Journal of Intelligent Systems and Internet of Things
Vol. 13, No. 01, PP. 225-233, 2024
DOI: https://doi.org/10.54216/JISIoT.130116

Improving Loan Status Prediction Accuracy with Generative Adversarial Networks: Addressing Data Scarcity and Bias

Enas A. Raheem[1,*], Ahmed M. Dinar[1], Mazin Abed Mohammed[2], Bourair AL-Attar[3]

[1]Computer Engineering Department, University of Technology –Baghdad, Iraq
[2]Department of Artificial Intelligence, College of Computer Science and Information Technology, University of Anbar, Anbar 31001, Iraq
[3]College of Medicine, University of Al-Ameed, Karbala 1238, Iraq
Emails: enas.a.raheem@uotechnology.edu.iq; ahmed.m.dinar@uotechnology.edu.iq; mazinalshujeary@uoanbar.edu.iq; bourair.alattar@alameed.edu.iq

Abstract

A precise and reliable loan status prediction is of the essence for financial institutions, However, the lack of real-world data and biases within that data can greatly impact the accuracy of machine learning models. Another challenge faced by loan status prediction models is class imbalance, where one category (such as approved loans) is much more common than another (such as defaulted loans), leading to skewed predictions towards the majority class. This study inspects Generative Adversarial Networks (GANs) to augment the data and improve the machine learning models' performance. Several machine learning (ML) models including but not limited to Support Vector Machines (SVM) and ensemble bagged trees were employed on a Kaggle loan dataset (380 samples). Baseline training and testing accuracies were 86.9% and 86.3% (SVM) and 84.5% and 82.1% (ensemble). ActGAN (Activating Generative Networks) was then utilized to generate synthetic data points for both accepted and rejected loans. Retraining the models with new augmented data showed remarkable improvements: SVM accuracies for training and testing rose to 94.4% and 93.4%, while ensemble models achieved 97.4% and 95.8%, respectively. Other ML models were also explored such as KNN, Decision tree and logistic Regression and showed promising results in terms of accuracy as compared to the state of art. These findings put forward that GAN-based data augmentation can enhance the performance of loan status prediction. Future research could explore GAN's impact of different architectures and assess the general applicability of this approach.

Received: September 21, 2023 Revised: January 28, 2024 Accepted: June 07, 2024

Keywords: loan status; Machine learning; Generative Adversarial Networks; Prediction

1. Introduction

Reliable and accurate loan status prediction is a keystone for financial institutions, it enables them to assess creditworthiness and make loan decisions [1]. However, achieving high prediction accuracy faces two main challenges: lack of real-world data and potential blunders (i.e. Biases) within existing data. These restrictions on the availability of loan data due to privacy concerns and competitive edges can make it more difficult to develop ML models for loan status prediction. Likewise, the currently available data might have biases reflecting previous lending practices that can lead to unfair loan decisions [2].

Additionally, class imbalance is another challenge presents in loan data. In general, the ratio of granted loans is significantly higher than declined ones. This scarcity may distort the results of ML predictions by favouring the majority class (approved loans) and may possibly affect missed defaults and make higher risks for financial institutions [3] [4].

This work explores the potential of Generative Adversarial Networks (GANs) to address these challenges and improve the performance of ML models for loan status prediction. GANs are a type of deep learning architecture such that AI-generated synthetic data points that look like real data can be produced. By utilizing GANs, we aim to:

- To Perform data augmentation on existing loan data. Generating new synthetic data points for both accepted and rejected loans to effectively address class imbalance and data scarcity.
- To overcome data bias. The new data generated by GANs can be less susceptive to the biases existing in the original dataset. Eventually resulting in fairer and more accurate predictions.
- This work examines the effects of GAN-based data augmentation on different ML models that are usually applied for loan status prediction. We use a publicly available loan dataset from Kaggle that tests our model architecture performance with the and without augmented data. Our contributions include:
- To explore ActGAN (Activating Generative Networks) utility of a loan data augmentation.
- • Examine the impact of data augmentation on the accuracy of different ML models, including ensemble bagged trees Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and decision trees.
- To ilustrate the possible use of GANs to overcome the data scarcity and imbalance on loan patterns forecast and decrease the possibility of bias in the decision process.

The outcomes of this study can be very valuable for financial institutions trying to improve the performance of their loan prediction assessment systems that is great for decision making.

2. Related Work

Machine learning (ML) models have been widely exploited for loan status prediction. The most commonly used models are Decision Trees (DT), K Nearest Neighbour (KNN), Logistic Regression, Random Forest, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes, Adaboost, XGBoost, Deep Neural Network (DNN), Gradient Boosting, , and Long Short Term Memory network (LSTM) [5][6][7][8]. These methods have demonstrated promising results, especially logistic regression achieving up to 86.4% accuracy for loan prediction systems. Utilizing ML model in processing loan requests, does not only improve the accuracy of this process, but also reduces the time required to make decisions which is a plus for both banks, and applicants.
One of the most significant challenges of loan prediction is the limited data of real-world scenarios, because of privacy issues and competition advantages. This limitation can hinder the use of ML models. Several methods have been used to overcome this challenge like oversampling or SMOTE
That duplicates data points to increase the minority class [9]. In Addition, multi-view loan application graphs (MLAGs) and multi-view graph convolution networks (MGCN) has been proposed as a way to improve the small sample data for loan default risk prediction, particularly when there are imbalanced data distributions [10]. Bias on the other hand can also provide unfair model predictions. Different strategies have been utilized to mitigate this issue in a trade off with accuracy [11][12][13]. These techniques aimed to solve the bias problem without sacrificing the accuracy of loan assessment systems, leading to fair and more accurate decisions in credit underwriting.
This study expands the previous research attempts to deal with class imbalances and data scarcity through the use of GAN. Employing ActGAN to generate synthetic loan data points. As compared to existing data augmentation algorithms, our approach brings about significant improvements in terms of accuracy of loan predictions in a trade of with class balance, revealing the effectiveness of GAN for this particular task.

3. Loan Prediction System

In this paper, a system for loan prediction was developed twice, one time using a standalone machine leaning algorithms, utilizing several ML models and second time after applying Generative Adversarial Networks (GANs) to augment the data and improve the ML models' performance. The details for what was conducted are explained next and the system's overall structure before and after GAN is illustrated in figure 1 below.

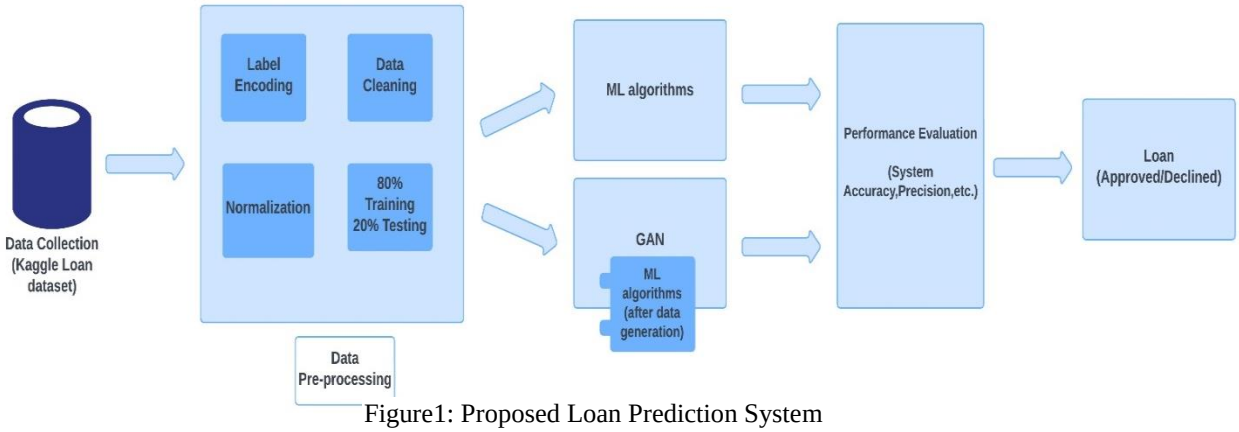


Figure1: Proposed Loan Prediction System

A. Data Collection and Pre-Processing

The dataset used in this paper was generally collected from the online website of Kaggle. 614 records with 13 different attributes (i.e. features) of the dataset are explained in Table 1. Pre-processing of data was done through several steps including:

i) Label encoding of categorical features, where categorical data was converted to numerical format using one hot encoding technique to ensure uniformity of dataset [14].

ii) Data cleaning for missing values to ensure dataset efficiency.

iii) Data Normalization using Z-Score was also conducted on dataset to scale data into a certain range [15].

iv) Data separation into 80% of the dataset for training and 20% for testing.

Table 1:Features of Kaggle Dataset

Feature	Description
Loan-id	An ID number unique for applicant
Gender	Female/Male
Marriage status	Married-Unmarried
Education	Grad/Undergrad.
Self-Employed	Yes/No
Dependents	No of dependents
Applicant-Income	Applicant's income
Co-Applicant-Income	Applicants partner income
Loan-Amount	Amount of Loan
Loan-Term	Repayment period of loan
Credit-History	Credit worthiness
Property-Area	Urban/rural/semi-urban
Loan-Status	Yes/No

B. ML Models

Several ML algorithms were utilized to test the performance of the prediction system before and after applying GAN, a significant improvement was noticed in the overall performance of the prediction system. In terms of accuracy and system's precision, the increment of data samples showed the impact of generating new data. ML models used in this study include machine learning models, like Support

Vector Machines (SVM), ensemble bagged trees, K-Nearest Neighbors (KNN), Decision Trees, and Logistic Regression, were employed for loan status prediction. The details for every model are illustrated below:

(i) Support Vector Machine: An SVM is a binary classifier where only two classes are involved to analyze data and distinguish the patterns. An SVM model is trained with data sets that are represented as points in space. A mapping is performed to divide the separated data categories by a certain gap and as wide as possible. Test data are then mapped into the same space and predictions are implemented as they belong to a certain class based on which side of the divided categories they fall on[16].

(ii) ensemble bagged trees: called a random forest classifier, enhances classification by combining several decision trees. It uses replacement to create random subsets of the training data, leveraging the idea of bagging (bootstrap aggregating). A collection of decision trees with some variety results from training a single decision tree for each distinct subset. A fresh data point is run through each tree in the ensemble during the prediction process. Each tree makes its own predictions, which are then aggregated by average (regression) or by majority vote (classification) to provide the final ensemble classification. Through the use of many learners' strengths, this method enhances overall accuracy while reducing the variance of individual trees, which are prone to overfitting[17].

(iii) K-Nearest Neighbors (KNN): (K-Nearest Neighbors) classifier. Utilizing a user-defined parameter, K, this non-parametric technique assigns a sample to the majority class among its K nearest neighbors[18].

(iv) Decision Trees: a machine learning technique that implements a flowchart as classification or regression algorithm. The hierarchy begins with an initial root node that stands for a complete dataset. The internal nodes on the tree, nodes which have splitting rules based on data features, are included as well. These procedures ask questions about one of the factors involved, and its response determines the next route - left or right. End node – leaves of the tree correspond to predicted outcome (class for classification, value for regression) [19].

C. Generative Adversarial Network (GAN):

The Generative Adversarial Networks (GANs) turned out to be a very efficient and critical method for creating synthetic data, especially in cases where actual data is limited or even non-existent. Here a GAN and its specific variant, called ActGAN, is discussed; these could be applied to a loan dataset and used to alleviate the data deficiency which results from limited resources.

i) Data Augmentation: The traditional machine learning models encounter difficulties when they are trained with overpriced datasets. GANs present a method through generating imaginary data that are not only realistic, but also replicate the real data's distribution. The Adversarial Approach: GANs involve 2 neural networks that are competing against each other, the winner gaining the numerical advantage.

-Generator Network (G): This interaction assembles fake data with equal probabilities of ending up with either real or fake data.

-Discriminator Network (D): This network turns into the right-searcher among actual data and falsely generated samples of G too. In the training process: G and D have separate optimizers too that are trained differently from each other. The generator determines a group of new data samples. A discriminator is responsible for splitting the data both real and fake and then the task is to classify them correctly (real or fake). Accumulated based on D's performance, weight updates for both (D to G) and (G to D) are propagated. This process is repetitive, wherein G continually enhances its capability to create realistic content, and D is heading towards becoming an expert in identifying fakes. Initially, G understands the data distribution and through the learning process succeeds in generating the important and high-quality data[21].

ii) ActGAN (Activating Generative Networks)

ActGAN is a further development of the foundation GAN model, which uses activation maps for additional information. These maps lay out what part of the data plays the most significant role in generating the output.
In the context of loan data augmentation, ActGAN can be particularly beneficial:
-Feature Extraction: Instead of that a second neural network processes pertinent pattern of characteristics out of real property data like income history.
credit score, and loan amount. However, in the future, lending businesses must strive to expand their range of services beyond traditional loans to attract a diverse pool of customers and maintain their competitive edge.
-Activation Map Generation: The functions which bring about the aforementioned maps are being done through this information processing, where the areas of concentration for generating artificial loan data is schematically depicted.
-Guiding the Generator: Such activation maps are passed alongside the Generator along with latent randomness vector. This extra bit of info implies that the Generator will yield to generate realistic figure only for the key elements of the made-up loan data.
Through equation of characteristics and activation maps, ActGAN has more specificity of data generation which is not typical in standard GANs. This in the generation of the loan data which very closely similar to the original data, can give statistics an ability to select the native data characteristics and properties. This additional data can be used to train more efficient models of machine learning since they will be able to recognize and make decisions in a wider scope, for example, loan applications or credit risk assessments[22].

D. System Performance Evaluation

The efficacy of the classification techniques was assessed by using suitable evaluation criteria, such as accuracy. To determine the accuracy of models, the test set, with both augmented and real data, was utilized. Furthermore, the HTER was calculated in order to measure the general error rate. The graphical results are also evaluated using area under curve (AUC). The formula to calculate accuracy and HTER is shown below :

$$\text{Accuracy} = (TP + TN)/ (TP + FN + TN + FP) \quad (1)$$

Where TP represents true positives, TN =true negatives, FP=false positives, and FN =false negatives [23]. HTER is calculated by the following equation:

$$\text{HTER}=100-\text{Accuracy} \quad (2)$$

The specific methodology proposed comprises the following steps, namely data collection, preprocessing, normalization, GAN usage for data augmentation, and implementation of ensembled tree, SVM, KNN classifiers for data classification.

4. Results and Discussion

This paper investigated the effectiveness of Networks (GANs) for data augmentation in loan status prediction models. Limited real-world data, class imbalance and biases in existing data were the main challenges that GAN was utilized to overcome. Several ML classifiers were used before and after GAN implementation to assess its impact on the prediction system. Table 2 show the results of those classifiers before GAN implementation.

Table 2: Classifiers Results Before GAN implementation.

Classifier	Accuracy(Training)	Accuracy(Testing)	HTER
SVM	85.9%	84%	16%
Ensemble Bagged Tree	86.5%	82%	18%
KNN	85.9%	85.6%	14.4%
Decision Tree	82.4%	80%	20%

The training accuracy represents the model's efficiency on the data which has been used for training it. High training accuracy implies good learning of patterns of the training data. Testing accuracy on the other hand reflects model's ability to generalize to unseen data. Indeed, all models have reasonable training accuracy (more than 80%). This indicates that they are able to learn well about the patterns of the data in the training set. SVM and KNN have the highest testing accuracy (about 84-86%), which provides proof that they are able to generalize well to unseen testing data.
Ensemble Bagged Tree's testing accuracy is lower (82%) than its high training accuracy, which could be the result of some overfitting probably.
Decision Tree is the model that shows the lowest overall accuracy amongst other algorithms, which may mean it does not display the underlying connections in the data as well as other algorithms. This leads to SVM and KNN exhibiting the smallest drop between training and testing accuracies, indicating good generalization abilities. Ensemble Bagged Tree alternatively shows a larger gap between training and testing accuracy, suggesting potential overfitting. A smaller HTER number means the bank was successful at approving the right loans and rejecting the wrong ones. The HTER value of KNN algorithm is the smallest one (14. 4%), which implies it may strike a perfect balance on minimizing false positives and false negatives. SVM and Ensemble Bagged Tree have the HTERs which are slightly higher. The models were then re-trained with the augmented dataset (original data + synthetic data). The Augmented data showed significant improvements in retraining, Results of classifiers after data augmentation using GAN are illustrated in table 3 below. The ROC graph for each classifier is also shown in figure 2.

Table 3: Classifiers Results After GAN

Classifier	**Accuracy(Training)**	**Accuracy(Testing)**	**HTER**	**AUC**
SVM	94.4%	93.4%	6.6%	0.97
Ensemble Bagged Tree	97.4%	95.8%	4.2%	0.98
KNN	93%	92.2%	7.8%	0.96
Decision Tree	96.5%	95.4%	4.6%	0.963

From the above table, several observations can be stated, one is accuracy boost for both training and testing data as they increase (8.6% - 17.4% and 10. 2% - 13.8% respectively). This suggests the models learn more accurate patterns upon augmented data. Yet, HTER reduction for all models (6.6% - 16%) implies that an attempt was made to properly predict the loans. AUC values close to 1 (0. 96-0. 98) were detected for a significant enhancement of the model's ability to distinguish between approved and rejected loans.

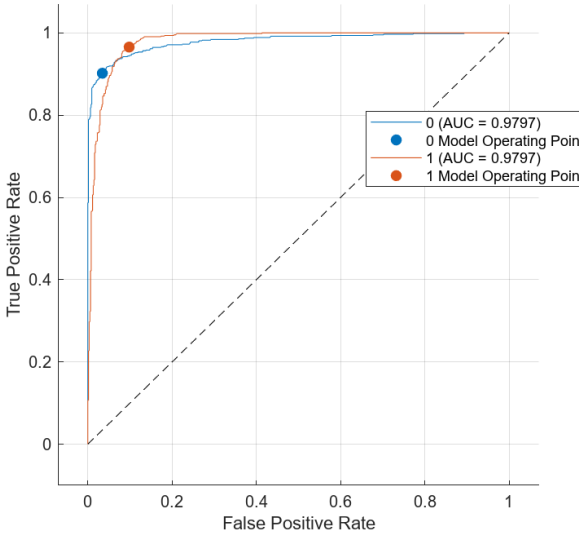
These results prove the efficiency of data augmentation via GAN to tackle the problem of lack of data. It also indicates an improved learning such that the models learn deeper representations because of the data's fine-tuning. Bias present in original data can be neutralized by the diverse yet synthetic data created by GAN. And finally, Models can maybe give a more generalized output by means of a more expanded training dataset. Using GAN to augment data substantially increases the efficiency of all the models in the context of this loan status prediction task. This analysis indicates that GAN-created data is a very helpful tool in the process of improving the accuracy of models, reducing misclassifications, and enhancing loan status prediction. Table 4 shows the results and improvements before and after data augmentation.

Table 4: Comparison of ML Results Before and after GAN

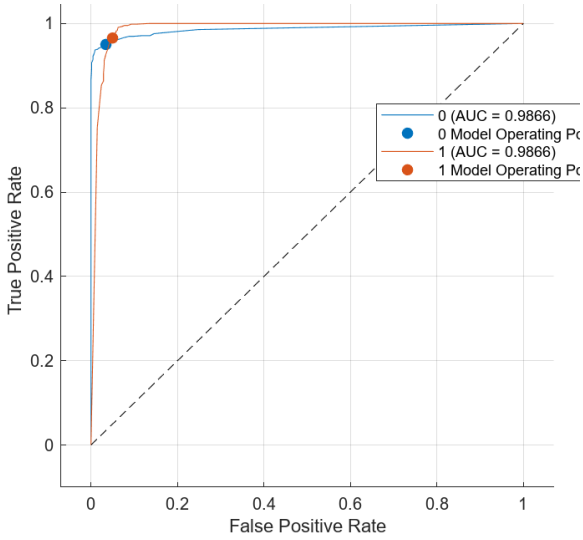
Metric	**Before Augmentation (Hypothetical)**	**After Augmentation (Current Results)**	**Improvement**
Training Accuracy	~ 80-85%	93.0% - 97.4%	+8.6% - 17.4%

Testing Accuracy	~ 80-82%	92.2% - 95.8%	+10.2% - 13.8%
HTER	~ 14% - 20%	4.0% - 7.8%	-6.2% - 16.0%

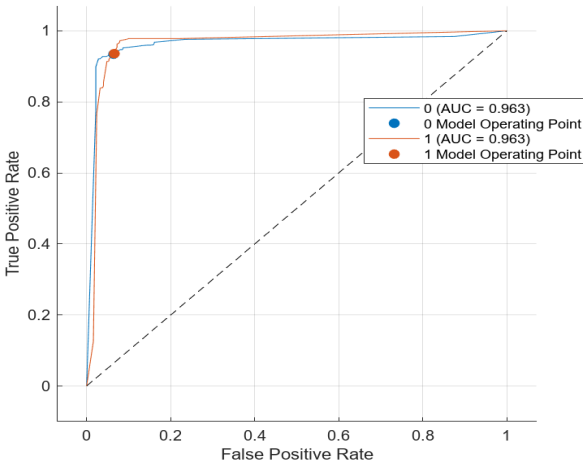
Figure 3 shows the ROC and AUC for all ML classifiers after GAN implementation. It reveals a notable increase in AUC metric as compared to the state of art. To evaluate the proposed loan status prediction system, the results were compared to existing work on the same dataset as a benchmark. Table 5 shows the results of our proposed system exceeding the performance of all benchmarks in terms of both accuracy and AUC. These findings prove that our model is successful in solving the problem of previous models and might bring a fresh approach to loan status prediction.



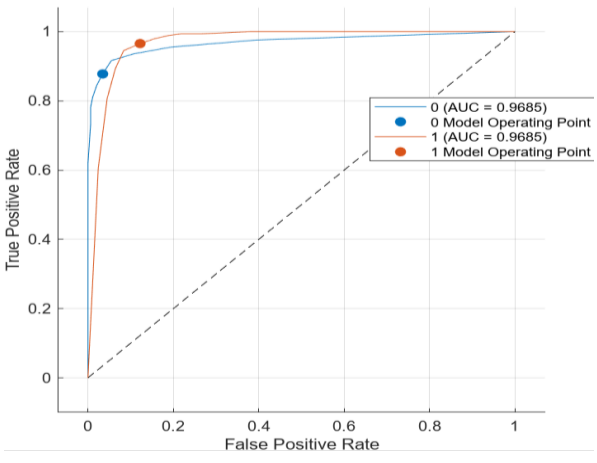
ROC for SVM



ROC for Ensemble BaggedTree



ROC for Decision Tree



ROC for KNN

Figure 3 : ROC of Different ML classifiers after GAN implementation.

5. Conclusion

This work explored the possibility of GANs to tackle the problems of machine learning models in loan status prediction. We identified the obstacles of insufficient data, bias, and imbalanced classes and have illustrated how generating datasets using GAN data augmentation can tackle such issues. Through the use of ActGAN to generate synthetic loan data points, we managed to improve the accuracy of several machine learning models to a great extent. Support Vector Machines (SVM) and bagged tree model ensemble were the best performers, the accuracy rates (both train and test) have been above 93%. The rest of the models such as K-Nearest Neighbors (KNN) and also decision trees, achieved a good result on the enhanced dataset. These results imply that the GAN-based augmentation technique provides a robust approach to enrich the performance of the loan status prediction models. The process of adjustment of data to reduce the bias and thus improve prediction accuracy will eventually benefit financial institutions in their loan assessment and decision-making process. In future research directions, it is possible to investigate the effect of GAN on different loan datasets and utilize various ML technologies for loan prediction.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] S. M. Fati, "a Loan Default Prediction Model Using Machine Learning and Feature Engineering," *ICIC Express Lett.*, vol. 18, no. 1, pp. 27–37, 2024.

[2] Z. Wang *et al.*, "Fairness-aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 10369–10378, 2022.

[3] M. Anand, A. Velu, and P. Whig, "Prediction of Loan Behaviour with Machine Learning Models for Secure Banking," *J. Comput. Sci. Eng.*, vol. 3, no. 1, pp. 1–13, 2022.

[4] J. L. Breeden, "A survey of machine learning in credit risk," *J. Credit Risk*, vol. 17, no. 3, pp. 1–62, 2021.

[5] A. S, "A Comparison of Various Machine Learning Algorithms and Deep Learning Algorithms for Prediction of Loan Eligibility," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 6, pp. 4558–4564, 2023.

[6] K. Bhatt, P. Sharma, M. Verma, and K. Agarwal, "Loan Status Prediction in the Banking Sector using Machine Learning," in *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, 2023, pp. 253–259.

[7] S. Wang, S. You, and S. Zhou, "Loan Prediction Using Machine Learning Methods," *Adv. Econ. Manag. Polit. Sci.*, vol. 5, no. 1, pp. 210–215, 2023.

[8] A. F. and M. M. Miraz Al Mamun, "Predicting Bank Loan Eligibility Using Machine Learning Models and Comparison Analysis," in *Proceedings of the 7th North American International Conference on Industrial Engineering and Operations Management, Orlando, Florida, USA,x*, pp. 1423–1432.

[9] G. Shingi, "A federated learning based approach for loan defaults prediction," in *2020 International Conference on Data Mining Workshops (ICDMW)*, 2020, pp. 362–368.

[10] L. Yu, X. Zhang, and H. Yin, "An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity," *Expert Syst. Appl.*, vol. 202, p. 117363, 2022.

[11] J. Liao, W. Wang, J. Xue, A. Lei, X. Han, and K. Lu, "Combating Sampling Bias: A Self-Training Method in Credit Risk Models," *Proc. 36th AAAI Conf. Artif. Intell. AAAI 2022*, vol. 36, pp. 12566–12572, 2022.

[12] A. Wu *et al.*, "Simultaneous Improvement of ML Model Fairnessand Performance by Identifying Bias in Data," *Nature*, vol. 388. pp. 1–14, 2020.

[13] A. Singh, J. Singh, A. Khan, and A. Gupta, “Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair,” *Mach. Learn. Knowl. Extr.*, vol. 4, no. 1, pp. 240–253, 2022.

[14] J. T. Hancock and T. M. Khoshgoftaar, “Survey on categorical data for neural networks,” *J. Big Data*, vol. 7, no. 1, 2020.

[15] R. H. Maharrani, P. D. Abda’u, and M. N. Faiz, “Clustering method for criminal crime acts using K-means and principal component analysis,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 34, no. 1, pp. 224–232, 2024.

[16] C. N. S.-T. J, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press, 2014.

[17] L. BREIMAN, “Bagging predictors,” in *Risks*, vol. 24, no. 3, 1996, pp. 123–140.

[18] A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja’afari, and S. Abdulwahed, “Diabetes classification based on KNN,” *IIUM Eng. J.*, vol. 21, no. 1, pp. 175–181, 2020.

[19] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, “Decision trees: from efficient prediction to responsible AI,” *Front. Artif. Intell.*, vol. 6, 2023.

[20] B. Caradima, A. Scheidegger, J. Brodersen, and N. Schuwirth, “Bridging mechanistic conceptual models and statistical species distribution models of riverine fish,” *Ecol. Modell.*, vol. 457, no. August, p. 109680, 2021.

[21] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, “Synthetic data generation for tabular health records: A systematic review,” *Neurocomputing*, vol. 493, pp. 28–45, 2022.

[22] T. P. Aki Koivu, Mikko Sairanen, Antti Airola, “Synthetic minority oversampling of vital statistics data with generative adversarial networks,” *J. Am. Med. Informatics Assoc.*, vol. 27, no. 11, pp. 1667–1674, 2020.

[23] Y. Dasari, K. Rishitha, and O. Gandhi, “Prediction of Bank Loan Status Using Machine Learning Algorithms,” *Int. J. Comput. Digit. Syst.*, vol. 14, no. 1, pp. 139–146, 2023.

[24] H. Li and W. Wu, “Loan default predictability with explainable machine learning,” *Financ. Res. Lett.*, vol. 60, 2024.

[25] D. Swapnesh Kumar Nayak, T. Swarnkar, and S. Kumari, “Loan Eligibility Prediction Using Machine Learning: a Comparative Approach,” *Glob. J. Model. Intell. Comput.*, vol. 3, no. 1, 2023.