Explainable Multistage Ensemble 1D Convolutional Neural Network for Trust Worthy Credit Decision

Pavitha N[1], Shounak Sugave[2]
PhD Research Scholar[1], Associate Professor[2]
Department of Computer Engineering and Technology[1, 2]
Dr. Vishwanath Karad MIT World Peace University Pune, Maharashtra, India[1, 2]

Abstract—**Banking is a dynamic industry that places significant importance on risk management, requiring accurate and interpretable AI models to make transparent lending decisions. This study introduces a groundbreaking approach that combines a multistage ensemble technique with a 1D convolutional neural network (CNN) architecture. The algorithm not only delivers superior classification performance but also offers interpretable explanations for its decisions. The algorithm is designed with multiple strategic steps to enhance model performance without sacrificing explainability. Thorough experiments were conducted using datasets from private banks and non-banking financial companies (NBFCs) in India to evaluate the algorithm's performance. It was compared against various state-of-the-art models, demonstrating remarkable precision, recall, F1 score, and accuracy values of 0.994, 0.992, 0.993, and 0.991, respectively. This outperformed competing models like homogeneous deep ensembles, 1D CNN, and Artificial Neural Networks (ANN). Furthermore, individual borrower dataset evaluations confirmed the proposed algorithm's consistency and efficiency, achieving precision, recall, F1 score, and accuracy values of 0.960, 0.961, 0.952, and 0.964, respectively. The research emphasizes the effectiveness of the explanatory integration decision process, wherein the Explainable Multistage Ensemble 1D CNN not only provides enhanced credit risk prediction but also facilitates transparent and interpretable lending decisions. The algorithm's ability to offer understandable explanations empowers financial institutions to make well-informed lending decisions, reduce credit risk, and foster a more stable and inclusive financial ecosystem.**

Keywords—Credit risk prediction; explainable AI; multistage ensemble; 1D convolutional neural network; interpretability; transparency; lending decisions; financial institutions

I. Introduction

In the realm of financial services, credit risk prediction plays a crucial role in enabling sound and responsible lending decisions [1], [2]. Accurate assessments of borrowers' creditworthiness are essential for financial institutions to mitigate risks, ensure fair lending practices, and maintain a stable financial ecosystem. With the advancements in Machine Learning (ML) and artificial intelligence (AI), there has been a surge in the development and adoption of advanced predictive models for credit risk assessment [3], [4]. These models, such as convolutional neural networks (CNNs) and artificial neural networks (ANNs), offer the ability to capture intricate patterns and dependencies within the data, leading to improved predictive accuracy. Despite their impressive performance, the use of complex AI models in the financial industry raises concerns about their inherent opacity and lack of interpretability. Often referred to as "black box" models, these approaches provide little insight into the factors that influence their decisions [5]. In highly regulated and sensitive domains like credit risk assessment, the lack of transparency can be a major obstacle, as stakeholders, including customers, regulators, and internal compliance teams, require explanations to trust and validate the model's decisions [6], [7].

To address these challenges and bridge the gap between predictive accuracy and interpretability, there has been a growing interest in explainable AI (XAI). XAI techniques aim to provide interpretable explanations for complex models, allowing stakeholders to understand how decisions are made and identify the key features driving predictions. In the context of credit risk prediction, XAI offers several advantages, including increased transparency, regulatory compliance, enhanced customer trust, and the ability to detect potential biases in decision-making [8], [9].

In this study, we introduce a technique aiming to incorporate explanations into the decision-making process. Our model utilizes multistage ensemble techniques, known for enhancing interpretability by offering explanations at various decision stages. By combining the strengths of multiple models, this ensemble approach improves predictive accuracy while retaining the capability to provide meaningful explanations for credit risk evaluations.

The proposed model's objective is to achieve high predictive accuracy while ensuring that the underlying decision process is transparent and understandable. By providing interpretable explanations, financial institutions can gain valuable insights into the model's risk assessments, understand the relative importance of different features, and detect potential biases, ultimately leading to more informed lending decisions.

To proposed Explainable Multistage Ensemble 1D CNN model, is evaluated for the effectiveness on two different data sets. We conducted comprehensive experiments on enterprise credit risk dataset and individual borrower credit dataset. We compared the performance of proposed model with other state-of-the-art approaches, including Homogeneous deep Ensembles (ANN and CNN), as well as standalone ANN and 1D CNN classifiers. The results demonstrate the superior predictive accuracy and interpretability of our proposed model, reinforcing its potential value in the domain of credit risk prediction.

The rest of this paper is structured as follows: In Section II, covers a comprehensive review of related studies in the fields of credit risk prediction, explainable AI, and ensemble techniques. Section III presents the methodology and architecture of our innovative model. Following that, in Section IV, we present and analyze the experimental results, highlighting the strengths of our approach compared to other existing methods. Finally, in Section V, we conclude the paper and discuss potential avenues for future research.

II. Literature Review

Assessing credit risk is a crucial undertaking in the financial sector, as it involves evaluating borrowers' creditworthiness to make well-informed lending choices. To develop predictive models for credit risk assessment, researchers have explored various AI and machine learning approaches over time. Furthermore, the growing need for transparency and interpretability in models has given rise to explainable AI (XAI) techniques, which aim to provide insights into the decision-making process of intricate models. In this section, we present a thorough examination of pertinent literature concerning credit risk prediction, explainable AI, and ensemble techniques.

A. Credit Risk Prediction

The literature on credit risk prediction is vast and diverse, with numerous studies focusing on developing accurate and reliable models. Traditional credit scoring methods, such as logistic regression and decision trees, have long been used in the industry. However, with advancements in machine learning, more sophisticated models, including neural networks, support vector machines (SVM), and gradient boosting algorithms, have gained popularity due to their ability to capture complex patterns in credit data [4], [10], [11], [12].

B. Explainable AI for Credit Risk Prediction

The need for model transparency and interpretability in credit risk prediction has led to the exploration of explainable AI techniques. Several studies have proposed methods to generate explanations for credit risk models, enabling stakeholders to understand the rationale behind model decisions. Approaches such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) have been applied to credit risk models to provide local and global interpretations [9], [13], [14]. Additionally, rule-based models and decision trees have been used as interpretable alternatives to black box models [15], [16].

C. Ensemble Techniques

Ensemble methods have demonstrated remarkable success in various machine learning tasks, including credit risk prediction. Ensemble approaches combine the predictions of multiple models to improve overall performance and robustness. Bagging and boosting techniques, such as Random Forest and Gradient Boosting Machines (GBM), have been widely employed in credit risk prediction [17]. Recent research has explored the benefits of using homogeneous and heterogeneous ensembles, where models from the same or different algorithm families are combined [8], [18], [19], [20], [21]. Researchers from various domains proved the effectiveness of ensemble techniques in their fields [22], [23], [24], [25], [26], [27], [28], [29], [30].

D. Multistage Ensemble Techniques

Multistage ensemble techniques offer a promising approach for improving both predictive accuracy and model interpretability. By combining multiple models at different stages of the decision-making process, these methods can provide valuable insights into the model's reasoning. Various studies have shown that multistage ensembles can outperform single model [22], [25], [30]. However, the application of multistage ensemble techniques in credit risk prediction remains relatively unexplored.

Credit risk prediction is a critical domain where model accuracy, transparency, and interpretability are of utmost importance. While various AI models have been employed for credit risk assessment, the emergence of explainable AI and ensemble techniques presents new opportunities to enhance predictive performance and provide interpretable explanations for model decisions. The proposed approach aims to leverage explainable multistage ensemble techniques to address the dual objectives of predictive accuracy and model transparency.

III. Proposed Methodology

A. Credit Risk Dataset used in the Experiments

The analysis focuses on two distinct borrower segments: individual borrowers and enterprise borrowers. Individual borrowers obtain loans in their personal capacity, while enterprise borrowers secure loans on behalf of their businesses. Data for both segments were collected under a Non-Disclosure Agreement (NDA). The individual borrower dataset was obtained from a private bank in India and comprises 105,163 records, with 100,497 records (95.6%) falling into the negative class (non-risky) and 4,665 records (4.4%) classified as positive class (risky). The enterprise dataset was collected from an NBFC (Enterprise) in India and consists of 97,451 records, with 92,900 classified (95.3%) as the negative class and 4,550 (4.7) as the positive class. The sample description is presented in Fig. 1. The target variable in this analysis is "risk," which is binary, and the other variables serve as independent variables. The dataset contains a mix of categorical and numerical variables. All data used in this analysis were collected following ethical guidelines and legal agreements to ensure confidentiality and privacy.

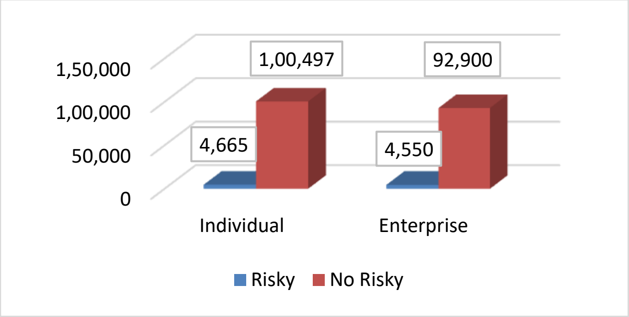

Fig. 1. Sample profile.

B. Multistage Ensemble Architecture Overview

The multistage ensemble classifier is designed to improve classification performance by employing a series of stages, each containing a 1D Convolutional Neural Network (CNN). Unlike a single CNN model, the ensemble classifier combines the outputs of all stages to make the final prediction. Each stage contributes its specialized knowledge to enhance the overall decision-making process. The choice of the number of stages will depend on the complexity of the classification task, the size of the dataset, and the available computational resources. Determining the appropriate number of stages for the ensemble classifier is crucial. Too few stages might limit the model's representational power, while too many stages could lead to excessive computational requirements and potential overfitting. The optimal number of stages is typically determined through experimentation and performance evaluation on the validation set. In the proposed setup five stages are used.

C. Stage-wise CNN Architecture

For each stage, design a 1D CNN architecture tailored to the specific characteristics of the dataset and classification problem. Each stage's CNN should consist of multiple convolutional layers, followed by activation functions and pooling layers. This design enables the CNN to learn hierarchical features from the 1D input data. Experimentation is done with different filter sizes, strides, and the number of filters in each layer to identify the configuration that yields the best results. Additionally, techniques like batch normalization are applied to accelerate training and improve convergence. To reduce overfitting, introduced dropout layers, which randomly deactivate neurons during training, preventing reliance on any single set of features. Table I represents the architecture of a 1D CNN-based ensemble classifier with multiple stages, where each stage consists of Conv1D layers, Batch Normalization, Max Pooling, LeakyReLU activation, and finally, an Average Pooling, Dropout, and Dense layer for classification.

TABLE I. ARCHITECTURE OF A 1D CNN-BASED ENSEMBLE CLASSIFIER WITH MULTIPLE STAGES

Type of Layer	Other Parameters
Conv1D **Batch Normalization** **Max Pooling 1D** **Activation**	Filters = 128 - Pool size = 4 LeakyReLU activation
Conv1D **Batch Normalization** **Max Pooling 1D** **Activation**	Filters = 256 - Pool size = 4 LeakyReLU activation
Conv1D **Batch Normalization** **Max Pooling 1D** **Activation**	Filters = 256 - Pool size = 4 LeakyReLU activation
Conv1D **Batch Normalization** **Max Pooling 1D** **Activation**	Filters = 512 - Pool size = 4 LeakyReLU activation
Average Pooling 1D **Flatten** **Dropout** **Dense**	Pool size = 2 - Rate 0.4 Regularizer L2 (0.001), Softmax activation

D. Conv1D Layer

This layer performs 1-dimensional convolution on the input data. The "Filters" parameter is set to 128 for the first Conv1D layer, 256 for the second and third Conv1D layers, and 512 for the fourth Conv1D layer. The number of filters determines the number of features maps the layer will learn. Higher filter values allow the model to learn more complex patterns but also increase computational complexity.

E. Batch Normalization Layer

Batch normalization normalizes the input of the layer, helping to stabilize and accelerate the training process. It improves convergence and prevents internal covariate shift, which occurs when the distribution of inputs to a layer change during training.

F. Max Pooling 1D Layer

Spatial dimensions of the data can be reduced by using max pooling while retaining the most important features. The "Pool size" parameter is set to 4 for all Max Pooling layers. This means the layer will take the maximum value within a sliding window of size 4 along the temporal dimension.

G. LeakyReLU Activation

Leaky ReLU (Rectified Linear Unit) is an activation function that introduces a small negative slope for negative input values, preventing the "dying ReLU" problem. The negative slope helps the model during backpropagation even for negative inputs, leading to improved gradient flow and avoiding potential dead neurons.

H. Average Pooling 1D Layer

After the last Conv1D layer, an Average Pooling layer is used instead of Max Pooling. Average pooling computes the average value of each feature map, reducing the data dimensionality and providing a global summary of the features.

I. Flatten Layer

The Flatten layer converts the 3-dimensional output from the previous layers into a 1-dimensional vector, preparing it for the fully connected layers.

J. Dropout Layer

Dropout is a technique that randomly drops out (sets to zero) a fraction of the neurons during training. The "Rate" parameter is set to 0.4, meaning 40% of the neurons will be dropped out during training. This helps prevent overfitting and encourages the model to learn more robust representations.

K. Dense Layer

The Dense layer is a fully connected layer, linking each neuron from the preceding layer to every neuron in this current layer. With a "Pool size" parameter of 2, this layer comprises 2 output neurons. To prevent overfitting, the layer employs L2 regularization with a coefficient of 0.001, penalizing large weights. By using the Softmax activation function, the final output values are transformed into probability scores for each class, making it suitable for multi-class classification.

L. Training the Stage-wise CNNs and Ensemble Combination

In the proposed approach, we utilize a multi-stage Convolutional Neural Network (CNN) architecture, where each stage's CNN is trained independently on the training set. Throughout the training process, we closely monitor the model's performance on the validation set to prevent overfitting. To achieve optimal results, we tune hyperparameters like learning rate, batch size, and the number of epochs. After training all stages' CNNs, we proceed to combine their outputs to create the final prediction. To evaluate the effectiveness of our ensemble classifier, we employ the test set and calculate various standard metrics, including accuracy, precision, recall, F1-score, and confusion matrices. Subsequently, we conduct a comparative analysis to assess how our proposed ensemble classifier performs in comparison to other state-of-the-art techniques. By doing so, we aim to demonstrate the superiority of our approach in making accurate predictions.

M. Explanation Generation

Adaptive Relevance Scaling for Layer-wise Relevance Propagation (ARSLRP) based explanations are employed to interpret the decision-making process of the multistage 1D CNN based ensemble classifier. Its primary objective is to attribute the model's prediction to the input features, offering a human-understandable explanation for the decision outcomes. ARSLRP operates on the principle of redistributing the model's output back to its input features, layer by layer, to identify the most influential features contributing to the final prediction.

The ARSLRP process starts from the final layer of the network, where the relevance is initialized based on the model's output (e.g., for a classification problem, relevance is initialized for the predicted class). Then, the relevance is propagated backward through the network layers using the Alpha-Beta rule until it reaches the input layer. The relevance scores obtained after the ARSLRP process indicate the importance of each input feature in influencing the model's decision.

IV. Results and Discussion

Multi Stage Heterogeneous Ensemble 1D CNN The proposed multistage ensemble 1D CNN model is evaluated based on various performance matrices namely precision, recall, f1-score and accuracy on two datasets namely enterprise data set and individual dataset. Table II illustrates the results for various performance matrices on enterprise dataset.

TABLE II. Explainable Ensemble 1 D CNN Performance on NBFC Data

Algorithm	Precision	Recall	F1-score	Accuracy
Proposed Algorithm (Multistage 1 D CNN)	0.994	0.992	0.993	0.991
Homogeneous deep Ensemble (ANN)	0.910	0.950	0.930	0.950
Homogeneous deep Ensemble (CNN)	0.900	0.950	0.930	0.950
ANN	0.891	0.893	0.892	0.894
1 D CNN	0.893	0.891	0.894	0.892

1) Precision: Multistage 1D CNN model achieves a precision of 0.994. This means that when the model predicts a customer as being at risk of defaulting on their credit, it is correct 99.4% of the time. A high precision value indicates that the model is effective in minimizing false positives, i.e., it rarely misclassifies customers who are not likely to default as high-risk, which is essential for banks to avoid unnecessary precautionary measures for low-risk customers.

2) Recall: The Multistage 1D CNN model achieves a recall of 0.992, meaning it successfully captures 99.2% of the customers who are genuinely at risk of defaulting on their credit. A high recall value indicates that the model has a low false negative rate, meaning it rarely misses identifying customers who are actually high-risk. This is crucial for banks to ensure that they do not overlook customers who pose a real credit risk.

3) F1-score: For the Multistage 1D CNN model, the F1-score is 0.993, which indicates an excellent balance between precision and recall. It demonstrates that the model is effective in achieving both accurate positive predictions and comprehensive identification of high-risk customers. A high F1-score suggests that the model is well-suited for credit risk prediction tasks where precision and recall need to be balanced.

4) Accuracy: The Multistage 1D CNN model achieves an accuracy of 0.991, which means it correctly predicts approximately 99.1% of all instances in the dataset. This high accuracy indicates that the model performs exceptionally well in making overall accurate predictions, regardless of the class distribution. A high accuracy value shows the reliability and effectiveness of the model in capturing credit risk patterns and making informed decisions.

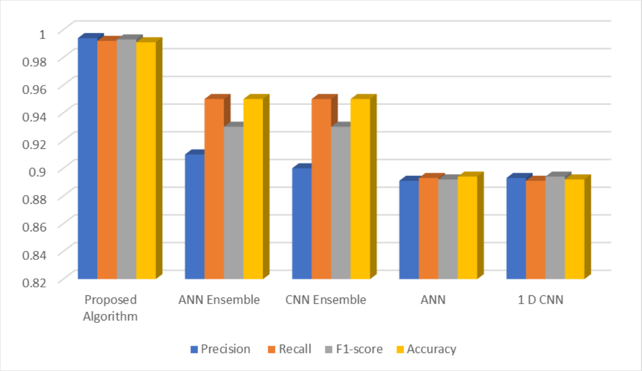

Fig. 2. Performance of proposed algorithm on NBFC dataset.

In summary, the results for the Multistage 1D CNN model in credit risk prediction are highly impressive. The model achieves exceptional precision, recall, F1-score, and accuracy values, demonstrating its ability to identify high-risk customers accurately while minimizing false predictions. This performance surpasses that of other algorithms tested in the study, making the Multistage 1D CNN model a promising choice for credit risk assessment tasks, and suggesting its potential for real-world implementation in financial institutions to enhance credit risk management and decision-making processes. Pictorial illustration for the same is shown in Fig. 2.

Similarly, Table III presents the performance metrics of various algorithms for individual borrower credit risk prediction dataset, with each row corresponding to a specific model. Among the algorithms tested, the "Proposed Algorithm" based on the Multistage 1D CNN stands out as the top-performing model across multiple evaluation metrics.

The Proposed Algorithm achieves a precision of 0.960, indicating that 96% of the predicted high-risk customers are genuinely at risk of defaulting on their credit. This demonstrates the model's effectiveness in minimizing false positives, ensuring that it correctly identifies most customers who pose an actual credit risk. Furthermore, the Recall for the Proposed Algorithm is 0.961, signifying that the model captures 96.1% of the actual high-risk customers present in the dataset. A high recall score suggests that the model has a low false negative rate, meaning it rarely misses identifying customers who are truly at risk of defaulting on their credit. This capability is crucial for financial institutions to avoid overlooking potential credit risks. The F1-score of 0.952 for the Proposed Algorithm reflects a balance between precision and recall, indicating a good overall performance. The F1-score is particularly valuable when there is an uneven distribution of classes in the dataset, making it a reliable measure for credit risk prediction tasks.

Lastly, the Proposed Algorithm achieves an accuracy of 0.964, implying that it makes accurate predictions for approximately 96.4% of all instances in the dataset. A high accuracy value indicates that the model's overall performance is strong, making it a reliable tool for credit risk assessment. In comparison, the other algorithms, including Homogeneous deep Ensemble (ANN), Homogeneous deep Ensemble (CNN), ANN, and 1D CNN, also show respectable results, but the Proposed Algorithm based on the Multistage 1D CNN consistently outperforms them across all metrics. In conclusion, the results demonstrate that the Multistage 1D CNN model proposed in the study is highly effective for credit risk prediction. Its balanced precision, recall, and F1-score, combined with its impressive accuracy, make it a promising approach for financial institutions seeking accurate and reliable credit risk assessment models. Pictorial illustration for the same is shown in Fig. 3.

B. Explanations / Interpretations

In this study, we explored the ARSLRP as an explainability technique for credit risk prediction models. The goal was to gain insights into how the model makes predictions and to provide transparent explanations to stakeholders, such as regulators, auditors, and customers, who need to understand the factors contributing to credit risk assessments. Fig. 4 and Fig. 5 illustrate the results generated by the model to provide explanations on enterprise and individual borrower dataset respectively.

TABLE III. EXPLAINABLE ENSEMBLE 1 D CNN ON INDIVIDUAL BORROWER DATASET

Algorithm	Precision	Recall	F1-score	Accuracy
Proposed Algorithm (Multistage 1 D CNN)	0.960	0.961	0.952	0.964
Homogeneous deep Ensemble(ANN)	0.918	0.952	0.939	0.958
Homogeneous deep Ensemble(CNN)	0.950	0.958	0.940	0.959
ANN	0.881	0.883	0.882	0.884
1 D CNN	0.883	0.881	0.884	0.882

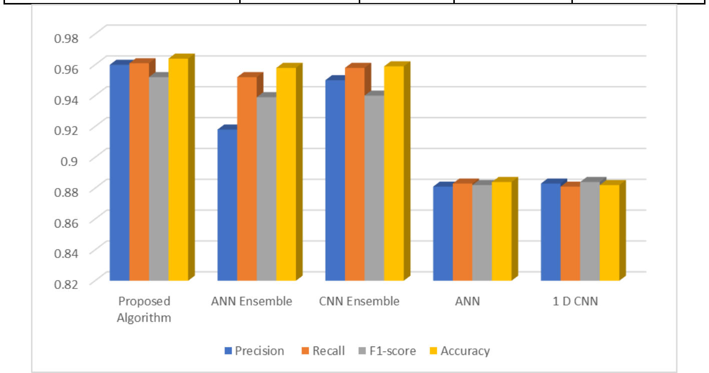

Fig. 3. Performance of proposed algorithm on individual borrower dataset.

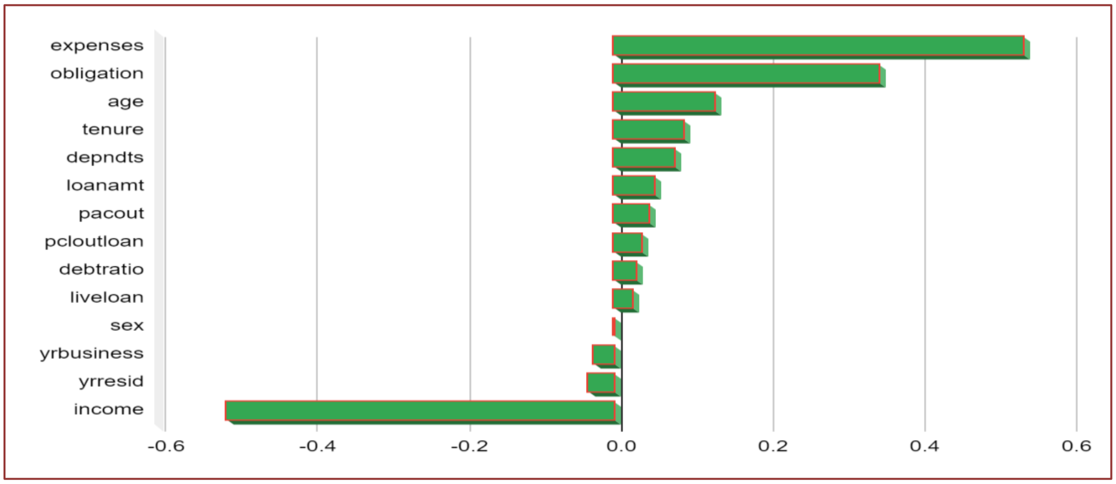

Fig. 4. Explanations / Interpretations on NBFC dataset.

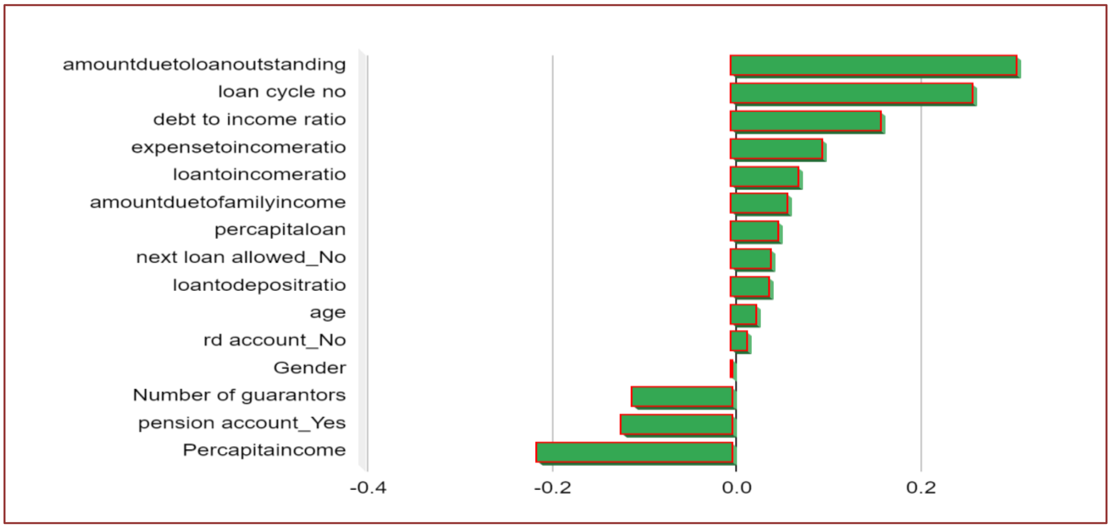

Fig. 5. Explanations / Interpretations on individual borrower dataset.

Fig. 4 illustrates the results generated by the model to provide explanations on enterprise (NBFC) dataset. The ARSLRP values for the NBFC dataset reveal key insights into the factors influencing the model's risk predictions. Notably, "Expenses" stands out as the most impactful variable with a substantial positive ARSLRP value of +0.5397. This implies that clients with higher reported family expenses are deemed riskier by the model. It suggests a correlation between increased spending and elevated risk in the context of the NBFC dataset. Similarly, "Obligation" contributes positively with an ARSLRP value of +0.3480, suggesting that clients with higher financial obligations are also considered riskier. These findings underscore the model's sensitivity to financial commitments and their association with increased risk. The positive contributions of demographic factors come into play with variables such as "Age" (+0.1320) and "Tenure" (+0.0915). The positive ARSLRP values indicate that older clients are perceived as riskier. Entrepreneurs with increased age may have higher family and social commitments as well as more expenses towards business and family. This could suggest that the model associates increased age and tenure with a higher likelihood of risk in the context of the dataset. The presence of dependents ("Dependents" with +0.0778) and certain loan-related variables, such as "Loan Amount" (+0.0521) and "Payout" (+0.0443), also contribute positively, indicating that clients with larger loan amounts, more dependents, and higher payouts are associated with increased risk according to the model. These associations in general reflect the model's perception of increased financial commitments and complexities contributing to higher default risk.

The variable "Sex" has a neutral impact with an ARSLRP value of 0.0000, suggesting that gender does not significantly contribute to the model's risk predictions. This implies that the model does not distinguish between male and female clients when assessing risk. On the negative side, variables such as "Year in Business" (-0.0258), "Year at Residency" (-0.0350), and "Income" (-0.5101) exert a negative influence on risk predictions. Longer durations in business and residency are associated with decreased risk, suggesting that recent businesses and residents are considered riskier by the model. The most impactful negative contributor, "Income," indicates that clients with higher incomes are perceived as less risky.

Fig. 5 illustrates the results generated by the model to provide explanations on individual borrower dataset. The ARSLRP values offer a comprehensive understanding of the factors influencing risk predictions in the individual borrower dataset. Starting with positive contributors, "Amount Due to Loan Outstanding" is the most influential variable with a positive ARSLRP value of +0.3076. This suggests that borrowers with higher number of dues of the loan outstanding are perceived as riskier by the model. Similarly, "Loan Cycle Number" and "Debt to Income Ratio" contribute positively, indicating that borrowers with higher loan cycles and those with elevated debt relative to income are associated with increased risk. These findings emphasize the model's sensitivity to the financial positions and credit history of the borrowers.

Additionally, various financial ratios such as "Expense to Income Ratio," "Loan to Income Ratio," and "Loan to Deposit Ratio" contribute positively, indicating that borrowers with higher expense and loan-to-deposit ratios are perceived as riskier. The model seems to prioritize a cautious approach towards borrowers with higher financial commitments and dependency on loans. The positive contribution of "Age" suggests that older borrowers are associated with increased risk, possibly indicating that the model considers factors related to the borrower's life stage in its risk assessment. The middle age borrowers may have higher family commitments towards children education, health and housing requirements which pushes them to higher level of debt.

Conversely, the neutral contribution of "Gender" suggests that gender does not significantly influence the model's risk predictions. The model does not differentiate between male and female borrowers in terms of perceived risk. Moving to negative contributors, "Number of Guarantors" negatively influences risk predictions, implying that borrowers with more guarantors are considered less risky. This suggests that having additional guarantors provides a sense of security in the model's assessment. The negative impact of "Pension Account (Yes)" as a contributor indicates that borrowers with pension accounts are considered less risky. This aligns with the notion that individuals with stable sources of income, such as a pension, may be perceived as more reliable borrowers. However, the most impactful negative contributor is "Per Capita Income" with a negative ARSLRP value of -0.2108. This implies that borrowers with increasing per capita income are seen as less risky by the model. It suggests that higher individual income levels play a significant role in mitigating perceived risk of defaults.

ARSLRP values in individual borrower dataset reveal that the model relies on a combination of financial ratios, historical borrowing patterns, and demographic factors to assess default risk. Positive contributors highlight the risk associated with higher outstanding amounts, specific financial ratios, and certain borrower characteristics. Negative contributors point to factors such as having more guarantors, possessing a pension account, and higher per capita income as indicators of lower perceived risk. These insights provide valuable guidance for refining risk assessment strategies and making informed decisions tailored to the nuances of individual borrower datasets.

The ARSLRP -based explanations can shed light on the model's decision-making process and highlight the features that are most influential in determining credit risk. Our findings indicate that ARSLRP provides meaningful and interpretable explanations for credit risk predictions. By propagating relevance through each layer of the model, we identified the features that contribute the most to the final prediction. This feature importance helps users comprehend the risk factors considered by the model, leading to enhanced transparency and trust in the credit risk assessment process. Moreover, ARSLRP enables us to analyze how the model handles both positive and negative instances. We observed that high-risk customers received higher relevance on features associated with past credit history, debt-to-income ratio, and payment delinquencies. On the other hand, low-risk customers obtained higher relevance on features like steady income, low credit utilization, and a history of timely payments. These findings align with domain knowledge and provide valuable insights for risk managers in understanding the decision-making process of the model.

Furthermore, the ARSLRP -based explanations revealed cases where the model's predictions deviated from conventional wisdom. In such instances, stakeholders can closely investigate the underlying factors and potentially identify areas for model improvement or data validation. For instance, if the model assigns high relevance to a seemingly irrelevant feature, such as a customer's occupation, it may raise concerns about data quality or the model's sensitivity to certain attributes. One of the strengths of ARSLRP is its ability to handle complex models, including deep learning architectures. Traditional linear models or decision trees often lack the capacity to capture intricate patterns in credit risk prediction, whereas deep learning models like CNNs and LSTMs can capture nonlinear relationships in the data. ARSLRP can handle such complex architectures, providing detailed explanations for individual predictions and overall model behavior. In conclusion, ARSLRP -based explanations offer a valuable tool for interpreting credit risk prediction models. The insights provided by ARSLRP facilitate understanding model predictions, identifying influential features, and assessing the model's performance.

V. Conclusion

Credit risk prediction is a vital aspect of the financial industry, where accurate assessments of borrowers' creditworthiness are crucial for making responsible lending decisions. This research explored the integration of explainable AI (XAI) techniques and ensemble methods to address the dual objectives of predictive accuracy and model interpretability in credit risk prediction. Specifically, the proposed approach leverages multistage ensemble techniques with a 1D CNN architecture to achieve both high performance and transparent decision-making. Our proposed approach aims to enhance credit risk prediction by integrating multistage ensemble techniques with a 1D CNN architecture. The model operates through multiple stages, providing interpretable explanations for its decisions, while maintaining high predictive performance. By offering transparency into the decision-making process, our proposed approach empowers financial institutions to understand and validate the model's risk assessments, ensuring fair lending practices and regulatory compliance. In conclusion, the proposed approach presents a novel contribution to the field of credit risk prediction. By combining the advantages of multistage ensemble techniques and XAI, our proposed model offers a balance between predictive accuracy and model interpretability, making it a valuable tool for credit risk assessment in the financial industry. Transferability of the trained model across two credit risk domains or financial institutions is investigated, assessing the model's generalizability and ability to provide meaningful explanations in diverse settings. As a future enhancement extend the model to handle time-series or sequential data that often appear in credit risk scenarios. Incorporating temporal dependencies and sequential patterns could enhance the model's predictive performance and provide more meaningful explanations.

References

[1] V. Ivashina and D. Scharfstein, “Bank lending during the financial crisis of 2008,” J financ econ, vol. 97, no. 3, pp. 319–338, Sep. 2010, doi: 10.1016/J.JFINECO.2009.12.001.

[2] J. C. K. Chow, “Analysis of Financial Credit Risk Using Machine Learning,” Feb. 2018, doi: 10.13140/RG.2.2.30242.53449.

[3] A. Bhattacharya, S. K. Biswas, and A. Mandal, “Credit risk evaluation: a comprehensive study,” Multimedia Tools and Applications 2022, pp. 1–51, Oct. 2022, doi: 10.1007/S11042-022-13952-3.

[4] D. Zhang, H. Huang, Q. Chen, and Y. Jiang, “A comparison study of credit scoring models,” in Proceedings - Third International Conference on Natural Computation, ICNC 2007, 2007, pp. 15–18. doi: 10.1109/ICNC.2007.15.

[5] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR,” Harv J Law Technol, vol. 31, no. 2, 2018, doi: 10.1177/1461444816676645.

[6] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine Learning Interpretability: A Survey on Methods and Metrics,” Electronics (Basel), vol. 8, no. 8, p. 832, Jul. 2019, doi: 10.3390/electronics8080832.

[7] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Feb. 2017, Accessed: Apr. 27, 2021. [Online]. Available: http://arxiv.org/abs/1702.08608

[8] M. P. Neto and F. V. Paulovich, "Explainable matrix - Visualization for global and local interpretability of random forest classification ensembles," IEEE Trans Vis Comput Graph, vol. 27, no. 2, pp. 1427–1437, Feb. 2021, doi: 10.1109/TVCG.2020.3030354.

[9] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "An Interpretable Model with Globally Consistent Explanations for Credit Risk," NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, Montréal, Canada. 2018.

[10] M. B. Goudzwaard, "Consumer Credit Charges and Credit Availability," South Econ J, vol. 35, no. 3, p. 214, Jan. 1969, doi: 10.2307/1056532.

[11] K. Jajuga, "Statistical Methods in Credit Risk Analysis," Prace Naukowe Akademii Ekonomicznej we Wrocławiu. Taksonomia, vol. 8, no. nr 906 Klasyfikacja i analiza danych : teoria i zastosowania, pp. 224–232, 2001.

[12] M. J. Furletti, "An Overview and History of Credit Reporting," SSRN Electronic Journal, Dec. 2011, doi: 10.2139/ssrn.927487.

[13] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat Mach Intell, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[14] J. Adams and H. Hagras, "A type-2 fuzzy logic approach to explainable ai for regulatory compliance, fair customer outcomes and market stability in the global financial sector," in IEEE International Conference on Fuzzy Systems, Institute of Electrical and Electronics Engineers Inc., Jul. 2020. doi: 10.1109/FUZZ48607.2020.9177542.

[15] S. Dash, O. Günlük, and D. Wei, "Boolean Decision Rules via Column Generation," in 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, 2018. Accessed: Apr. 27, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/743394beff4b1282ba735e5e3723ed74-Paper.pdf

[16] Y. Hayashi and N. Takano, "One-dimensional convolutional neural networks with feature selection for highly concise rule extraction from credit scoring datasets with heterogeneous attributes," Electronics (Switzerland), vol. 9, no. 8, pp. 1–15, Aug. 2020, doi: 10.3390/electronics9081318.

[17] W. Liu, H. Fan, and M. Xia, "Step-wise multi-grained augmented gradient boosting decision trees for credit scoring," Eng Appl Artif Intell, vol. 97, p. 104036, Jan. 2021, doi: 10.1016/j.engappai.2020.104036.

[18] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and S. Z. Li, "Ensemble-based discriminant learning with boosting for face recognition," IEEE Trans Neural Netw, vol. 17, no. 1, pp. 166–178, 2006, doi: 10.1109/TNN.2005.860853.

[19] S. Yamashkin, A. Yamashkin, M. Radovanović, M. Petrović, and E. Yamashkina, "Classification of Metageosystems by Ensembles of Machine Learning Models," International Journal of Engineering Trends and Technology, vol. 70, pp. 258–268, 2022, doi: 10.14445/22315381/IJETT-V70I9P226.

[20] M. P. Neto and F. V. Paulovich, "Explainable matrix - Visualization for global and local interpretability of random forest classification ensembles," IEEE Trans Vis Comput Graph, vol. 27, no. 2, pp. 1427–1437, 2021, doi: 10.1109/TVCG.2020.3030354.

[21] V. García, A. I. Marqués, and J. S. Sánchez, "Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction," Information Fusion, vol. 47, pp. 88–101, May 2019, doi: 10.1016/j.inffus.2018.07.004.

[22] Y. Song, Y. Wang, X. Ye, D. Wang, Y. Yin, and Y. Wang, "Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending," Inf Sci (N Y), vol. 525, pp. 182–204, Jul. 2020, doi: 10.1016/j.ins.2020.03.027.

[23] D. Reddy Edla, · Diwakar Tripathi, R. Cheruku, and V. Kuppili, "An Efficient Multi-layer Ensemble Framework with BPSOGSA-Based Feature Selection for Credit Scoring Data Analysis," Arab J Sci Eng, vol. 43, pp. 6909–6928, 2018, doi: 10.1007/s13369-017-2905-4.

[24] H. He and Y. Fan, "A novel hybrid ensemble model based on tree-based method and deep learning method for default prediction," Expert Syst Appl, vol. 176, Aug. 2021, doi: 10.1016/j.eswa.2021.114899.

[25] Y. Jin, W. Zhang, X. Wu, Y. Liu, and Z. Hu, "A Novel Multi-Stage Ensemble Model with a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data," IEEE Access, vol. 9, pp. 143593–143607, 2021, doi: 10.1109/ACCESS.2021.3120086.

[26] S. Wei, D. Yang, W. Zhang, and S. Zhang, "A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning," IEEE Access, vol. 7, pp. 99217–99230, 2019, doi: 10.1109/ACCESS.2019.2930332.

[27] W. Zhang, D. Yang, and S. Zhang, "A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring," Expert Syst Appl, vol. 174, p. 114744, Jul. 2021, doi: 10.1016/J.ESWA.2021.114744.

[28] J. Nalić, G. Martinović, and D. Žagar, "New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers," Advanced Engineering Informatics, vol. 45, p. 101130, Aug. 2020, doi: 10.1016/j.aei.2020.101130.

[29] A. Gicić and A. Subasi, "Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers," Expert Syst, vol. 36, no. 2, p. e12363, Apr. 2019, doi: 10.1111/exsy.12363.

[30] S. Guo, H. He, and X. Huang, "A Multi-Stage Self-Adaptive Classifier Ensemble Model With Application in Credit Scoring," IEEE Access, vol. 7, pp. 78549–78559, 2019, doi: 10.1109/ACCESS.2019.2922676.