan default risks, positively leading to better-informed lending decisions. The hybrid model's ability to handle features extraction and complex interactions makes it reliable in data-intensive financial environments. Our findings that removing age and gender from the hybrid model has a non-significant impact on predictive accuracy suggest that achieving fairness without compromising the model's performance is possible. This is an encouraging result, indicating that models can be designed to reduce reliance on discriminatory features while still maintaining high predictive accuracy. However, a limitation of this study is that it primarily focuses on a specific dataset and context. Future research could explore applying this hybrid model to other lending scenarios, such as corporate or consumer lending, using diverse datasets. Additionally, future studies might investigate integrating various machine-learning techniques to identify effective strategies



for balancing classification accuracy with interpretability and fairness. This broader exploration could further enhance our understanding of how to build fair and accurate models across different financial contexts.

## Appendix

### Data availability

No datasets were generated or analysed during the current study.

Received: 28 April 2024; Accepted: 1 October 2024

Published online: 24 October 2024

## References

- Gramegna, A. & Giudici, P. SHAP and LIME: an evaluation of discriminative power in credit risk. *Front. Artif. Intell.* **4** (2021). Available at <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.752558/full>. Accessed on 11/10/2024.
- Babaei, G., Giudici, P. & Raffinetti, E. Explainable artificial intelligence for crypto asset allocation. *Finance Res. Lett.* **47**, 102941 (2022).
- Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* (2017). Available at <https://arxiv.org/pdf/1705.07874>. Accessed on 11/10/2024.
- Zhu, B., Yang, W., Wang, H. & Yuan, Y. A hybrid deep learning model for consumer credit scoring. *IEEE*. <https://doi.org/10.1109/icaib.2018.8396195> (2018).
- Li, Y., Lin, X., Wang, X., Shen, F. & Gong, Z. Credit risk assessment algorithm using deep neural networks with clustering and merging. doi: (2017). <https://doi.org/10.1109/cis.2017.00045>
- Tripathi, D., Edla, D. R. & Cheruku, R. Hybrid credit scoring model using neighborhood rough set and multi-layer ensemble classification. *J. Intell. Fuzzy Syst.* **34**, 1543–1549 (2018).
- Suhadolnik, N., Ueyama, J. & Da Silva, S. Machine learning for enhanced credit risk assessment: an empirical approach. *J. Risk Financial Manage.* **16**, 496 (2023). Available at <https://www.mdpi.com/1911-8074/16/12/496>. Accessed on 11/10/2024.
- Nwafor, C. N. & Nwafor, O. Z. Determinants of non-performing loans: an explainable ensemble and deep neural network approach. *Finance Res. Lett.* **56**, 104084 (2023).
- Dumitrescu, E., Hué, S., Hurlin, C. & Tokpavi, S. Machine learning or econometrics for credit scoring: Let's get the best of both worlds. *Social Sci. Res. Netw.* <https://doi.org/10.2139/ssrn.3553781> (2020).
- Shen, F., Zhao, X., Kou, G. & Alsaadi, F. E. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. *Appl. Soft Comput.* **98**, 106852 (2021).
- Petch, J., Di, S. & Nelson, W. Opening the black box: the promise and limitations of explainable machine learning in Cardiology. *Can. J. Cardiol.* **38**, 204–213 (2022).
- Bussmann, N., Giudici, P., Marinelli, D. & Papenbrock, J. Explainable machine learning in credit risk management. *Comput. Econ.* **57**, 203–216 (2020).
- Shapley, L. S. 17. A value for N-Person games. *Princet. Univ. Press. eBooks*. 307–318. <https://doi.org/10.1515/9781400881970-018> (1953).
- Hayashi, Y. & Takano, N. One-dimensional convolutional neural networks with feature selection for highly concise rule extraction from Credit Scoring datasets with heterogeneous attributes. *Electronics*. **9**, 1318 (2020).
- Jabeur, S. B., Gharib, C., Mefteh-Wali, S. & Arfi, W. B. CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecast. Social Change/Technological Forecast. Social Change*. **166**, 120658 (2021).
- Silva, E. C. E., Lopes, I. C., Correia, A. & Faria, S. A logistic regression model for consumer default risk. *J. Appl. Stat.* **47**, 2879–2894 (2020).
- Wachter, S., Mittelstadt, B. & Russell, C. Bias Preservation in Machine Learning: the legality of fairness Metrics under EU Non-discrimination Law. *Social Sci. Res. Netw.* <https://doi.org/10.2139/ssrn.3792772> (2021).
- Huang, M. H. & Rust, R. T. A strategic framework for artificial intelligence in marketing. *J. Acad. Mark. Sci.* **49**, 30–50 (2020).
- Lambrecht, A. & Tucker, C. Algorithmic Bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Manage. Sci.* **65**, 2966–2981 (2019).
- Nwafor, O., Okafor, E., Aboushady, A. A., Nwafor, C. & Zhou, C. Explainable Artificial intelligence for prediction of non-technical losses in electricity distribution networks. *IEEE Access*. **11**, 73104–73115 (2023).
- Dastile, X. & Celik, T. Making deep learning-based predictions for credit scoring explainable. *IEEE Access*. **9**, 50426–50440 (2021).
- George, N. All Lending Club Loan Data. (2019). <https://www.kaggle.com/datasets/wordsforthewise/lending-club> Accessed on 12/05/2021.
- Bradshaw, T. J., Huemann, Z., Hu, J. & Rahmim, A. A guide to cross-validation for artificial intelligence in medical imaging. *Radiol. Artif. Intell.* **5**, e220232 (2023).
- Feurer, M. & Hutter, F. Hyperparameter Optimization. in *The  $\alpha$ Springer series on challenges in machine learning* 3–33 doi: (2019). [https://doi.org/10.1007/978-3-030-05318-5\\_1](https://doi.org/10.1007/978-3-030-05318-5_1)
- Shetty, A. M., Aljunid, M. F., Manjaiah, D. H. & Afzal, A. M. S. Hyperparameter optimization of machine learning models using grid search for Amazon Review sentiment analysis. *Lecture Notes Networks Syst.* 451–474. [https://doi.org/10.1007/978-981-99-7814-4\\_36](https://doi.org/10.1007/978-981-99-7814-4_36) (2024).
- Jabeur, S. B., Sadaoui, A., Sghaier, A. & Aloui, R. Machine learning models and cost-sensitive decision trees for bond rating prediction. *J. Oper. Res. Soc.* **71**, 1161–1179 (2019).
- Zou, M., Jiang, W. G., Qin, Q. H., Liu, Y. C. & Li, M. L. Optimized XGBoost Model with small dataset for Predicting relative density of Ti-6Al-4V Parts manufactured by selective laser melting. *Materials*. **15**, 5298 (2022).
- Szegedy, C. et al. Going deeper with convolutions. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1409.4842> (2015).
- Khosa, C. K., Mars, L., Richards, J. & Sanz, V. Convolutional neural networks for direct detection of dark matter. *J. Phys. G Nuclear Part. Phys.* **47**, 095201 (2020).
- Hand, D. J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach. Learn.* **77**, 103–123 (2009).
- Hofmann, H. Statlog (German Credit Data). *UCI Mach. Learn. Repository*. <https://doi.org/10.24432/C5NC77> (1994).
- Hellström, T., Dignum, V. & Bensch, S. Bias in Machine Learning -- what is it good for? *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2004.00686> (2020).
- United States Code (U.S.C.). Title 15 - COMMERCE AND TRADE. (2011). <https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm>
- Equality Act : guidance. GOV.UK (2010). <https://www.gov.uk/guidance/equality-act-2010-guidance>. (2015).
- Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Science & Business Media, 2013).

36. Zhang, Z., Niu, K. & Liu, Y. A deep learning based online credit scoring model for P2P lending. *IEEE Access*. **8**, 177307–177317 (2020).
37. Ahelegbey, D. F., Giudici, P. & Hadji-Misheva, B. Factorial network models to improve P2P credit risk management. *Front. Artif. Intell.* **2**, 456584 (2019). Available on <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2019.00008/full>. Accessed on 11/10/2024.
38. Lyócsa, Š., Vašaničová, P., Misheva, B. H. & Vateha, M. D. Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. *Financial Innov.* **8**, 32 (2022).
39. Chang, A. H., Yang, L. K., Tsaih, R. H. & Lin, S. K. Machine learning and artificial neural networks to construct P2P lending credit-scoring model: a case using Lending Club data. *Quant. Finance Econ.* **6**, 303–325 (2022).

## Author contributions

C and S wrote the main manuscript text. O and C prepared the figures and run the Python code. All the authors reviewed the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-75026-8>.

**Correspondence** and requests for materials should be addressed to C.N.N.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024