ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230532

Tab-Attention: Self-Attention-Based Stacked Generalization for Imbalanced Credit Default Prediction

Yandan Tan [a], Hongbin Zhu[b;*], Jie Wu[a,b] and Hongfeng Chai[a,b]

[a]School of Computer Science, Fudan University, Shanghai, China
[b]Institute of FinTech, Fudan University, Shanghai, China
ORCiD ID: Yandan Tan https://orcid.org/0000-0002-9629-1118

Abstract. Accurately credit default prediction faces challenges due to imbalanced data and low correlation between features and labels. Existing default prediction studies on the basis of gradient boosting decision trees (GBDT), deep learning techniques, and feature selection strategies can have varying degrees of success depending on the specific task. Motivated by this, we propose Tab-Attention, a novel self-attention-based stacked generalization method for credit default prediction. This approach ensembles the potential proprietary knowledge contributions from multi-view feature spaces, to cope with low feature correlation and imbalance. We organize multi-view feature spaces according to the latent linear or nonlinear strengths between features and labels. Meanwhile, the $f1$ score assists the model in imbalance training to find the optimal state for identifying minority default samples. Our Tab-Attention achieves superior $Recall_1$ and $f1_1$ of default intention recognition than existing GBDT-based models and advanced deep learning by about 32.92% and 16.05% on average, respectively, while maintaining outstanding overall performance and prediction performance for non-default samples. The proposed method could ensemble essential knowledge through the self-attention mechanism, which is of great significance for a more robust future prediction system.

1 Introduction

Credit risk is a critical aspect of financial risks, where credit default is a major manifestation. Credit defaults pose a significant impact on financial market stability and economic well-being [5]. Rising defaults have been observed globally, due to slowed economic growth, escalating geopolitical risks, and trade frictions [8]. These defaults have far-reaching consequences, including sluggish economic growth, diminished investor confidence, and substantial losses for financial institutions. Therefore, it is imperative to develop efficient credit risk management strategies to mitigate the adverse impact of credit defaults on both the financial market and the economy.

However, credit default prediction is a challenging task due to the data imbalance, high sparsity, and weak correlation between features and labels. In recent years, recent research endeavours have primarily focused on statistical learning and machine learning techniques to mitigate the risk of credit default.

Statistical learning: Traditionally, credit ratings play a crucial role in determining credit risk. However, the credit rating process is costly and can be influenced by subjective factors, such as the differences in expert experience and consideration standards[12, 24]. To overcome these issues, linear discriminant analysis (LDA) was introduced. However, LDA requires data to meet basic assumptions, such as multivariate normality and equal covariance matrices. Logistic regression (LR) is a more flexible method than traditional linear regression [19], but it is still limited in its ability to handle complex nonlinear relationships. LR is also sensitive to outliers and missing data, and it is difficult to train LR models on large datasets. Despite these limitations, LR is a widely used method for credit default prediction due to its good interpretability. However, the global financial crisis reveals some shortcomings of LR methods, such as their slow adaptability to changing economic conditions and their limited ability to model complex nonlinear interactions among economic, financial, and credit variables.

Machine learning techniques: The methods are crucial in adapting to changes in time and available data, while accommodating a large number of feature sets. Decision tree (DT) could capture the interactions and non-linear relationships between variables, thereby providing good discrimination between default and non-default cases [19]. However, repeated attribute segmentation for DT is susceptible to overfitting. Ensemble learning addresses this issue by growing multiple trees and computing their mean value [16]. For example, gradient-boosting decision trees (GBDT) learn a series of weak learners to predict outputs, where weak learners are typically non-differentiable standard DTs. Chen *et al*. [7] demonstrated that GBDT outperforms other methods in tabular data applications. Several GBDT algorithms, such as XGBoost [6, 14], LightGBM [29], and CatBoost [23], have been developed to predict credit risk successfully. Although these ensemble algorithms exhibit differences, their performance is often similar across various tasks [23].

Deep learning has emerged as a promising approach for credit default prediction due to its ability to capture complex patterns and nonlinear relationships between variables. Tan *et al*. [27] and Yang *et al*. [28] proposed improved DNN for superior credit default prediction than machine learning and CNN. Luo *et al*. [18] developed credit risk assessment models using deep belief networks, which excel in learning effective feature representations and capturing latent patterns in the data. Fu *et al*. [13] combined DNN and bidirectional long short-term memory to accurately identify potential credit risks. These studies demonstrate that deep learning techniques can significantly enhance the predictive performance of credit default models, making them valuable tools in credit risk management.

* Corresponding Author. Email: zhuhb@fudan.edu.cn

Furthermore, some novel methods are based on multi-view feature organization to improve model performance. Song *et al*. [25] designed multi-view-based feature sampling to optimize imbalanced learning. Tan *et al*. [27] proposed that multi-view learning of features based on discrete, continuous, and their correlation sign differences is an effective strategy to improve prediction accuracy. In addition, some related studies have shown that feature selection based on genetic algorithms [21], random forest (RF) [3], or regularization strategies [9] is conducive to model performance, as the feature selection process eliminates irrelevant and redundant features.

In general, existing works have shown that GBDT [6, 7, 14, 23, 29], deep learning [13, 18, 27, 28], effective feature selection strategies [3, 21], and multi-view ensemble [25, 27] could significantly improve model performance. However, there is a limited exploration on how to combine these optimization strategies to obtain better solutions.

For tabular data, Arik *et al*. [2] proposed an attention-based TabNet, which adaptively selects less important features to apply to tabular data, resulting in improved generalization performance. Additionally, Popov *et al*. [22] designed neural oblivious decision ensembles (NODE), which combines the hierarchical decision-making process of decision trees with the representation learning ability of deep learning networks. This integration allows NODE to capture complex interactions between input features.

To comprehensively deal with the challenge of the low correlation between features and labels and data imbalance, we propose a Tab-Attention, a self-attention-based stacked generalization learning approach, for predicting credit defaults. The method mainly includes the following three key modules:

- **Multi-view feature spaces**: For low correlation between features and labels in credit data, feature selection plays an important role in filtering key information. Various feature selection strategies can obtain feature sets with different advantages. Therefore, to focus on critical information and increase the informative diversity, we develop multiple feature importance-based strategies to organize multi-view feature spaces, mainly including linear correlation, nonlinear relationship, and similarity measures. Then, we design multi-layer perceptions (MLPs) to obtain outputs in different views as local-view knowledge. Besides, for all features, MLP was directly constructed to obtain global feature space. Notably, the feature sets with different importance provide key local view knowledge, which can reduce interference from less important information.
- **Self-attention-based stacked generalization learning**: The self-attention mechanism can dynamically weigh and combine key information. In this work, we develop a self-attention-based stacked generalization learning to efficiently ensemble information from multiple views. This approach learns ensemble mechanisms that can effectively combine the predictions of multiple models to improve the overall performance of the model.
- **Imbalanced learning based on $f1$ score**: To tackle the issue of imbalanced data, we adopt an $f1$ score to assist in training Tab-Attention, to find the optimal state in identifying default users, so as to achieve the goal of imbalanced learning.

2 Preliminaries

Credit default prediction constitutes a vital aspect of risk assessment and management within the financial sector. Default prediction evaluates the likelihood of borrowers fulfilling their repayment obligations by considering various factors, such as personal information, credit history, and loan characteristics, shown in Figure 1. In this work, credit default prediction is framed as a binary classification task, distinguishing between default (label 1) or no default (label 0). By developing a robust credit default prediction model, financial institutions could make informed decisions regarding credit approvals based on estimating default probabilities.

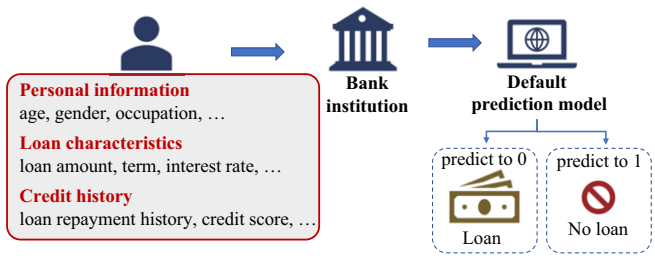

Figure 1. Credit approval process.

3 Methodologies

The low correlation and imbalance in credit data make it challenging to predict defaults accurately. To focus on important information, we first organize multi-view feature spaces by selecting features based on various linear or nonlinear importance indicators. We then employ MLP to obtain predictions from these local views. Simultaneously, for the global view of all features, we obtain a global feature space using MLP. To ensemble key knowledge for robust credit predictions, we develop self-attention modules to stack generalizing these obtained local and global outputs. Furthermore, we set the $f1$ score as training monitor indicators to adjust the learning rate to assist the model in finding the optimal state for identifying defaulting users, addressing the challenges posed by data imbalance. The code can be found in [1].

3.1 Multi-view feature spaces

In daily life, a general concept of "strong alliance" refers to the idea that by combining individuals with unique strengths, advantageous results can be obtained. Similarly, in feature selection, different approaches can lead to distinct advantageous views. To learn the multi-view information provided by these "strong alliance" feature sets, an automatic multi-view feature space generation method has been designed based on multiple screening strategies, shown in Figure 2. The following four types of views are considered:

- **Non-linearly correlated view**: This approach obtains the local-view feature spaces with the top 30% feature importance based on DT, GBDT, and RF methods. These models filter feature sets with non-linearly correlated importance based on information gain, which is important for default prediction models to focus on nonlinear mappings between labels and features.

  For $D = \{x_1, x_2, ..., x_n, y\}$, we give the example that calculates feature importance based on a DT using ID_3:

  1. Calculate the entropy of the target variable y as:

$$\Gamma(y) = -\sum(\Theta(y)/l) * \Theta(\Theta(y)/l), \tag{1}$$

  where $\Theta(y)$ refers to the number of occurrences of default and non-default samples in y, and l is the total number of instances.

[1] https://github.com/jinxtan/Tab_Attention/

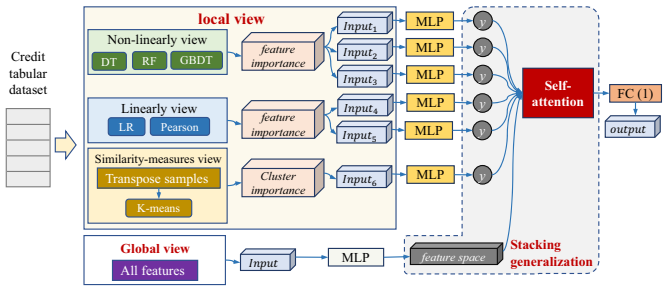

Figure 2. Multi-view feature spaces generation in Tab-Attention.

2. For each feature x_i, we calculate the information gain $x_{i_{IG}}$ as:

$$x_{i_{IG}} = \Gamma(y) - \sum \left(\frac{\Theta(x_i)}{l} \right) \cdot \Gamma(y|x_i), \tag{2}$$

where $\Gamma(y|x_i)$ is the entropy of y given feature x_i.

3. Select the feature with the highest $x_{i_{IG}}$ as the current division node.
4. Create a branch for each possible value of the selected x_i and recursively repeat steps 1-3 on the subsets of data corresponding to each branch.
5. Stop the recursion when:

(a) All samples belong to the same class.
(b) There are no more features to select.
(c) The maximum depth of the tree is reached.
(d) No enough instances could be split into a node.

6. Assign the majority class of the instances in the current subset as the class label for the leaf node.
7. Calculate the feature importance for x_i by accumulating its information gain $x_{i_{IG}}$ in the process of constructing the DT.

- **Linearly correlated view**: Linear correlation helps the model learn the linear mapping capability between features and the targets. Pearson correlation coefficient and LR are designed to select the features with the top 30% linear correlation importance, forming two local views.

For $D = \{x_1, x_2, ..., x_n, y\}$, the linear feature importance is calculated as,

1. For LR, the optimal model is obtained as,

$$y_{pred} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + ... + \beta_n \cdot x_n, \tag{3}$$

where β_0, β_1,..., β_n are the coefficients for each feature. We select the top 30% of features based on the absolute value of β.

2. For Pearson method, we obtain the correlation coefficient ι_i for feature x_i as,

$$\iota_i = cov(x_i, y)/(\sigma_{x_i} * \sigma_y), \tag{4}$$

where $cov(x_i, y)$ is the covariance between x_i and y, σ means the standard deviation. We also select the features based on the absolute of ι_i.

- **Similarity-measures view**: We employ the K-means to measure the similarity among features, then selects the features with the cluster with the largest number of features as a local view. Here, $K = 4$ is set to obtain multiple similarity feature sets that help the model learn information from a specific perspective rather than global information. The calculation process is as,

1. Transpose $D = \{x_1, x_2, ..., x_n, y\}$.
2. Utilize the K-means to partition the features into four clusters.
3. Enumerate the number of features within each cluster.
4. Select the cluster with the maximum count of features as the similarity-measures view.

- **Global view**: This view contains all features and offers the possibility of global optimization.

For each local view, MLPs are then designed to capture various possibilities. The global view aims to obtain a global feature space, while the local views aim to learn local knowledge. Here the MLPs model structure of the global view is larger than that of the local view.

3.2 *Self-attention-based stacked generalization*

Compared to tabular data, financial data typically exhibits lower feature correlations, higher sparsity, and greater imbalance, making it more challenging to accurately identify default users. Actually, uncovering critical knowledge is essential for enhancing predictive performance. Attention mechanisms, which focus on essential data based on activation significance, have led to significant breakthroughs in computer vision [4, 15] and natural language processing [10, 20]. Self-attention, a specific case of attention, learns the attention mechanism from the information provided by the data itself.

To address these challenges, we propose self-attention-based Tab-Attention to stack the knowledge from different feature views, as shown in Figure 3. Firstly, for different feature views $F = \{input_G, input_1, input_2, \ldots, input_6\}$ obtained in Section 3.1, we employ MLPs to acquire knowledge. The result is then normalized to reduce the risk of overfitting, as:

$$F^1 = Norm(Relu(W_1 * F + b_1)). \tag{5}$$

For the global view, we then also connect with the fully connected layer with fewer neurons to extract the global feature spaces as:

$$F_G^2 = Relu\left(W_2 * F_G^1 + b_2\right). \tag{6}$$

For the local views, we also connect a fully connected layer and then feed it into the output layer to obtain the predictions as:

$$F_i^3 = Sigmoid\left(W_3 * Relu\left(W_2 * F_G^1 + b_2\right) + b_3\right). \tag{7}$$

Here, we optimize the cross-entropy between real labels and F_i^3 to learn the exclusive nonlinear or linear mapping knowledge from the local view. Next, we concatenate the acquired local knowledge F_i^3 and global feature space F_G^2 into the self-attention layer to learn the importance of different elements, where the concatenated vectors F^* are linearly mapped to query matrix Q, key matrix K, and value matrix V, computed as,

$$F^* = concat(F_G^2, F_1^3, F_2^3, \ldots, F_6^3), \tag{8}$$

$$Q = W_Q * F^* + b_Q, \tag{9}$$

$$K = W_K * F^* + b_K, \tag{10}$$

$$V = W_V * F^* + b_V. \tag{11}$$

Then, we calculate the attention value α between every two input vectors as:

$$A = (\alpha)_{(i,j)} = softmax(Q * K). \tag{12}$$

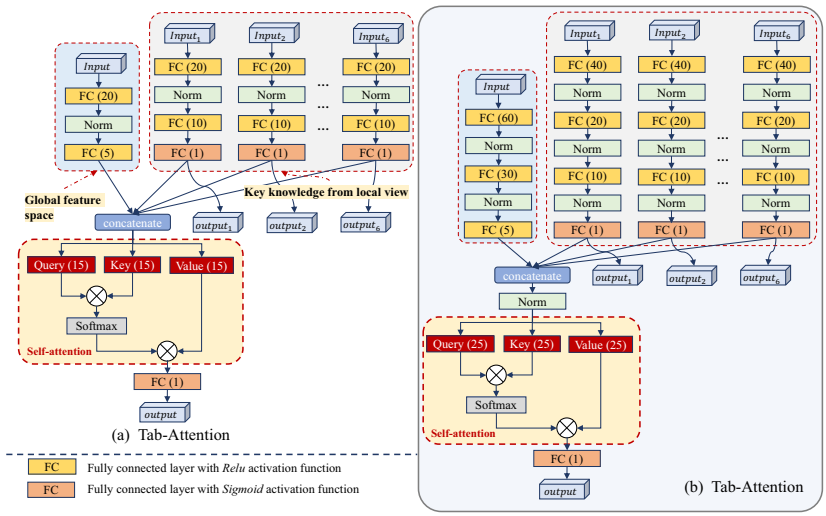

Figure 3. Tab-Attention model structure. (a) and (b) are the model structures of Tab-Attention for smaller and larger data sizes. The model described in the article is the structure of (a).

Based on the attention matrix A, we calculate the output vector $F^{attention}$ of the self-attention layer corresponding to each input vector α as,

$$F^{attention} = A * V. \tag{13}$$

Note that each element of $F^{attention}$ contains the associated information of other elements in F^* so as to capture a wider compositional mechanism. Finally, we employ a fully connected layer to learn the mapping relationship to the real labels.

$$y^{predict} = sigmoid(W_4 * F^{attention} + b_4). \tag{14}$$

For larger and smaller datasets, we configure different Tab-Attention model structures, as shown in Figure 3. Compared to smaller datasets, the Tab-Attention for larger ones involves a more complex model structure and normalization in all layers except for the output layer.

3.3 *Imbalanced learning based on $f1$ score*

Imbalanced credit data presents a significant challenge in default prediction. Traditional over-sampling [26] and under-sampling [17] methods are prone to cause overfitting or underfitting. Additionally, the performance of weighted cross-entropy methods [1] is limited, while generative model-based methods [11] are difficult to train and exhibit high complexity.

Traditionally, stopping training neural networks often relies on the convergence of overall loss or overall accuracy. However, this approach can inadvertently cause the model to neglect minority class samples. The $f1$ score, as the harmonic mean of $Precision$ and $Recall$, takes into account the prediction accuracy of default samples, calculated as,

$$f1 = 2 * (Precision * Recall)/(Precision + Recall), \tag{15}$$

where $Precision = TP/(TP+FP)$, $Recall = TP/(TP+FN)$. The TP and FP are the default samples with default and non-default prediction labels, respectively. The FN means non-default samples with default prediction labels.

To address the imbalance challenges, we optimize the training process of Tab-Attention based on $f1$ score as a training monitoring indicator. Specifically, when the $f1$ does not improve further after 10 epochs, we reduce the learning rate to 10% of the previous value, enabling the model to continue local optimization based on the current optimal state. The specific optimization process of imbalanced training is shown in Algorithm 1. Compared to the optimization process of traditional neural network training, this strategy facilitates the model to converge more rapidly to its optimal state of identifying default users.

Algorithm 1 Training Tab-Attention based on $f1$ score.

Require:
The initialized weights W, b and learning rate ξ_0 of neural networks, the best $f1$ score $f1_{best} = 0$, and $epoch_{count} = 0$.

for each training epoch i **do**
  Train Tab-Attention on training set using the Adam optimizer with ξ_0.
  if $f1 > f1_{best}$ **then**
    $f1_{best} = f1$
    $epoch_{count} = epoch_{count} + 1$.
  else
    $epoch_{count} = 0$.
  end if
  if $epoch_{count} = 10$ **then**
    $\xi_{i+1} = \xi_i * 0.1$.
    Update the learning rate in the Adam optimizer.
  end if
  if $epoch_{count} = 20$ **then**
    Break the loop.
  end if
end for
Output Tab-Attention.

4 Experimental results and analysis

4.1 *Dataset and preprocessing*

We utilized six credit datasets to validate the effectiveness of our model, shown in Table 1. We examined the data imbalance rate (IR) and the Pearson correlation quantiles (PCQ) between the features and labels, identifying challenges related to both imbalance and low correlation. We removed columns with ID attributes and over 60%

Table 1. Statistics for all credit default datasets.

Dataset	Data Size	Feature Num	PCQ 25%	PCQ 75%	IR
Zhongyuan	10,000	34	0.011	0.064	1:4.94
Taiwan	30,000	23	0.018	0.191	1:3.52
South German	1,000	20	0.033	0.155	1:2.33
Statlog	1,000	20	0.033	0.143	1:2.33
LC2018	197,178	84	0.015	0.071	1:3.18
LC2017	314,368	84	0.012	0.072	1:3.75

missing values and imputed the remaining missing values with the mode. The target variable in these datasets is the default status (1 for default and 0 for non-default).

- **Zhongyuan credit dataset**[2] (10,000 records, 34 features): from the 2021 CCF competition, focusing on individual credit default records and including loan records, user information, occupation, age, and marital status.
- **Taiwan credit dataset**[3] (30,000 samples, 23 attributes): donated by Chung Hua University and Tamkang University in 2016, including user information, education level, repayment records, and credit card bills.
- **South German credit dataset**[4] (1,000 records, 20 features): donated by Beuth University of Applied Sciences Berlin in 2019, including user information, loan amount, term, purpose, collateral, existing assets, and credit history.
- **Statlog credit dataset**[5] (1,000 records, 20 features): donated in 1994, including loan-related information, borrower personal information, and credit history.
- **LendingClub credit dataset**[6] (more than 197,178 samples, 84 attributes): containing data from the lending history of the US online lending platform LendingClub, including borrower personal information, loan information, credit rating and history, and loan purpose. We used data from 2018 and 2017.

Table 1 presents that the IRs of the utilized datasets exceed 1:2, indicating data imbalance. Furthermore, the 75% PCQ between the features and labels are below 0.2, indicating a weak correlation.

4.2 Evaluation metrics

Default prediction aims to minimize losses for financial institutions by denying loans to potential defaulters while ensuring access to creditworthy individuals. Therefore, we evaluate the model based on overall performance and class-specific performance:

- **Overall performance**: We evaluate accuracy (Acc), area under the curve (AUC), and Kolmogorov-Smirnov (KS). Acc measures the total number of samples classified correctly. AUC can comprehensively assess model performance with the class-imbalanced dataset. KS, a widely used evaluation metric in the financial domain, measures the risk discriminatory ability by calculating the maximum difference between cumulative positive and negative instances, calculated as,

$$KS = max(TPR - FPR), \tag{16}$$

where $TPR = TP/(TP+FN)$ and $FPR = FP/(FP+TN)$ are the cumulative results of two class samples, respectively.
- **Class-specific performance**: for defaulting and non-defaulting samples, we assessed the $Recall$, $Precision$, and $f1$ score to analyze the prediction capability for two classes.

4.3 Experimental setting

The experiments were conducted on an Ubuntu 20.04 server with a 5.8 Linux kernel, featuring Matrox G200eW3 GPU drivers operating at 3.90 GHz and 16 GB of main memory. In this study, each model configuration was executed 20 times, with the random state for splitting the data into training and testing sets set to values between 0 and 20 to mitigate the differences arising from data bias.

For the LC datasets, the batch size is configured to 300, 100 for the Taiwan dataset, 30 for Statlog and South German datasets, and 50 for Zhongyuan dataset. The initial learning rate for all datasets was set to 0.01. All data were normalized to minimize the impact of differences in scale.

Baseline: We compared our methods with three types of models: deep learning models (TabNet[2], NODE[22], and DNN), ensemble learning models (XGBoost[6, 14], AdaBoost, GBDT[7], and RF[16]), and machine learning-based single predictors (DT and LR [19]).

Parameter configurations: For TabNet, the number of neurons is set to 40 for the LC dataset, with 20 neurons for decision-making steps and 3 decision steps. For other datasets, the number of neurons is set to 20, with 10 neurons for decision-making steps. For the NODE model, we configured the number of neurons per layer to be 20 and employed 3 layers. Tree-based models have a maximum depth of 10, with 50 weak classifiers for ensemble models. The LR was optimized using L2 regularization with a regularization strength C of 0.1.

4.4 Performance comparison

Figure 4 displays the comparison results of model performance, and the analysis is as,

- For the Zhongyuan dataset, Tab-Attention and ensemble learning models (XGBoost, AdaBoost, GBDT, RF) achieve similar overall advantageous performance (Acc, AUC, KS). In terms of primary default prediction tasks, TabNet and NODE fail to recognize defaulting users, whereas Tab-Attention reaches a $Recall_1$ of 53.3% for default samples, outperforming other approaches by 28.7%, while maintaining outstanding $Precision_1$ of 50.4%. Furthermore, Tab-Attention attains $f1_1 = 51\%$, which is 16.4% better than other models.
- For the Taiwan dataset, similarly, Tab-Attention's $Recall_1 = 44.5\%$ surpasses other approaches by 20.9%, with a $f1_1$ of 50.4%, which is 7.2% better than other approaches. At the same time, Tab-Attention's overall performance and non-default sample prediction performance remain at a comparable level to the better results of other models.
- For the South German and Statlog datasets with smaller sample sizes, Tab-Attention could also recall more default users while maintaining a superior precision ($Precision_1 \approx 0.6$), which also upholds an advantageous level for non-default user prediction and overall performance.
- For the LC dataset, Tab-Attention's $Recall_1$ result exceeds other models by more than 65.8%, and its $f1_1$ score is over 31.7% better than other schemes. Meanwhile, the overall performance of

[2] https://www.datafountain.cn/competitions/530/datasets
[3] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[4] https://archive.ics.uci.edu/ml/datasets/South+German+Credit
[5] https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
[6] https://www.lendingclub.com

Tab-Attention is at a comparably superior level to DNN, TabNet, XGBoost, AdaBoost, GBDT, RF, and LR.

In general, compared to existing advanced methods, Tab-Attention achieves superior default prediction performance ($Recall_1$ and $f1_1$ are 32.92% and 16.05% better than other models on average, respectively), while maintaining outstanding overall performance and $Recall_0$ and $Precision_0$ for non-default users. Noteworthily, Tab-Attention's $Recall_0$ is lower than other models, but its $Precision_0$ is superior, implying that Tab-Attention is less likely to mistakenly predict default users as non-default ones with massive samples.

4.5 Ablation study

We develop multi-view feature spaces and self-attention stacking generalization components to assist the model in mining key knowledge for more robust default predictions. In addition, we designed f1-based imbalance learning. To verify the effectiveness of these components, we performed the following ablation study.

Multi-view feature space: Figure 5 shows the comparison results as more views feature spaces participate in the collaboration. As more views are incorporated, the KS progressively improves. Compared to Tab-Attention ($View = 0$) without multi-view feature spaces, the KS of Tab-Attention increases by 1.6%, 3.6%, 3.4%, and 3.8% for the Zhongyuan, Taiwan, South German, and Statlog datasets, respectively, while maintaining comparable levels of Acc and AUC. For the primary task of default identification, compared to Tab-Attention ($View = 0$), $Recall_1$ of Tab-Attention increases by 2.7%, 1.4%, 12.2%, and 13.9% for the four datasets, respectively. $Precision_1$ remains roughly equivalent level, and $f1_1$ displays an increasing trend, with the addition of more views. Additionally, $Recall_0$ gradually decreases as the views increase, while $Precision_0$ presents an increasing trend, which indicates that the addition of different views can reduce the risk of the model overlooking minority default users.

Overall, the cooperation of various views contributes to the model's ability to identify more default users and enhances the distinction between the two types of samples, resulting in the gradual improvement of the KS.

Self-attention-based stacked generalization: Then, we validated the necessity of ensembling important knowledge using the self-attention mechanism, as shown in Table 2. Overall, for the four datasets, the inclusion of self-attention increased the $Recall_1$ of default samples by more than 5%, and the $f1_1$ improved by 3.45% on average. Furthermore, Tab-Attention shows comparable overall performance (Acc, AUC, and KS) and the prediction results for non-default samples, compared to Tab*. This demonstrates that the self-attention mechanism is capable of ensembling essential knowledge to help the model better predict default users.

Imbalanced learning based on $f1$ score: Lastly, to validate the necessity of $f1$ in imbalanced learning, we compared it with AUC and Acc as monitoring indicators for the learning rate adjustment, shown in Table 2. Firstly, for the Zhongyuan dataset, it can be observed that the $Recall_1$ and $f1_1$ of the $f1$ training scheme are more than 17.8% and 6.3% better than that of Acc and AUC, respectively, while maintaining an excellent $Precision_1 = 0.5$. Meanwhile, $Recall_0$ of Acc and AUC training schemes is significantly higher than that of $f1$, but their $Precision_0$ is lower than that of $f1$, indicating that Acc and AUC schemes are more likely to misclassify default samples as normal samples. In addition, there is no significant difference in the Acc, AUC, and KS results among the three training schemes. For the Taiwan dataset, similarly, the $Recall_1$ and $f1_1$ of the $f1$ scheme are 25% and 6.4% better than other schemes, respectively. Furthermore, Acc and AUC schemes also expose the risk of recalling more normal samples as default samples (with superior $Recall_0$ but inferior $Precision_0$). For the South German and Statlog datasets, the performance of $f1$ in various indicators is significantly superior to that of $f1$ and AUC. Moreover, $f1$ imbalance learning converges more than 32.1% faster than the other two strategies, i.e. with fewer epochs.

Overall, the $f1$ training strategy is beneficial for Tab-Attention to accurately identify default users while maintaining superior overall Acc, AUC, KS, and non-default user prediction performance.

Table 2. Ablation comparison of f1-based imbalanced training and self-attention-based stacked strategies.

	Train metric	$f1$	AUC	Acc	Tab*
Zhongyuan	Acc	0.83 ± 0.01	$\mathbf{0.84 \pm 0.00}$	$\mathbf{0.84 \pm 0.00}$	0.84 ± 0.00
	AUC	0.86 ± 0.00	0.86 ± 0.00	$\mathbf{0.87 \pm 0.00}$	0.87 ± 0.00
	KS	$\mathbf{0.64 \pm 0.01}$	$\mathbf{0.64 \pm 0.01}$	0.62 ± 0.01	0.64 ± 0.01
	$Precision_0$	$\mathbf{0.91 \pm 0.01}$	0.89 ± 0.00	0.89 ± 0.00	0.89 ± 0.01
	$Recall_0$	0.89 ± 0.02	$\mathbf{0.93 \pm 0.01}$	0.91 ± 0.01	0.91 ± 0.01
	$f1_0$	0.90 ± 0.00	$\mathbf{0.91 \pm 0.00}$	0.90 ± 0.00	0.90 ± 0.00
	$Precision_1$	0.50 ± 0.01	$\mathbf{0.53 \pm 0.01}$	0.51 ± 0.01	0.51 ± 0.01
	$Recall_1$	$\mathbf{0.53 \pm 0.05}$	0.41 ± 0.02	0.45 ± 0.02	0.47 ± 0.04
	$f1_1$	$\mathbf{0.51 \pm 0.02}$	0.46 ± 0.01	0.48 ± 0.01	0.48 ± 0.02
	$epoch$	$\mathbf{38 \pm 5}$	77 ± 6	95 ± 5	40 ± 7
Taiwan	Acc	0.81 ± 0.00	$\mathbf{0.82 \pm 0.00}$	$\mathbf{0.82 \pm 0.00}$	0.81 ± 0.00
	AUC	0.76 ± 0.00	$\mathbf{0.77 \pm 0.00}$	$\mathbf{0.77 \pm 0.00}$	0.76 ± 0.01
	KS	0.41 ± 0.01	$\mathbf{0.42 \pm 0.00}$	$\mathbf{0.42 \pm 0.00}$	0.41 ± 0.01
	$Precision_0$	$\mathbf{0.85 \pm 0.00}$	0.84 ± 0.00	0.84 ± 0.00	0.85 ± 0.00
	$Recall_0$	0.91 ± 0.01	$\mathbf{0.95 \pm 0.00}$	$\mathbf{0.95 \pm 0.00}$	0.92 ± 0.01
	$f1_0$	0.88 ± 0.00	$\mathbf{0.89 \pm 0.00}$	$\mathbf{0.89 \pm 0.00}$	0.89 ± 0.00
	$Precision_1$	0.59 ± 0.02	$\mathbf{0.67 \pm 0.01}$	$\mathbf{0.67 \pm 0.01}$	0.61 ± 0.01
	$Recall_1$	$\mathbf{0.45 \pm 0.02}$	0.36 ± 0.01	0.36 ± 0.01	0.42 ± 0.01
	$f1_1$	$\mathbf{0.50 \pm 0.01}$	0.46 ± 0.01	0.47 ± 0.01	0.50 ± 0.01
	$epoch$	$\mathbf{31 \pm 1}$	66 ± 6	64 ± 9	32 ± 1
South German	Acc	$\mathbf{0.75 \pm 0.01}$	0.74 ± 0.01	0.73 ± 0.01	0.75 ± 0.01
	AUC	$\mathbf{0.75 \pm 0.01}$	$\mathbf{0.75 \pm 0.01}$	0.71 ± 0.01	0.75 ± 0.01
	KS	$\mathbf{0.43 \pm 0.02}$	0.42 ± 0.02	0.37 ± 0.02	0.42 ± 0.03
	$Precision_0$	$\mathbf{0.80 \pm 0.01}$	0.79 ± 0.01	0.79 ± 0.01	0.79 ± 0.01
	$Recall_0$	0.85 ± 0.02	$\mathbf{0.86 \pm 0.02}$	0.83 ± 0.01	0.87 ± 0.01
	$f1_0$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.82 \pm 0.01}$	0.81 ± 0.01	0.83 ± 0.01
	$Precision_1$	$\mathbf{0.60 \pm 0.02}$	0.59 ± 0.03	0.55 ± 0.02	0.60 ± 0.03
	$Recall_1$	$\mathbf{0.50 \pm 0.02}$	0.49 ± 0.02	0.49 ± 0.03	0.47 ± 0.02
	$f1_1$	$\mathbf{0.54 \pm 0.02}$	0.53 ± 0.01	0.51 ± 0.02	0.52 ± 0.02
	$epoch$	$\mathbf{56 \pm 6}$	74 ± 5	121 ± 12	59 ± 5
Statlog	Acc	$\mathbf{0.75 \pm 0.01}$	0.74 ± 0.01	0.73 ± 0.01	0.75 ± 0.01
	AUC	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.77 \pm 0.01}$	0.73 ± 0.01	0.77 ± 0.01
	KS	$\mathbf{0.46 \pm 0.02}$	$\mathbf{0.46 \pm 0.02}$	0.40 ± 0.03	0.45 ± 0.02
	$Precision_0$	$\mathbf{0.81 \pm 0.01}$	$\mathbf{0.81 \pm 0.01}$	0.8 ± 0.01	0.80 ± 0.01
	$Recall_0$	$\mathbf{0.84 \pm 0.02}$	0.83 ± 0.01	0.82 ± 0.01	0.85 ± 0.01
	$f1_0$	$\mathbf{0.83 \pm 0.01}$	0.82 ± 0.01	0.81 ± 0.01	0.83 ± 0.01
	$Precision_1$	$\mathbf{0.60 \pm 0.02}$	0.58 ± 0.02	0.55 ± 0.02	0.60 ± 0.02
	$Recall_1$	$\mathbf{0.55 \pm 0.03}$	0.54 ± 0.03	0.52 ± 0.03	0.51 ± 0.03
	$f1_1$	$\mathbf{0.57 \pm 0.02}$	0.56 ± 0.02	0.53 ± 0.02	0.55 ± 0.02
	$epoch$	$\mathbf{56 \pm 5}$	76 ± 7	117 ± 10	54 ± 5

[1] Tab* means Tab-Attention without self-attention.

5 Conclusion

In this work, to address the imbalanced and low feature-related credit default prediction, we proposed Tab-Attention, a novel credit default prediction framework that stacks important knowledge from multi-view feature spaces. Compared with the existing advanced GBDT-based and deep learning-based models, Tab-Attention improves the $Recall_1$ and $f1_1$ of default users by about 32.92% and 16.05% on average, respectively. We also observed that the collaboration of multi-view feature spaces helps Tab-Attention to better identify default risk, with an increasing trend of KS, $Recall_1$ and $f1_1$. In addition, the addition of self-attention is able to better ensemble important knowledge to further achieve superior $Recall_1$ and $f1_1$. Furthermore, we employed an $f1$-based imbalanced training strategy to assist the model to converge faster to the optimal state for identifying defaulting users. In future work, we would derive a theoretical justification on why Tab-Attention can better identify default users.

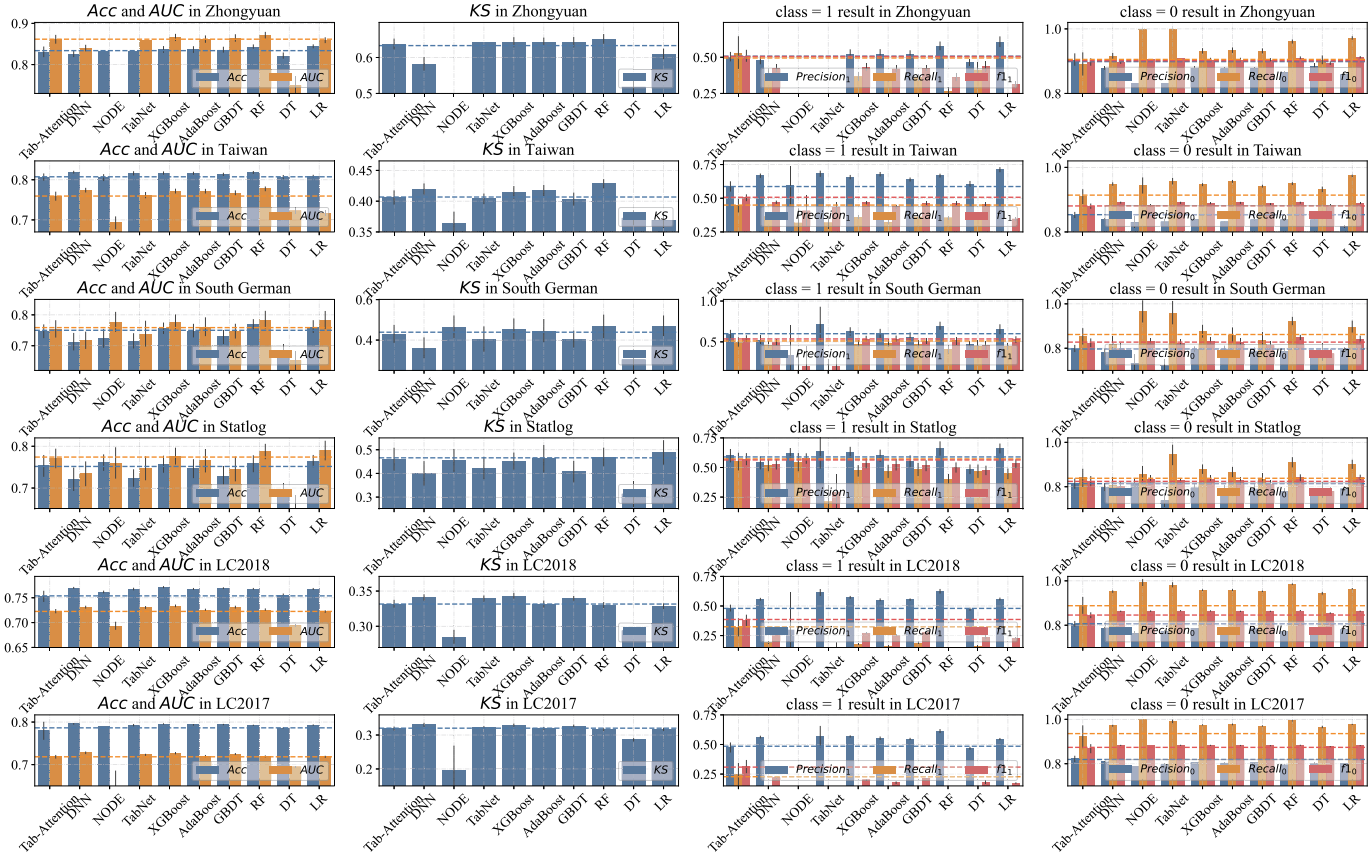

Figure 4. Performance comparison of Tab-Attention and current advanced models.

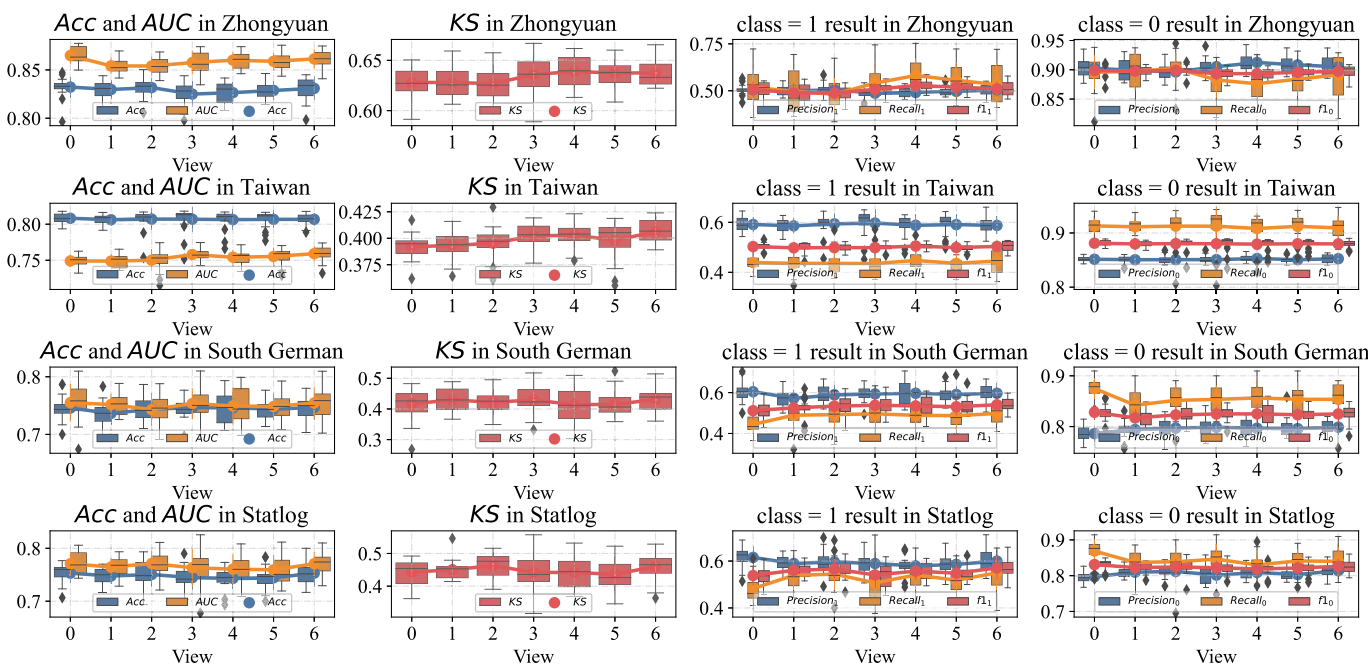

Figure 5. Results as the view increases. $View = 6$ means Tab-Attention, and $View = 0$ means DNN without the local-view model. 1, 2, 3, 4, 5, and 6 represent increasing views based on DT, GBDT, RF, LR, K-means, and Pearson, respectively.

Acknowledgements

We would like to thank the referees for their comments, which helped improve this paper considerably. The work is supported by the National Key Research and Development Program of China (2021YFC3300600).

References

[1] Mohamed Akil, Rachida Saouli, Rostom Kachouri, et al., 'Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy', *Medical image analysis*, **63**, 101692, (2020).

[2] Sercan Ö Arik and Tomas Pfister, 'Tabnet: Attentive interpretable tabular learning', *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(8), 6679–6687, (2021).

[3] Nisha Arora and Pankaj Deep Kaur, 'A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment', *Applied Soft Computing*, **86**, 105936, (2020).

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, 'Is space-time attention all you need for video understanding?', in *ICML*, volume 2, p. 4, (2021).

[5] Claudio Borio, 'The financial cycle and macroeconomics: What have we learnt?', *Journal of Banking & Finance*, **45**, 182–198, (2014).

[6] Yung-Chia Chang, Kuei-Hu Chang, and Guan-Jhih Wu, 'Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions', *Applied Soft Computing*, **73**, 914–920, (2018).

[7] Tianqi Chen and Carlos Guestrin, 'Xgboost: A scalable tree boosting system', in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, (2016).

[8] Stijn Claessens and M Ayhan Kose, 'Frontiers of macrofinancial linkages', *BIS Paper*, (95), 0–199, (2018).

[9] Lixin Cui, Lu Bai, Yanchao Wang, Xin Jin, and Edwin R Hancock, 'Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection', *Pattern recognition*, **114**, 107835, (2021).

[10] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan, 'Coatnet: Marrying convolution and attention for all data sizes', *Advances in Neural Information Processing Systems*, **34**, 3965–3977, (2021).

[11] Xi Fan, Xin Guo, Qi Chen, Yishuang Chen, Tongyao Wang, and Yuxin Zhang, 'Data augmentation of credit default swap transactions based on a sequence gan', *Information Processing & Management*, **59**(3), 102889, (2022).

[12] Carol Ann Frost, 'Credit rating agencies in capital markets: A review of research evidence on selected criticisms of the agencies', *Journal of Accounting, Auditing and Finance*, **22**(3), 469–492, (2007).

[13] Xiangling Fu, Tianxiong Ouyang, Jinpeng Chen, and Xiaopeng Luo, 'Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks', *Information Processing & Management*, **57**(4), 102236, (2020).

[14] Björn Rafn Gunnarsson, Seppe vanden Broucke, Bart Baesens, María Óskarsdóttir, and Wilfried Lemahieu, 'Deep learning for credit scoring: Do or don't?', *European Journal of Operational Research*, **295**(1), 292–305, (2021).

[15] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu, 'Beyond self-attention: External attention using two linear layers for visual tasks', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–13, (2022).

[16] Nan-Chen Hsieh and Lun-Ping Hung, 'A data driven ensemble classifier for credit scoring analysis', *Expert systems with Applications*, **37**(1), 534–545, (2010).

[17] Fayaz Itoo and Satwinder Singh, 'Comparison and analysis of logistic regression, naïve bayes and knn machine learning algorithms for credit card fraud detection', *International Journal of Information Technology*, **13**, 1503–1511, (2021).

[18] Cuicui Luo, Desheng Wu, and Dexiang Wu, 'A deep learning approach for credit scoring using credit default swaps', *Engineering Applications of Artificial Intelligence*, **65**, 465–470, (2017).

[19] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, and Yong Shi, 'Credit card churn forecasting by logistic regression and decision tree', *Expert Systems with Applications*, **38**(12), 15273–15285, (2011).

[20] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu, 'A review on the attention mechanism of deep learning', *Neurocomputing*, **452**, 48–62, (2021).

[21] Stjepan Oreski and Goran Oreski, 'Genetic algorithm-based heuristic for feature selection in credit risk assessment', *Expert systems with applications*, **41**(4), 2052–2064, (2014).

[22] Sergei Popov, Stanislav Morozov, and Artem Babenko, 'Neural oblivious decision ensembles for deep learning on tabular data', *arXiv preprint arXiv:1909.06312*, (2019).

[23] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, 'Catboost: unbiased boosting with categorical features', *Advances in neural information processing systems*, **31**, (2018).

[24] Akos Rona-Tas and Stefanie Hiss, *The role of ratings in the subprime mortgage crisis: The art of corporate and the science of consumer credit rating*, volume 30, 115–155, Emerald Group Publishing Limited, 2010.

[25] Yu Song, Yuyan Wang, Xin Ye, Dujuan Wang, Yunqiang Yin, and Yanzhang Wang, 'Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in p2p lending', *Information Sciences*, **525**, 182–204, (2020).

[26] Jie Sun, Jie Lang, Hamido Fujita, and Hui Li, 'Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates', *Information Sciences*, **425**, 76–91, (2018).

[27] Yandan Tan and Guangcai Zhao, 'Multi-view representation learning with kolmogorov-smirnov to predict default based on imbalanced and complex dataset', *Information Sciences*, **596**, 380–394, (2022).

[28] Mei Yang, Ming K. Lim, Yingchi Qu, Xingzhi Li, and Du Ni, 'Deep neural networks with l1 and l2 regularization for high dimensional corporate credit risk prediction', *Expert Systems with Applications*, **213**, 118873, (2023).

[29] Qiliang Zhu, Wenhao Ding, Mingsen Xiang, Mengzhen Hu, and Ning Zhang, 'Loan default prediction based on convolutional neural network and lightgbm', *International Journal of Data Warehousing and Mining (IJDWM)*, **19**(1), 1–16, (2023).