J Inf Process Syst, Vol.20, No.6, pp.812~826, December 2024
https://doi.org/10.3745/JIPS.01.0110
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Research on Transformation and Interpretability in Credit Classification

Jihong Kim and Nammee Moon*

Abstract

The modern financial industry demands rapid decision-making based on diverse information from dynamic environments. Predicting outcomes from such data is complex due to rapid shifts influenced by numerous factors. Despite advancements in artificial intelligence technology that offer sophisticated analytical models, accurately predicting outcomes and providing sufficient justification for these predictions remain challenging, particularly with fragmented model constructions. In this paper, we propose a novel approach for efficient processing of available public personal credit data, deriving new analysis elements, and comparing prediction interpretations. Specifically, we develop 11 prediction models that can be categorized into two types: data image transformation and time-series transformation. The models undergo standardization, preprocessing, and cross-validation for optimization, with their predictive performances compared and validated. Models leveraging convolutional neural network (CNN) and convolutional neural network-long short-term memory (CNN-LSTM) architectures demonstrate strong performance across both categories. To fully interpret the classification process, SHAP is applied to compare and explain the prediction results for each model type.

Keywords

Big Data Processing and Analysis, Credit Risk Prediction, Deep Learning, eXplainable AI

1. Introduction

With the advancement of the Internet and digital technologies, financial services have increasingly moved online, giving rise to new financial platform models that are reshaping traditional financial transactions. In particular, peer-to-peer (P2P) lending platforms have introduced a new change by facilitating direct lending and investment between participants, providing an environment for borrowing funds without intermediaries [1]. This has redefined the relationship between financial institutions and borrowers within the capital market [2].

Since financial data consists of many variables, selecting the right variables is important for improving classification performance. Convolution neural network (CNN) is considered as one of the most effective feature extraction algorithms and a highly efficient neural network learning model. Sohangir et al. [3] confirmed that big data-formatted financial data, due to its complexity, benefits from deep learning models using CNN and long short-term memory (LSTM) for prediction. Kim and Cho [4] demonstrated the strong performance of CNN models in extracting complex features and identifying patterns.

Manuscript received February 2, 2024; first revision April 25, 2024; accepted June 18, 2024.

* **Corresponding Author:** Nammee Moon (nammee.moon@gmail.com)
Dept. of Computer Science and Engineering, Hoseo University, Asan, Korea (jh.wisard@gmail.com, nammee.moon@gmail.com)

Previous studies on credit risk prediction [3-5] have not directly employed high-performance CNN models for financial prediction, instead using them primarily for feature extraction. The extracted features were then applied to other artificial intelligence (AI) models for prediction. Recent studies [6,7] have shown the excellent performance of CNN models in encoding tabular numerical data into images. In this paper, we aim to directly evaluate the prediction performance of CNN models using two algorithms that transform the data itself into images.

In addition, time dependence in loans is a crucial factor for both financial institutions and borrowers. To address this, we apply multivariate time series classification using CNN-LSTM. Although there have been time series analysis studies using LSTM [8,9], they focused on aggregating P2P individual loan sequential data on a monthly basis to predict monthly average default rates.

Regarding model interpretability, previous studies [10-12] mostly focused on developing models or techniques that offer interpretability. While post-hoc explanation for a single model is essential, this paper focuses on comparing model interpretations from different perspectives using the same data. The comparative analysis of model interpretability is a novel contribution in this study.

The rest of this paper is structured as follows: In Section 2 reviews existing research in the field. Section 3 provides an explanation of the research design, including analysis model construction, evaluation methods for each model, explainable artificial intelligence (XAI) design, and the overall research procedure. Section 4 details the data collection process, preprocessing, image transformation, time series transformation, and the training process for the analysis models. It also includes an analysis and summary of performance comparisons by model type and the interpretability of predictive models. Finally, Section 5 discusses the research results and limitations, and suggests directions for future research.

2. Related Works

As financial data becomes increasingly complex and larger in scale, research utilizing effective algorithms has gained prominence. Among these, CNN has shown outstanding performance as a deep learning model for feature extraction from financial data, largely due to their remarkable performance in image recognition. Hoseinzade and Haratizadeh [13] demonstrated significant performance improvement by applying CNN for automatic feature selection and market prediction, highlighting the importance of feature extraction in market forecasting. Cao and Wang [14] constructed a stock index prediction model based on CNN and CNN-SVM through rule-based methods to forecast future trends in financial activities. Their study confirmed the feasibility and effectiveness of both prediction models, concluding that utilizing neural networks for financial forecasting and handling continuous and categorical predictive variables can yield favorable results. Qian et al. [7] proposed a soft reordering 1D CNN architecture designed to adaptively reshape tabular data for CNN training, and showed that it outperforms existing machine learning and deep learning models for credit scoring.

Research employing LSTM or recurrent neural network (RNN) has also emerged to understand the characteristics and predict financial time series. Lee and Oh [15] collected 2,362 non-financial industry data spanning from 2000 to 2017, predicting credit ratings with a total of 35 variables. By applying nine different machine learning models, the study found that the LSTM model, commonly used in time series prediction models, performed the best. Liang and Cai [9] developed LSTM, ARIMA, support vector machine (SVM), and artificial neural network (ANN) models to examine the monthly default rate of new loans on a P2P lending platform. Their results indicated that the LSTM model significantly enhanced prediction accuracy and trend accuracy across various time series cross-validations.

Research has also introduced hybrid CNN-LSTM models that combine the strengths of CNN and LSTM. Liu et al. [16] proposed a CNN-LSTM model to analyze quantitative strategies in the stock market. By utilizing CNN for trend analysis to establish stock selection strategies and LSTM for formulating timing strategies to enhance profitability, the study achieved better returns than basic momentum strategies and benchmark indices. Eapen et al. [17] developed a CNN-BiLSTM model that improved predictive performance by 9% in a single-pipeline deep learning model. This study implemented various modifications, including different CNN kernel sizes and bidirectional LSTM (BiLSTM) units, minimizing overfitting while enhancing prediction accuracy. Additionally, Lu et al. [18] proposed a CNN-BiLSTM-AM model to forecast the closing price of the Shanghai Composite Index. CNN was employed to extract features from input data, BiLSTM used these extracted features to predict the next day's stock closing price, and an attention mechanism (AM) was used to enhance prediction accuracy. When comparing the proposed model to seven other models including CNN, RNN, and LSTM, the proposed model demonstrated superior performance.

Machine learning models' high accuracy often comes at the cost of interpretability. While these technologies are widely utilized in the financial sector, the use of complex models makes it difficult to interpret the process of result derivation and to explain the correlation between model outputs and the underlying variables. XAI techniques, such as LIME [19] and SHAP [20], aim to make machine learning models interpretable, resembling traditional linear models. Bussmann et al. [21] demonstrated the use of the XGBoost model for predicting corporate defaults using a dataset from a specialized credit rating agency for small and medium-sized enterprise (SME) commercial loans. Their study showcased both high accuracy in corporate default prediction and improved interpretability through SHAP. Misheva et al. [22] applied LIME and SHAP methods to XGBoost and random forest (RF) models based on P2P lending platform, providing consistent explanations aligned with financial logic. Gramegna and Giudici [23] trained an XGBoost model on Italian SME data and demonstrating the superiority of SHAP values in credit default prediction by applying LIME and SHAP. Kim [24] studied the differences in loan classification by analyzing interpretation across various model types. Chen et al. [25] suggested a global interpretable model that is decomposable into sub-models for additional risk assessment. This decomposed model generated from subgroups of features, produces default probabilities, offering both global interpretability through rules and local interpretability through case-based explanations.

3. Research Structure Design

In this paper, we undertake the following research steps. First, in order to process complex variables in financial data, soft rearrangement is performed to reorganize the tabular data for enhanced CNN learning. The reordering mechanism used at this time is based on the foundational concepts of Zhu et al. [26] and Sharma et al. [27], with partial modifications to create a 2D-matrix type dataset (For better understanding, the modified algorithm is expressed as Simple-Gen and Simple-Matrix).

Next, we proceed with CNN-LSTM-based multivariate time series classification to address the time dependence problem of personal loans. To achieve this, the dataset is transformed into a multivariate time series to measure time-related variables, while remaining independent variables are converted into images.

Lastly, in order to strengthen the explanatory power of the research model, we perform a performance comparison between the general classification model and the time series classification model. The best-performing model for each type is selected, and XAI is applied to provide predictive analysis for each model, allowing for simultaneous comparison and analysis of the results.

3.1 Research Model Design

To construct a dataset suitable for the credit risk prediction model proposed in this paper, a prediction dataset is reorganized into four types of datasets through a transformation process. The four final datasets are used for general classification prediction, general classification prediction converted to images, time series classification prediction, and time series classification prediction converted to images. The architecture of the research model is illustrated in Fig. 1.

To prevent overfitting that may affect model prediction performance and to create a generalized model, cross-validation is performed by distinguishing between general classification prediction and time series classification prediction. Additionally, to improve prediction performance, optimal parameters for each classification model are determined through hyperparameter adjustment.

After training, the prediction results are categorized into general classification prediction and time series classification prediction, followed by performance evaluation for each model. XAI is applied to the top-performing models for each type to compare and analyze the interpretability of the model prediction process.

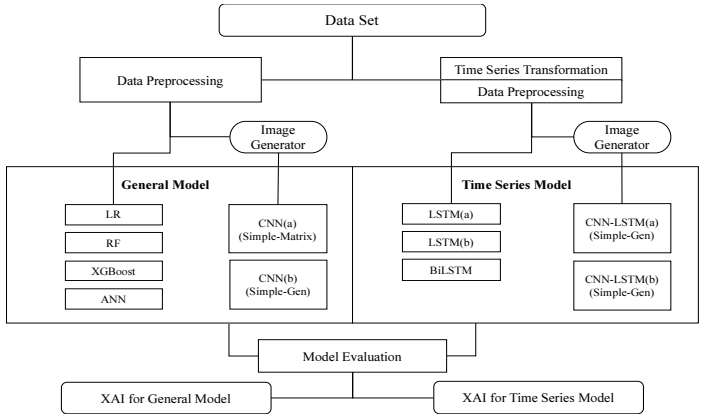


Fig. 1. Research architecture.

For consistency in terminology used in this paper, there are two model types: General Model for general classification prediction model and Time Series Model for time series prediction classification model.

3.2 Model Configuration

The General Model was constructed by selecting an algorithm known for its strong performance in binary classification, along with a CNN algorithm recognized for its excellent performance in image analysis. The Time Series Model is composed of LSTM algorithms that excel in time series analysis. The objective is to determine whether CNN's impressive performance is maintained in classification learning, and to explore any differences between time series classification prediction and the existing binary classification model when applied to the same dataset. The model configuration is shown in Table 1.

3.3 Evaluation Indicators

Performance evaluation of a classification model generally confirms the results by comparing actual and predicted values. This study uses three evaluation metrics: accuracy, F1-score, and area under the curve (AUC) [28].

Table 1. Model configuration

Model	Base model	Data format
General Model		
LR	Logistic regression	Tabular data
RF	Random forest	Tabular data
XGBoost	XGBoost	Tabular data
ANN	ANN	Tabular data
CNN(a)	CNN	Image data
CNN(b)	CNN	Image data
Time Series Model		
LSTM(a)	LSTM	Time series data
LSTM(b)	LSTM	Time series data
BiLSTM	Bidirectional LSTM	Time series data
CNN-LSTM(a)	1-dimension CNN, LSTM	Image-time series data
CNN-LSTM(b)	2-dimension CNN, LSTM	Image-time series data

3.4 XAI Design

The performance of the General Model and the Time Series Model are compared, and the SHAP method is applied to the model that showed the best performance. SHAP provides explanatory power in two ways: First, Global Interpretation quantifies and visualizes the contribution of each variable, providing an analysis of their relative importance from a model-wide perspective. Second, Local Interpretation focuses on specific data and allows you to visually examine the variables contributed to the prediction result. The visual representations of the interpretation for each model type are expressed based on the same data.

4. Experiment and Result Analysis

4.1 Experiment Data

This study used personal loan-related data released by the U.S. Lending Club. It contains information about the loan applicants including their repayment histories. Our objective is to classify the likelihood of individual defaults.

The dataset spans from 2007 to 2020 and contains a total of 2,925,493 entries. It comprises 141 independent variables, including credit limit balance, average account balance, loan amount, credit rating, with the dependent variable being 'loan_status.' The individuals in the data reside in the United States and the dataset includes details on the company, location, and credit information for loans.

4.2 Data Processing

4.2.1 Preprocessing

Loan data consists of numerous variables, presenting challenges when applying analysis models. Consequently, it is essential to reduce the number of data variables.

First, we investigate missing values for each variable. When variables with a high rate of missing values are assigned with substitute values, it can impact predictions. Therefore, in this study, such variables will

be excluded. Additionally, variables that are not significantly associated with the prediction of default will be excluded. This includes variables related to pre-loan execution, investors in the loan, and any factors that do not have a meaningful connection with the load itself.

The principal component analysis calculates the principal component with the highest eigenvalue, thereby selecting the most important variables. Furthermore, we intentionally choose the final variable based on their importance and scalability in predictions. Categorical variables are converted to numeric values using the one-hot-encoding method [29]. Among them, 'emp_length' and 'grade' are converted to numeric data through variable abstraction. Ultimately, this process results in a refined dataset consisting of 47 variables.

The dependent variable is the 'loan_status' data, which is categorized into nine types based on the degree of delinquency and repayment status, as shown in Table 2. This study aims to predict the risk of current borrowers, using "Fully Paid" and "Charged Off" as dependent variable categories. Loans marked as "Fully Paid" is assigned a value of 0, totaling 1,497,783 records, while "Charged Off" loans are assigned a value of 1, with a total of 362,548 records. This process results in an organized dataset comprising a total of 1,860,331 records.

Table 2. Dependent variable (loan status)

Loan status	Count
Fully Paid	1,497,783
Current	1,031,016
Charged Off	362,548
Late (31–120 days)	16,154
In Grace Period	10,028
Late (16–30 days)	2,719
Issued	2,062
Does not meet the credit policy	2,749
Default	433

4.2.2 Data conversion

The final 46 independent variables, excluding the key variables 'id' and 'issue_d,' are transformed using 2D matrix conversion algorithms (Simple-Gen, Simple-Matrix) to create a single channel image dataset with a shape of [7,7]. In this square-shaped 2D matrix, missing cells are filled with 0. The referenced algorithms are originally designed to predict associations between genomic information and drug. For this study, the algorithm has been partially modified and adapted for our specific application.

4.2.3 Preparing for time series analysis

A time series is a sequence of observations recorded at some time intervals. In this study, the issuance date of loans is used to determine the start date of the time series. In the preprocessing of time series data, it is important to check the stationarity. If the time series is not stationary, it is first transformed into a stationary time series to ensure that the mean and variance remain constant over time. To achieve this, the augmented Dickey-Fuller (ADF) test [30] is applied to each individual variable within the time series. The ADF test results indicate that all variables are stationary over time, confirming that their mean and variance remain constant over time, with no identifiable trends or seasonality associated with the variables.

4.2.4 Cross-validation

To minimize data overfitting and bias, The General Model applies 5-fold cross-validation. The entire dataset is split into training and test sets in a 7:3 ratio, and the training data is randomly and evenly divided into 5 folds. One fold is further divided into 5 folds, creating training and validation datasets for model learning. The remaining folds undergo the same process with the validation dataset rotated among the different folds, and this procedure continues until all folds have served as validation datasets.

For Time Series Model, a 5-split time series cross-validation is applied. Similar to the General Model, the entire dataset is equally divided into training and test sets in a 7:3 ratio. In the 5-split iterative process, the partitioned datasets are evenly distributed and at each step, the training set is expanded to facilitate model learning.

4.2.5 Hyperparameter

General machine learning models exhibit significant performance variations depending on parameter and hyperparameter settings. However, since this study focuses on comparing classification outcomes between CNN-based model and time series-based model, hyperparameter tuning is performed using basic grid search [31]. Deep learning models use hyperparameters such learning rate, mini-batch size, and optimization algorithm. The hyperparameters used in each model are summarized in Table 3.

Fig. 2 provides a pseudo-code that outlines the adopted algorithm detailing the steps and implementations of the proposed procedure.

4.3 Results Analysis

4.3.1 Comparison of model performance

The results measured with the test data are summarized in Table 4. In the General Model, accuracy is generally high across all models, with the exception of logistic regression (LR). The CNN-based models show superior performance compared to non-CNN-based models, with an average accuracy that is 5.7% higher, an F1-score 33.6% higher, and an AUC that is 1.9% higher. Notably, the CNN(b) model, which applies the Simple-Gen transformation, achieves the best performance among other classification models,

Table 3. Model hyperparameters

Model	Hyperparameter
LR	max_iter=500, multi_class='ovr', random_state=42
RF	criterion=entropy, n_estimators=100
XGBoost	n_estimators=100, colsample_bytree=0.5, max_depth=7, random_state=42
ANN	learning_rate=0.001, batch_size=128, optimizer=Adam, loss=binary_crossentropy
CNN(a)	learning_rate=0.001, batch_size=128, optimizer=Adam, loss=binary_crossentropy
CNN(b)	learning_rate=0.001, batch_size=128, optimizer=Adam, loss=sparse_categorical_crossentropy
LSTM(a)	learning_rate=0.001, batch_size=64, optimizer=RMSprop, loss=sparse_categorical_crossentropy
LSTM(b)	learning_rate=0.001, batch_size=64, optimizer=Adam, loss=sparse_categorical_crossentropy
BiLSTM	
CNN-LSTM(a)	learning_rate=0.0005, batch_size=64, optimizer=Adam, loss=sparse_categorical_crossentropy
CNN-LSTM(b)	

```
1. Inputs: tabular data, Outputs: accuracy of the forecasted data
2. Split data into 70% training and 30% testing data
3. Data preprocessing
   Reducing data variables
   Complex sampling for unbalanced dataset
   ADF test for time series analysis
4. Convert numeric data to [7,7] image shape using image generator(Simple-Gen, Simple-Matrix)
5. General Model configuration
   LR = LogisticRegression(random_state=42, max_iter=500, multi_class='ovr')
   RF = RandomForestClassifier(criterion=entropy, n_estimators=100)
   XGBoost = XGBClassifier(n_estimators=100, colsample_bytree=0.5, max_depth=7, random_state=42)
   ANN = Keras custrom model
   CNN(a) = Pytorch CNN model adapting Simple-Matrix
   CNN(b) = Keras CNN model adapting Simple-Gen
6. Time Series Model configuration
   LSTM(a) = Keras LSTM mode using RMSprop optimizer
   LSTM(b) = Keras LSTM model using Adam optimizer
   BiLSTM = Keras LSTM model
   CNN-LSTM(a) = Keras 1 Dimension CNN-LSTM model adapting Simple-Gen
   CNN-LSTM(b) = Keras 2 Dimension CNN-LSTM model adapting Simple-Gen
7. Model training
   General Model using 5-fold cross validation
   Time Series Model using 5-split time series cross validation
8. Model evaluation using accuracy, F1-score and AUC
9. SHAP Deep explainer to visualize global interpretation and local interpretation of the model
```

Fig. 2. Pseudo-code for the proposed algorithm.

with an average accuracy 6.1% higher, an F1-score 33.8% higher, and an AUC 2.2% higher.

In Time Series Model, the CNN-LSTM based models outperform the standard LSTM models, with an average accuracy 1.1% higher, an F1-score 7.5% higher, and an AUC 3.3% higher. Furthermore, the CNN-LSTM(b) model exhibits the best performance among the LSTM-based models, achieving an average accuracy 1.5% higher, an F1-score 10.1% higher, and an AUC 4.5% higher. Fig. 3 visualizes parts of the prediction results for better understanding.

4.3.2 Comparison of general model vs. time series model and non-CNN-based models vs. CNN-based models

Excluding LR, which showed low accuracy within the General Model, we conduct a comparative analysis between the General Model and the Time Series Model.

Table 4. The performance results by model

Model	Accuracy	F1-score	AUC
General Model			
LR	0.7499	0.6026	0.8442
RF	0.8278	0.6710	0.8554
XGBoost	0.8457	0.6906	0.8598
ANN	0.8227	0.6356	0.8114
CNN(a)	0.8559	0.8670	0.8559
CNN(b)	**0.8611**	**0.8699**	**0.8613**
Time Series Model			
LSTM(a)	0.8598	0.5751	0.7072
LSTM(b)	0.8728	0.6040	0.7191
BiLSTM	0.8630	0.5531	0.6924
CNN-LSTM(a)	0.8712	0.6053	0.7210
CNN-LSTM(b)	**0.8782**	**0.6359**	**0.7381**

The bold font indicates the best performance in each test.

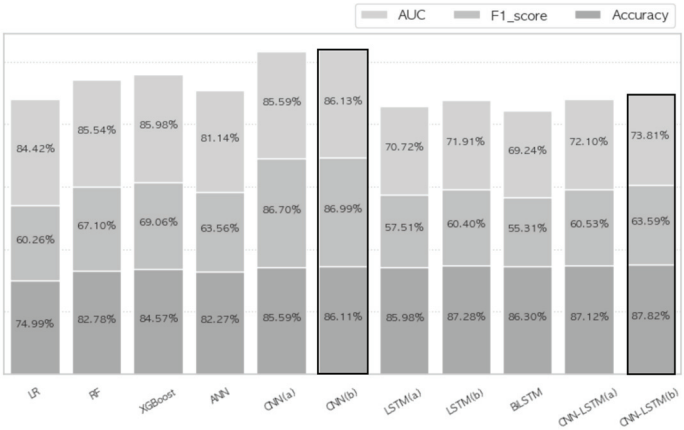


Fig. 3. The performance results by model.

Table 5 summarizes the results of the comparison. The Time Series Model demonstrates an average accuracy that is 3.1% higher than the General Model; however, the F1-score and AUC are low by -20.3% and -15.7%.

CNN-based models outperform their non-CNN counterparts, with all metrics showing increases: average accuracy improved by 2.1%, F1-score enhanced by 19.8%, and AUC rising by 2.6%. Fig. 4 visualizes the performance results for clarity.

The performance of both General Model and Time Series Model varies based on their evaluation criteria, and it can be confirmed that the CNN-based models generally have better performance than the non-CNN-based models.

Table 5. The performance results by type

Type	Accuracy	F1-score	AUC
General Model	0.8426	**0.7468**	**0.8488**
Time Series Model	**0.8690**	0.5947	0.7156
Non-CNN-based models	0.8486	0.6216	0.7742
CNN-based models	**0.8666**	**0.7445**	**0.7941**

The bold font indicates the best performance in each test.

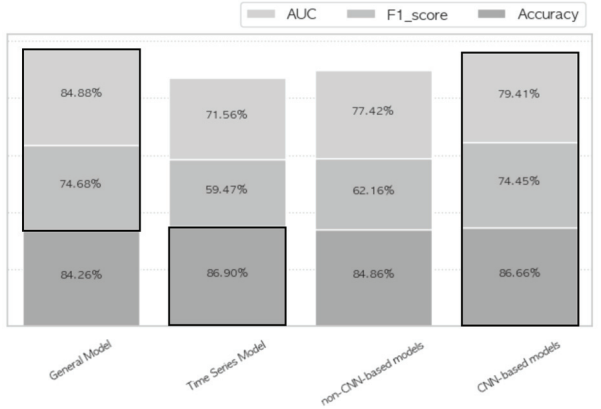


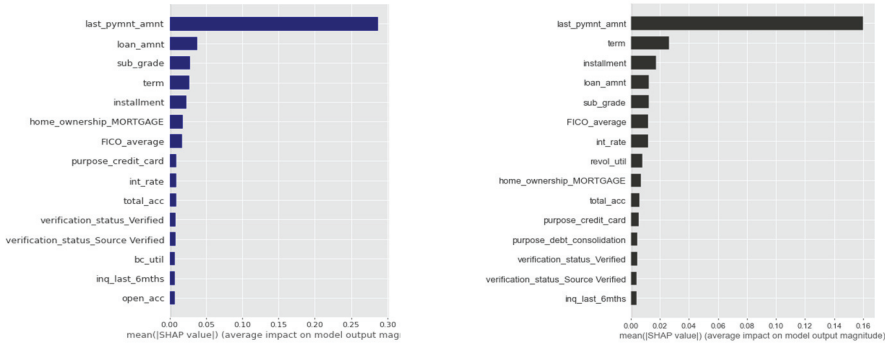
Fig. 4. The performance results by type.

4.3.3 Comparison of model explanation

After evaluating the performance of the classification models, post hoc explanations using the SHAP method were applied to the CNN(b) and CNN-LSTM(b) models, which demonstrated excellent results for enhanced interpretability. As the target models are deep learning models, deep SHAP was used.

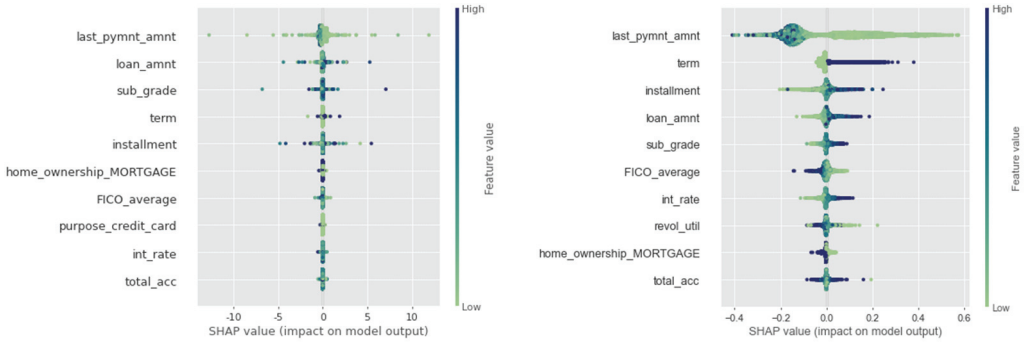
First, a global interpretation was conducted to assess the overall classification influence of the models. We examined variable importance to check the impact of all independent variables on the dependent variable ('loan_status'). SHAP variable importance represents the average of the absolute SHAP value in the classification prediction process. The magnitude of these values indicates how much each variable contributed to the prediction process.

Fig. 5 is a visualization of the top 15 variables ranked by the importance in descending order. For CNN(b) model, the variable importance is ranked in the following order: 'last_pymnt_amnt,' 'loan_amnt,' 'sub_grade,' 'term,' and 'installation.' Meanwhile, the CNN-LSTM(b) model, ranked the



(a) (b)

Fig. 5. Global Interpretation (variable importance): (a) CNN(b) model and (b) CNN-LSTM(b) model.



(a) (b)

Fig. 6. Global Interpretation (scatter plot with density estimation): (a) CNN(b) model and (b) CNN-LSTM(b) model.

variable importance as: 'last_pymnt_amnt,' 'term,' 'installation,' 'loan_amnt,' and 'sub_grade.' It is observed that the top elements in variable importance, including 'last_pymnt_amnt,' 'term,' and 'installment,' are consistent across both models. However, there are notable differences in the ranking order between the two models' interpretation processes.

Fig. 6 presents scatter plot that combine variable importance with variable effects. The x-axis represents SHAP values, and the y-axis represents important variables. The color of the feature value indicates the abundance (dark) or scarcity (light) of variable values.

In the case of CNN(b), most of the important variables, such as 'last_pymnt_amnt' and 'loan_amnt,' show unclear scatter plot colors, making interpretation difficult.

For the CNN-LSTM(b) model, when the value of 'last_pymnt_amnt' is high (dark color), it indicates a negative impact on "Charged Off," while a low value (light color) has a positive impact. Similarly, lower values (light color) of variables such as 'term,' 'installment,' and 'loan_amnt,' results in a negative impact on "Charged Off," and a high value (dark color) yield positive impacts. Overall, 'last_pymnt_amnt' shows a negative correlation with "Charged Off," while 'term,' 'installment,' and 'loan_amnt' are positively correlated. Similar interpretations can be applied to other variables, and it must be noted that interpretation is possible when the colors of scatter plot are clear.

The following is an interpretation of individual data classification for the applied models (Local Interpretation).

Fig. 7 visualizes individual model predictions, allowing us to understand the variable influences for each data point. In this case (data A), both the actual and predicted outcomes reflect successful "Fully Paid" status. For CNN(b), the SHAP prediction value is 0.21. The variables that negative impact on "Charged Off" (positive impacting "Fully Paid") are 'sub_grade,' 'FICO_average,' and 'int_rate.' Conversely, variables that has a positive impact on "Charged Off" (negative impact on "Fully Paid") are 'last_pymnt_amnt' and 'revol_util.' For CNN-LSTM(b), the SHAP prediction value is 0.01 with 'loan_amnt,' 'installment,' and 'sub_grade' negatively impacting "Charged Off," while 'last_pymnt_amnt' positively impacts it.

While the variables affecting "Charged Off" do not precisely match between the General Model and the Time Series Model, there is consistency in the directional impact of variables such as 'last_pymnt_amnt' and 'sub_grade.'

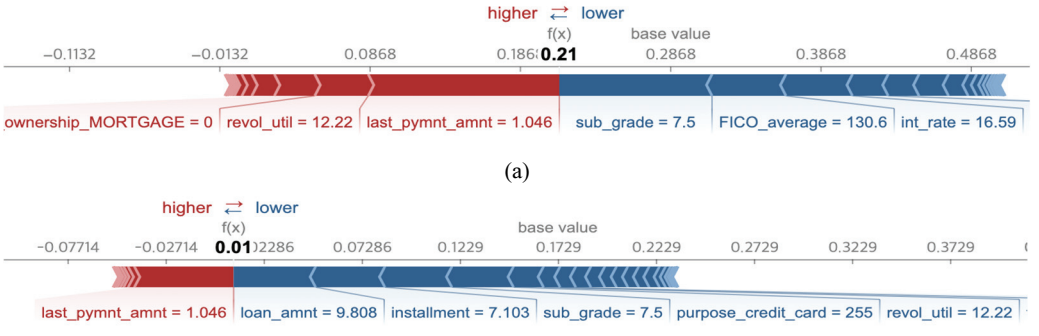


Fig. 7. Local Interpretaion (loan_status=0) of data A: (a) CNN(b) model and (b) CNN-LSTM(b) model.

Fig. 8 presents the prediction visualization for data B where both the actual and predicted values indicate a "Charged Off."

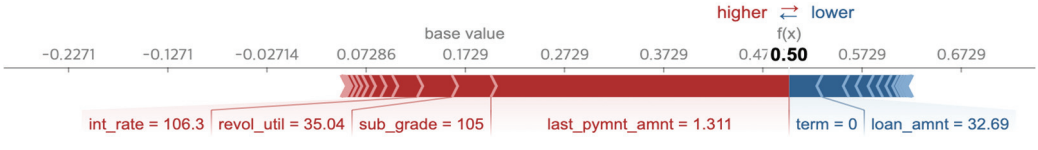
For CNN(b), the SHAP prediction value is 0.50. The variables that negatively impact "Charged Off" include 'term' and 'loan_amnt.' While 'last_pymnt_amnt,' 'sub_grade,' 'revol_util,' and 'int_rate' positively impact it. For CNN-LSTM(b), the SHAP prediction value is 0.74. Variables negatively

impacting “Charged Off” are ‘purpose_other’ and ‘loan_amnt,’ whereas ‘last_pymnt_amnt,’ ‘sub_grade’ and ‘home_ownership_MORTGAGE’ positively impact it.

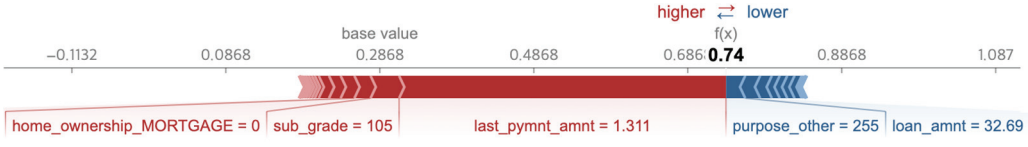
While the specific variables influencing “Charged Off” vary between the General Model and the Time Series Model, there is consistency in the directionality of the variable ‘last_pymnt_amnt,’ ‘sub_grade,’ and ‘loan_amnt,’

In the overall classification interpretation (Global Interpretation), the top common variables in importance for both models (General Model and Time Series Model) are ‘last_pymnt_amnt,’ ‘term,’ and ‘installment.’ However, differences were observed in the interpretation process between the two models, making it challenging to identify similarities through the combined analysis of variable importance, variable effects, and interaction analysis among independent variables.

For the individual data classification interpretation (Local Interpretation), regardless of the data type (“Fully Paid” or “Charged Off”), it was found that variables that positively or negatively influenced “Charged Off” during the prediction process did not align precisely. Nevertheless, there were instances where the prediction directionality remained consistent.



(a)



(b)

Fig. 8. Local Interpretaion (loan_status=1) of data B: (a) CNN(b) model and (b) CNN-LSTM(b) model.

5. Conclusion

This paper employs two algorithms for 2D matrix transformation to visualize data for credit risk prediction models. It also extracts time series characteristics to ensure stationarity for time series analysis on basic dataset. In the General Model, 5-fold cross-validation is applied to four non-CNN-based models and two CNN-based models for training, validation, and testing. For Time Series Models, time series cross-validation is applied, targeting three LSTM-based models and two CNN-LSTM-based models for training, validation, and testing to generate prediction results.

The General Model assessed prediction performance by differentiating between basic data and image data, revealing that models utilizing image data outperformed their counterparts. Similarly, the Time Series Model evaluated performance by distinguishing between time series basic data and time series image data, confirming the excellent performance of the models based on time series image data.

To identify differences in predictive interpretation between the General Model and the Time Series Model, the best-performing models from each category were selected, and SHAP was applied to compare

their predictive explanatory power. While the overall data and individual data interpretations showed similar predictive trends, variations in the ranking of importance for each variable were observed.

The research model in this paper aims to explore the potential of classification performance through the image visualization of tabular data, the efficiency of extracting time series elements from basic data, and understanding the predictive explanatory power differences across various model types. To accomplish this, a comprehensive research model was constructed, integrating data imaging, time series transformation, cross-validation, and XAI design. In addition, the process of refining each component to enhance prediction accuracy and the model analysis was intrinsically valuable as the empirical study progressed.

The results of model reliability using CNN, LSTM models and SHAP application can be useful for developing models that are robust and reliable across diverse data types, particularly in decision-making processes for credit rating agencies and other financial institutions.

In future research, we aim to expand various data processing capabilities for financial prediction and evaluate the effectiveness of different explainable AI techniques by establishing relevant evaluation metrics. Through this approach, we hope to confirm the utility of innovative methods and foster the development of transparent and trustworthy AI applications in the financial field by addressing the shortcomings of deep learning.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2021R1A2C2011966).

References

[1] A. Bachmann, A. Becker, D. Buerckner, M. Hilker, F. Kock, M. Lehmann, P. Tiburtius, and B. Funk, “Online peer-to-peer lending: a literature review,” *Journal of Internet Banking and Commerce*, vol. 16, no. 2, pp. 1-18, 2011.

[2] G. G. Parker, M. W. Van Alstyne, and S. P. Choudary, *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. New York, NY: W. W. Norton & Company, 2017.

[3] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, “Big data: deep learning for financial sentiment analysis,” *Journal of Big Data*, vol. 5, no. 1, article no. 3, 2018. https://doi.org/10.1186/s40537-017-0111-6

[4] J. Y. Kim and S. B. Cho, “Towards repayment prediction in peer-to-peer social lending using deep learning,” *Mathematics*, vol. 7, no. 11, article no. 1041, 2019. https://doi.org/10.3390/math7111041

[5] W. Jiao, X. Hao, and C. Qin, "The image classification method with CNN-XGBoost model based on adaptive particle swarm optimization," *Information*, vol. 12, no. 4, article no. 156, 2021. https://doi.org/10.3390/info12040156

[6] J. M. T. Wu, Z. Li, N. Herencsar, B. Vo, and J. C. W. Lin, "A graph-based CNN-LSTM stock price prediction algorithm with leading indicators," *Multimedia Systems*, vol. 29, no. 3, pp. 1751-1770, 2023. https://doi.org/10.1007/s00530-021-00758-w

[7] H. Qian, P. Ma, S. Gao, and Y. Song, "Soft reordering one-dimensional convolutional neural network for credit scoring," *Knowledge-Based Systems*, vol. 266, article no. 110414, 2023. https://doi.org/10.1016/j.knosys.2023.110414

[8] C. Wang, D. Han, Q. Liu, and S. Luo, "A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM," *IEEE Access*, vol. 7, pp. 2161-2168, 2018. https://doi.org/10.1109/ACCESS.2018.2887138

[9] L. Liang and X. Cai, "Forecasting peer-to-peer platform default rate with LSTM neural network," *Electronic Commerce Research and Applications*, vol. 43, article no. 100997, 2020. https://doi.org/10.1016/j.elerap.2020.100997

[10] T. Gramespacher and J. A. Posth, "Employing explainable AI to optimize the return target function of a loan portfolio," *Frontiers in Artificial Intelligence*, vol. 4, article no. 693022, 2021. https://doi.org/10.3389/frai.2021.693022

[11] A. Hanif, "Towards Explainable Artificial Intelligence in banking and financial services," 2021 [Online]. Available: https://arxiv.org/abs/2112.08441.

[12] I. D. C. Arifah and I. U. Nihaya, "Artificial intelligence in credit risk management of peer-to-peer lending financial technology: systematic literature review," in *Proceedings of 2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, Lombok, Indonesia, 2023, pp. 329-334. https://doi.org/10.1109/IC2IE60547.2023.10331487

[13] E. Hoseinzade and S. Haratizadeh, "CNNPred: CNN-based stock market prediction using several data sources," 2018 [Online]. Available: https://arxiv.org/abs/1810.08923.

[14] J. Cao and J. Wang, "Stock price forecasting model based on modified convolution neural network and financial time series analysis," *International Journal of Communication Systems*, vol. 32, no. 12, article no. e3987, 2019. https://doi.org/10.1002/dac.3987

[15] H. S. Lee and S. Oh, "LSTM-based deep learning for time series forecasting: the case of corporate credit score prediction," *The Journal of Information Systems*, vol. 29, no. 1, pp. 241-265, 2002. https://doi.org/10.5859/KAIS.2020.29.1.241

[16] S. Liu, C. Zhang, and J. Ma, "CNN-LSTM neural network model for quantitative strategy analysis in stock markets," in Neural Information Processing. Cham, Switzerland: Springer, 2017, pp. 198-206. https://doi.org/10.1007/978-3-319-70096-0_21

[17] J. Eapen, D. Bein, and A. Verma, "Novel deep learning model with CNN and bi-directional LSTM for improved stock market index prediction," in *Proceedings of 2019 IEEE 9th annual computing and communication workshop and conference (CCWC)*, Las Vegas, NV, USA, 2019, pp. 264-270. https://doi.org/10.1109/CCWC.2019.8666592

[18] W. Lu, J. Li, J. Wang, and L. Qin, "A CNN-BiLSTM-AM method for stock price prediction," *Neural Computing and Applications,* vol. 33, no. 10, pp. 4741-4753, 2021. https://doi.org/10.1007/s00521-020-05532-z

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778

[20] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural*

Information Processing Systems, vol. 30, pp. 4765-4774, 2017.

[21] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable AI in fintech risk management," *Frontiers in Artificial Intelligence*, vol. 3, article no. 26, 2020. https://doi.org/10.3389/frai.2020.00026

[22] B. H. Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, and S. F. Lin, "Explainable AI in credit risk management," 2021 [Online]. Available: https://arxiv.org/abs/2103.00949.

[23] A. Gramegna and P. Giudici, "SHAP and LIME: an evaluation of discriminative power in credit risk," *Frontiers in Artificial Intelligence*, vol. 4, article no. 752558, 2021. https://doi.org/10.3389/frai.2021.752558

[24] J. Kim, "A study on credit risk predicting base on XAI using transformation of time series data," Ph.D. dissertation, Department of Convergence Technology, Hoseo University, Cheonan, Korea, 2024.

[25] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, and T. Wang, "An interpretable model with globally consistent explanations for credit risk," 2018 [Online]. Available: https://arxiv.org/abs/1811.12615.

[26] Y. Zhu, T. Brettin, F. Xia, A. Partin, M. Shukla, H. Yoo, Y. A. Evrard, J. H. Doroshow, and R. L. Stevens, "Converting tabular data into images for deep learning with convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, article no. 11325, 2021. https://doi.org/10.1038/s41598-021-90923-y

[27] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture," *Scientific Reports*, vol. 9, no. 1, article no. 11399, 2019. https://doi.org/10.1038/s41598-019-47765-6

[28] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168-192, 2021. https://doi.org/10.1016/j.aci.2018.08.003

[29] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 7-9, 2017. https://doi.org/10.5120/ijca2017915495

[30] X. Fu, S. Zhang, J. Chen, T. Ouyang, and J. Wu, "A sentiment-aware trading volume prediction model for P2P market using LSTM," *IEEE Access*, vol. 7, pp. 81934-81944, 2019. https://doi.org/10.1109/ACCESS.2019.2923637

[31] J. Fan, X. Wang, F. Zhang, X. Ma, and L. Wu, "Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data," *Journal of Cleaner Production*, vol. 248, article no. 119264, 2020. https://doi.org/10.1016/j.jclepro.2019.119264

Jihong Kim **https://orcid.org/0000-0002-1380-5700**

He received B.S. degree from the Dept. of Mathematics at Kyung Hee University in 1994 and M.S. degree from IT MBA at Hanyang Cyber University in 2016. Since March 2021, he is with the Dept. of Convergence Technology from Hoseo University as a Ph.D. candidate. His current research interests include big data processing and analysis, explainable AI, deep learning, NLP, financial engineering.

Nammee Moon **https://orcid.org/0000-0003-2229-4217**

She received B.S., M.S., and Ph.D. degrees from the School of Computer Science and Engineering at Ewha Womans University in 1985, 1987, and 1998, respectively. She served as an assistant professor at Ewha Womans University from 1999 to 2003, then as a professor of digital media at the Graduate School of Seoul Venture Information, from 2003 to 2008. Since 2008, she has been a professor of Computer Science and Engineering at Hoseo University. Her current research interests include social learning, HCI and user-centric data, deep learning, and big data processing and analysis.