A Comparative Analysis of Data Balancing Techniques:SMOTE and ADASYN in Machine Learning for Enhancing Bank Loan Default Risk Predictions

Manoj Rai[1],Bishal Ghimire[2],Nishant Uprety[3]Rubim Shrestha[4]

manoj.kiranti@gmail.com,[a] b4shal@gmail.com [b],[c]me.nishant2002@gmail.com,[d]rubimshrestha@kcc.edu.np
Bishal Ghimire,Kantipur City College, Kathmandu,44600,Nepal
Manoj Rai,Kantipur City College, Kathmandu,44600,Nepal
Nishant Uprety ,Thapathali Engineering Campus,, Kathmandu,44600,Nepal
Rubim Shrestha,Kantipur City College, Kathmandu,44600,Nepal

Abstract

Due to substantial technological advancements, people's needs have expanded. Consequently, there has been an increase in the number of loan approval requests in the banking industry. Several criteria are considered while selecting a candidate for loan approval in order to ascertain the loan's status. Banks encounter major challenges in evaluating loan applications and mitigating the risks linked to prospective borrower defaults. Due to the need to thoroughly assess the eligibility of every borrower for a loan, banks consider this process as notably burdensome. First the balancing of dataset will the performed. Recognizing the gravity of this task, the present study undertook the balancing of datasets as an imperative precursor, employing two oversampling techniques, SMOTE and ADASYN, for comparative analysis. The investigation aimed to discern the most efficacious balancing strategy for loan approvals by harnessing the analytical capabilities of algorithms such as Logistic Regression and Support Vector Machines (SVM). This exploration highlighted the distinct advantages and limitations intrinsic to each technique, underscoring the significance of aligning the choice with the dataset's unique attributes and the financial institution's objectives. Compellingly, the results of the study demonstrated that SMOTE, when paired with SVM, emerged as the superior method, yielding the highest accuracy rate of 93.55%, thereby recommending its application as a robust and generalizable strategy for enhancing the accuracy and reliability of loan approval processes in the banking sector.

Keywords: ADASYN; Loan; ML; Prediction; SMOTE; SVM

1. Introduction

In an era of unprecedented technological progress, the banking industry is experiencing a transformative shift to meet the diverse needs of its clientele. With a surge in loan approval requests driven by demands for homeownership, education, and entrepreneurship, there is a pressing need for sophisticated credit risk assessment methods. Traditional manual underwriting methods have proven inadequate, prompting financial institutions to adopt machine learning (ML) algorithms to enhance efficiency and accuracy in loan approvals.

Accurate prediction of loan eligibility has become critical, traditionally based on manual assessment but now significantly enhanced by ML techniques. A major challenge is the class imbalance in datasets, where instances

[1]
[2]
[3]
[4]

of loan defaults are far fewer than approvals, potentially skewing predictive models and leading to unfair decisions. Addressing this imbalance is crucial for technical accuracy, ethical standards, and regulatory compliance.

To mitigate class imbalance issues, techniques like SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling) have been developed. SMOTE creates samples by interpolating existing minority instances, while ADASYN focuses on generating samples near harder-to-learn minority instances, creating a more adaptive balance. While these techniques have shown potential in various fields, their impact on bank loan eligibility prediction requires further exploration.

1.1. Statement of Problem

In the banking sector, the decision-making process for loan eligibility is crucial. Despite advancements in efficiency and accuracy through machine learning (ML) techniques, a significant challenge persists: imbalanced datasets. Typically, loan eligibility datasets are skewed, with one class (e.g., loan defaults or rejections) being underrepresented compared to another (e.g., loan approvals). This imbalance can lead to biases in predictive models, favoring the majority class and resulting in unfair or inaccurate loan decisions.

The challenge is magnified when considering different ML models, such as Support Vector Machine (SVM) and Logistic Regression, each with unique data analysis and prediction approaches. The effectiveness of these models in handling imbalanced datasets in loan eligibility contexts is not fully explored. Moreover, while techniques like SMOTE (Synthetic Minority Over-Sampling Technique) and ADASYN (Adaptive Synthetic Sampling) aim to address class imbalance, their impact on the predictive accuracy and bias of these models in the banking sector remains unclear.

1.2. Research Objectives

- To compare SMOTE and ADASYN balancing technique with the help of machine learning algorithms.
- To build a robust model for bank loan default risk predictions

1.3. *Research Questions*

- How do SMOTE and ADASYN balancing techniques affect the performance of machine learning algorithms in predicting bank loan default risk?
- Which combination of SMOTE, ADASYN, and machine learning algorithms results in the most accurate and fair bank loan default risk prediction model?

2. Literature Review

In 1997, [1] et al. examined the challenges of imbalanced training sets in machine learning and proposed "one-sided selection" to improve classifier performance by selectively choosing training examples.In 1998, [2] et al. found that Naive Bayes performs well despite the assumption of feature independence, especially when feature distributions have low entropy .In 2002, [3] et al. demonstrated that combining over-sampling of the minority class with under-sampling of the majority class enhances classifier performance .In 2005, [4] et al. introduced borderline-SMOTE1 and borderline-SMOTE2, which showed superior performance in managing the minority class compared to traditional SMOTE .In 2008, [5] He et al. proposed ADASYN, which adaptively generates synthetic samples for challenging minority class instances, improving learning outcomes from imbalanced datasets .In 2009, [6] et al. developed EasyEnsemble and BalanceCascade, which use multiple subsets of the majority class to improve classifier performance while maintaining comparable training times .In 2009, [7] et al. emphasized the need for new principles and algorithms to handle imbalanced learning challenges in vast data-driven environments .In 2009, [8] et al. enhanced information retrieval systems by using balancing algorithms to improve classification performance .In 2009, [9] et al. introduced novel methods to improve the Naive Bayes model's precision and detection rate in email spam detection .In 2009, [10] et al. provided a comprehensive review of challenges and solutions associated with classifying imbalanced data .In 2010, [11] et al. found that generative oversampling significantly enhances results with linear SVMs for text datasets .In 2011, [12] et al. introduced the Structure Preserving Over-Sampling (SPO) method for imbalanced time series data classification, improving performance over traditional techniques .In 2014, [13] et al. introduced IRUSRT, which significantly outperformed existing methods for addressing class imbalance in 23 real-world datasets .In 2019, [14] et al. highlighted the SVM's evolution and its effectiveness in high-dimensional data applications like document classification .In 2021, [15] et al. developed a predictive model for early recovery from post-prostatectomy incontinence using preoperative MRI data .In 2021, [16] et al. showed that XGBoost provides robust predictive capabilities for identifying high-risk loan customers .In 2021, [17] et al. found that Random Forest performed best among several algorithms for predicting loan approvals and default risks .In 2022, [18] et al. used SMOTE-ENN to improve the performance of the Random Forest Classifier in predicting heart failure survivability .

In 2022, [19] et al. demonstrated the high accuracy of Random Forest and other machine learning algorithms in enhancing the loan approval process .In 2023, [20], [21] et al. developed a theoretical analysis of SMOTE, providing insights into the representativeness of generated samples .In 2023, [22] et al. found that random oversampling and hybrid approaches are effective for different degrees of class imbalance in educational datasets .In 2023, [23] et al. provided a comparative analysis of oversampling techniques like SMOTE, Borderline-SMOTE, and ADASYN for various machine learning tasks .In 2023, [24] et al. cautioned against the indiscriminate use of dataset balancing techniques due to varying impacts on different evaluation sets .In 2023, [25] highlighted the potential of machine learning models in revolutionizing the loan approval process in the banking sector .In 2023, [26] et al. identified the Naïve Bayes algorithm as the most effective for enhancing the loan approval process with an accuracy of 83.73% .In 2020, [27] et al. introduced a hybrid method combining SMOTE with ensemble machine learning models for bankruptcy prediction .In 2020, [28] et al. explored using SVMs for dimension reduction, showing improvements in estimation accuracy .In 2020, [29] et al. found that boosting techniques significantly outperformed traditional decision trees in loan approval predictions .In 2016, [30]et al. introduced GASMOTE, improving over traditional SMOTE by enhancing F-measure and G-mean metrics .In 2016, [31] et al. proposed new cost functions to improve logistic discrimination for imbalanced datasets In 2015, [32] et al. introduced a dynamic over-sampling approach using SMOTE and back-propagation to optimize neural network training for imbalanced datasets .In 2015, [33] et al. introduced MDOBoost, a technique combining boosting and over-sampling to enhance learning from multi-class imbalanced datasets .In 2023, [34] et al. showed that ensemble methods, particularly ExtraTrees with SMOTEENN, significantly improve the prediction of child health outcomes . In 2023, [34] et al. investigated the use of ensemble machine learning classifiers and class imbalance techniques like SMOTE, SMOTEENN, and SMOTETomek to predict diarrhoea in children under 5 years old. The study found that ensemble methods, particularly the ExtraTrees classifier with SMOTEENN, significantly outperformed traditional classifiers, achieving high recall, accuracy, precision, and F1-scores. These findings highlight the potential of ensemble methods to improve child health outcomes and aid policymakers in developing effective interventions.

3. Research Methodology

1.4. Data Collection

In the course of this research, the datasets were sourced from Kaggle, encompassing a total of 4,268 instances(approved – 2656 and rejected – 1613). These datasets comprise eleven input features, alongside a single output feature that categorizes the likelihood of a loan default into two distinct classes: negative and positive. Here, 'negative' signifies either the denial of a loan application or an instance of loan default, while 'positive' denotes the sanctioning of a loan.

1.5. Data Preprocessing

In the data pre-processing stage, categorical variables were first converted into a machine-readable format by encoding according to the label i.e. assigning unique integer identifiers to each category. The dataset was then split into training and testing subsets with an 80:20 ratio. Next, numerical features were standardized to ensure all features contributed equally to the model's predictions. Finally, class imbalance was addressed using ADASYN and SMOTE by generating synthetic instances of the minority class. ADASYN focused on hard-to-classify instances, while SMOTE interpolated between existing minority instances, thereby balancing the class distribution and reducing model bias towards the majority class

1.6. Machine Learning Algorithms

In this research, Support Vector Machines (SVM) were used for their powerful and versatile classification capabilities. SVMs constructed an optimal hyperplane to separate different class labels with the maximum margin, enhancing the classifier's generalization ability. They handled both linear and non-linear data through kernel functions, proving suitable for complex, high-dimensional datasets. The reliance on support vectors made SVMs memory efficient and robust, particularly in applications like bioinformatics and text classification. Logistic Regression (LR) was employed to model the probability of a binary outcome using the sigmoid function, transforming real-valued inputs into a range between 0 and 1. The model created a decision boundary to classify data into two categories based on a probability threshold. Logistic Regression provided interpretable coefficients and performed well when there was a linear relationship between features and the log-odds of the outcome. It was particularly effective with a smaller number of features, reducing the risk of overfitting.

Both SVM and LR were utilized to verify the effectiveness of the applied balancing techniques, such as SMOTE and ADASYN, in removing biases inherent to one model. By evaluating both models, the research aimed to ensure that the balancing techniques improved predictive performance and fairness across different algorithms, demonstrating the robustness of the methods in handling class imbalance.

1.7. Balancing Techniques

In the realm of data analysis, particularly where dataset imbalance is a significant challenge, balancing techniques play a crucial role. Two prominent techniques used in this research are SMOTE (Synthetic Minority Over-Sampling Technique) and ADASYN (Adaptive Synthetic Sampling). Both techniques generate synthetic instances of the minority class to balance the class distribution, which is common in applications such as fraud detection, medical diagnosis, and sentiment analysis.

SMOTE addresses class imbalance by generating synthetic instances of the minority class through interpolation between existing minority instances and their nearest neighbors. This technique balances the class distribution, reducing the model's bias towards the majority class and enhancing predictive performance.

ADASYN extends SMOTE by focusing on minority instances that are harder to learn, generating more synthetic samples for these challenging instances based on their local distribution. This adaptive approach ensures that the oversampling process targets areas where additional data is most beneficial, further improving model performance and fairness. Both SMOTE and ADASYN were employed in this research to evaluate and enhance the effectiveness of machine learning models in handling imbalanced datasets, ensuring unbiased and accurate predictions.

4. Results and Analysis

1.8. Dataset after Balancing

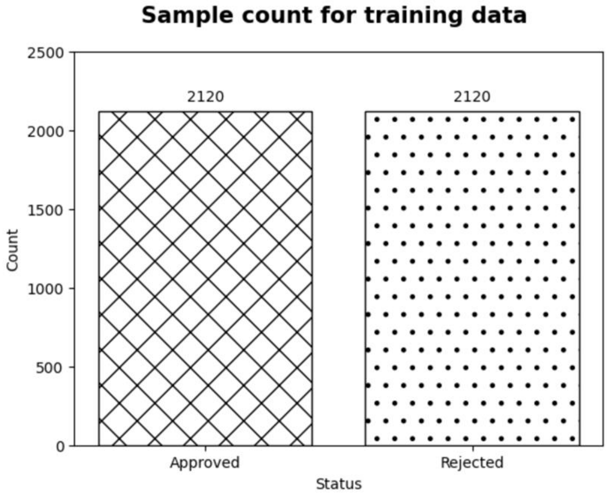

Figure 1: Dataset after SMOTE

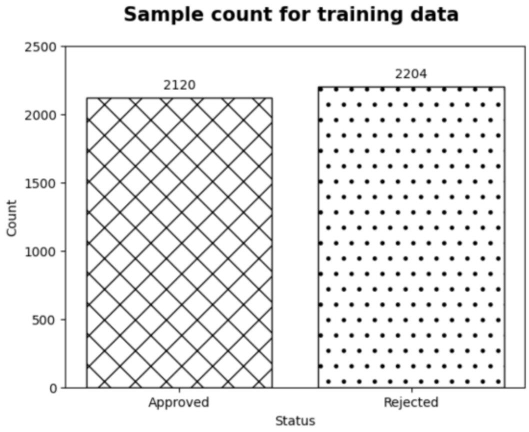

Figure 2: Dataset after ADASYN

1.9. Experiment Results

Table 1: Experiment Results

Data Condition	Technique	Classifier	Accuracy
Unbalanced	-	LR	90.39
Unbalanced	-	SVM	88.76
Balanced	SMOTE	LR	91.68
Balanced	ADASYN	LR	91.33
Balanced	SMOTE	SVM	93.55
Balanced	ADASYN	SVM	93.20

The results clearly illustrate the benefits of using oversampling techniques to manage class imbalance. Both SMOTE and ADASYN not only improved the accuracy of LR and SVM models compared to their performance on the unbalanced dataset but also highlighted the nuanced differences in their effectiveness. SMOTE consistently provided a slight edge in performance for both models, indicating its robustness as a general-purpose oversampling technique.

1.10. SMOTE vs ADASYN

SMOTE's approach to synthesizing new samples by interpolating between existing minority class samples was found to be marginally more effective than ADASYN. This effectiveness is attributed to SMOTE's method of creating synthetic samples that are not just replicas of existing minority instances but are new points along the line segments joining any/all of the k minority class nearest neighbors. Hence, SMOTE tends to expand the feature space where the minority class is underrepresented, allowing the classifiers to draw more generalized decision boundaries. ADASYN also contributes to balancing class distribution but does so by creating more synthetic data for those minority class samples that are harder to learn. This adaptivity means that ADASYN can focus on the regions of the feature space where the classifier is most likely to benefit from more information. However, this targeted approach may not always capture the broader underlying patterns as effectively as SMOTE, which could explain why SMOTE achieved slightly better performance metrics in this study

1.11. Best Model: SVM + SMOTE

Table 2: Classification Metrics

Class	Precision	Recall	F1-score	Support
0	0.96	0.93	0.95	536
1	0.89	0.94	0.92	318

The classification report for the SMOTE + SVM model demonstrates its effectiveness in predicting loan defaults. For non-default loans (class 0), the model achieved a precision of 0.96, recall of 0.93, and an F1-score of 0.95, indicating a low false positive rate and strong overall performance. For default loans (class 1), it achieved a precision of 0.89, recall of 0.94, and an F1-score of 0.92, highlighting its ability to accurately identify the majority of default cases with a slight trade-off in precision. These results underscore the model's robustness and reliability in classifying both loan defaults and non-defaults, making it a valuable tool for credit risk assessment.

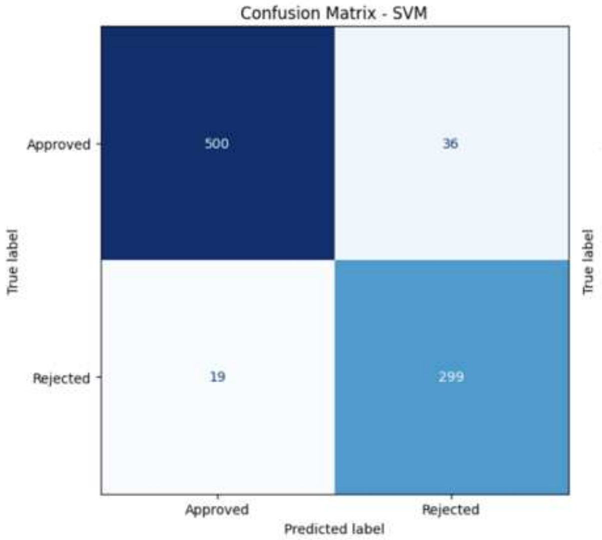

Figure 3: Confusion Matrix for Best Model

The confusion matrix for the SMOTE + SVM model shows that it effectively identifies loan defaults with 500 true positives and non-defaults with 299 true negatives. It has 19 false positives (non-default loans incorrectly predicted as defaults) and 36 false negatives (default loans incorrectly predicted as non-defaults). These results highlight the model's robustness and reliability in credit risk assessment, demonstrating its ability to accurately classify the majority of both default and non-default loan cases with minimal misclassifications.

5. Discussion on Results

Oversampling techniques like SMOTE and ADASYN significantly improved the accuracy of Logistic Regression (LR) and Support Vector Machine (SVM) models compared to their performance on an unbalanced dataset. SMOTE consistently provided a slight edge in performance, demonstrating its robustness as a general-purpose oversampling technique.

1.12. *Advantages of Oversampling Techniques*

The class imbalance problem is a well-known challenge in machine learning, particularly in scenarios such as loan default prediction, where the minority class (defaulting loans) is of great interest but is underrepresented in the dataset. Oversampling techniques like SMOTE and ADASYN work by generating synthetic examples of the minority class, thereby creating a more balanced class distribution. This balance allows for more effective learning, as the classifiers are exposed to sufficient examples of both classes during training, reducing the bias towards the majority class that typically occurs with imbalanced data.

1.13. *SMOTE vs ADASYN*

SMOTE's approach to synthesizing new samples by interpolating between existing minority class samples was found to be marginally more effective than ADASYN. This effectiveness is attributed to SMOTE's method of creating synthetic samples that are not just replicas of existing minority instances but are new points along the line segments joining any/all of the k minority class nearest neighbors. Hence, SMOTE tends to expand the feature space where the minority class is underrepresented, allowing the classifiers to draw more generalized decision boundaries.

ADASYN also contributes to balancing class distribution but does so by creating more synthetic data for those minority class samples that are harder to learn. This adaptivity means that ADASYN can focus on the regions of the feature space where the classifier is most likely to benefit from more information. However, this targeted approach may not always capture the broader underlying patterns as effectively as SMOTE, which could explain why SMOTE achieved slightly better performance metrics in this study

1.14. *Implication for Predictive Modeling*

The subtle yet consistent superiority of SMOTE in this research suggests that when dealing with imbalanced datasets, SMOTE could be the preferred initial choice for model training enhancement. The fact that the SVM model, tuned with its optimal parameters, showed the highest accuracy when used in conjunction with SMOTE, further supports this claim. However, it is important to recognize that while SMOTE improved accuracy, the ultimate choice of oversampling technique should also consider other performance metrics such as precision, recall, and F1-score, particularly in applications where the costs of false positives and false negatives are highly disproportionate.

6. Conclusion

The conducted research analyzed the impact of oversampling techniques on class imbalance, focusing on SMOTE and ADASYN. Both techniques significantly improved the accuracy of Logistic Regression (LR) and Support Vector Machine (SVM) models compared to their performance on an unbalanced dataset. SMOTE yielded higher accuracies, with the LR model achieving 91.68% and the SVM model 93.55%, highlighting its effectiveness in enhancing model robustness. ADASYN also enhanced model performance, with the LR model achieving 91.334% and the SVM model 93.20%.. The comparison indicated that SMOTE was marginally more effective, particularly with the SVM model, which emerged as the most proficient combination for handling class imbalance and maximizing classification accuracy.

Acknowledgements

I would like to extend my heartfelt appreciation to my exceptional supervisor, Rubim Shrestha, whose unwavering support and encouragement inspired me to pursue this paper. His invaluable guidance and advice were instrumental in making this achievement possible. I am also deeply grateful to Dr. Gajendra Sharma for his support and direction. Furthermore, I wish to thank my parents and sister for their moral support, and my friend Nishant Uprety for his expertise and recommendations..

References

[1] Miroslav Kubát, M. Kubat, Stan Matwin, and S. Matwin, “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection.,” pp. 179–186, Jan. 1997.

[2] Andrew McCallum, A. McCallum, Kamal Nigam, and K. Nigam, “A comparison of event models for naive bayes text classification,” *AAAI Conference on Artificial Intelligence*, pp. 41–48, Jan. 1998.

[3] Nitesh V. Chawla *et al.*, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, Jan. 2002, doi: 10.1613/jair.953.

[4] Hui Han, H. Han, Wenyuan Wang, W. Wang, Binghuan Mao, and B. Mao, “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning,” pp. 878–887, Aug. 2005, doi: 10.1007/11538059_91.

[5] Haibo He *et al.*, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *IEEE World Congress on Computational Intelligence*, pp. 1322–1328, Jun. 2008, doi: 10.1109/ijcnn.2008.4633969.

[6] Xuying Liu, X.-Y. Liu, Jianxin Wu, J. Wu, Zhi-Hua Zhou, and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," vol. 39, no. 2, pp. 539–550, Apr. 2009, doi: 10.1109/tsmcb.2008.2007853.

[7] Haibo He, H. He, E.A. Garcia, and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/tkde.2008.239.

[8] Pablo Bermejo *et al.*, "Comparison of balancing techniques for multimedia IR over imbalanced datasets," pp. 674–679, Oct. 2009, doi: 10.1109/iscis.2009.5291904.

[9] Yang Song, Y. Song, Yang Song, A. Kolcz, Aleksander Kolcz, and C. L. Giles, "Better Naive Bayes classification for high-precision spam detection," *Software - Practice and Experience*, vol. 39, no. 11, pp. 1003–1024, Aug. 2009, doi: 10.1002/spe.v39:11.

[10] Yanmin Sun *et al.*, "CLASSIFICATION OF IMBALANCED DATA: A REVIEW," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, Jun. 2009, doi: 10.1142/s0218001409007326.

[11] Alexander Liu *et al.*, "Effects of Oversampling Versus Cost-Sensitive Learning for Bayesian and SVM Classifiers," pp. 159–192, Jan. 2010, doi: 10.1007/978-1-4419-1280-0_8.

[12] Hong Cao *et al.*, "SPO: Structure Preserving Oversampling for Imbalanced Time Series Classification," pp. 1008–1013, Dec. 2011, doi: 10.1109/icdm.2011.137.

[13] Chunxia Zhang *et al.*, "IRUSRT: A Novel Imbalanced Learning Technique by Combining Inverse Random Under Sampling and Random Tree," *Communications in Statistics - Simulation and Computation*, vol. 43, no. 10, pp. 2714–2731, Jun. 2014, doi: 10.1080/03610918.2013.765467.

[14] V. K. Chauhan *et al.*, "Problem formulations and solvers in linear SVM: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 803–855, Aug. 2019, doi: 10.1007/s10462-018-9614-6.

[15] Seongkeun Park, S.-H. Park, S. Park, Jieun Byun, J. Byun, and Jieun Byun, "A Study of Predictive Models for Early Outcomes of Post-Prostatectomy Incontinence: Machine Learning Approach vs. Logistic Regression Analysis Approach," *Applied Sciences*, vol. 11, no. 13, p. 6225, Jul. 2021, doi: 10.3390/app11136225.

[16] M I Omogbhemhe, M. I. Omogbhemhe, Momodu I.B.A., and M. I.B.A., "Model for Predicting Bank Loan Default using XGBoost," *International Journal of Computer Applications*, vol. 183, no. 32, pp. 1–4, Oct. 2021, doi: 10.5120/ijca2021921705.

[17] Krishan Kumar Pandey *et al.*, "Predictive Analysis of Classification Algorithms on Banking Data," Sep. 2021, doi: 10.1109/gucon50781.2021.9573792.

[18] Mirza Muntasir Nishat *et al.*, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," *Scientific Programming*, vol. 2022, pp. 1–17, Mar. 2022, doi: 10.1155/2022/3649406.

[19] Ugochukwu Orji *et al.*, "Machine Learning Models for Predicting Bank Loan Eligibility," *2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON)*, Apr. 2022, doi: 10.1109/nigercon54645.2022.9803172.

[20] Anthony Anggrawan, Hairani Hairani, and Christofer Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *International Journal of Information and Education Technology*, vol. 13, no. 2, pp. 289–295, Jan. 2023, doi: 10.18178/ijiet.2023.13.2.1806.

[21] Dina Elreedy, Amir F. Atiya, and Firuz Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine-mediated learning*, Jan. 2023, doi: 10.1007/s10994-022-06296-4.

[22] Tarid Wongvorachan, Surina He, and Okan Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, pp. 54–54, Jan. 2023, doi: 10.3390/info14010054.

[23] Ishani Dey and Vibha Pratap, "A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques using Different Classifiers," *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, Mar. 2023, doi: 10.1109/icsmdi57622.2023.00060.

[24] Robert C. Moore, Daniel P. W. Ellis, Eduardo Fonseca, Shawn Hershey, Aren Jansen, and Manoj Plakal, "Dataset Balancing Can Hurt Model Performance," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Jun. 2023, doi: 10.1109/icassp49357.2023.10095255.

[25] Archana Archana, "A Comparison of Various Machine Learning Algorithms and Deep Learning Algorithms for Prediction of Loan Eligibility," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 6, pp. 4558–4564, Jun. 2023, doi: 10.22214/ijraset.2023.54495.

[26] V. V, R. A. C, V. K N, and A. G, "Prediction of Loan Approval in Banks Using Machine Learning Approach." Rochester, NY, Aug. 04, 2023. Accessed: Jan. 20, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4532468

[27] Hossam Faris *et al.*, "Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case from the Spanish market," *Progress in Artificial Intelligence*, vol. 9, no. 1, pp. 31–53, Mar. 2020, doi: 10.1007/s13748-019-00197-9.

[28] Luke Smallman, L. Smallman, Andreas Artemiou, and A. Artemiou, "A study on imbalance support vector machine algorithms for sufficient dimension reduction," *Communications in Statistics-theory and Methods*, vol. 46, no. 6, pp. 2751–2763, Mar. 2017, doi: 10.1080/03610926.2015.1048889.

[29] Mohamed Alaradi, M. Alaradi, Mohamed Alaradi, Sawsan Hilal, S. Hilal, and Sawsan Hilal, "Tree-Based Methods for Loan Approval," 2020, doi: 10.1109/icdabi51230.2020.9325614.

[30] Kun Jiang, K. Jiang, Jing Lu, J. Lu, Kuiliang Xia, and K. Xia, “A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE,” *Arabian Journal for Science and Engineering*, vol. 41, no. 8, pp. 3255–3266, May 2016, doi: 10.1007/s13369-016-2179-2.

[31] Huaping Guo *et al.*, “Logistic discrimination based on G-mean and F-measure for imbalanced problem,” *Journal of Intelligent and Fuzzy Systems*, vol. 31, no. 3, pp. 1155–1166, Jan. 2016, doi: 10.3233/ifs-162150.

[32] R. Alejo, R. Alejo, Vicente García, V. García, J. H. Pacheco-Sánchez, and J. H. Pacheco-Sánchez, “An Efficient Over-sampling Approach Based on Mean Square Error Back-propagation for Dealing with the Multi-class Imbalance Problem,” *Neural Processing Letters*, vol. 42, no. 3, pp. 603–617, Dec. 2015, doi: 10.1007/s11063-014-9376-3.

[33] Lida Abdi, L. Abdi, Sattar Hashemi, and S. Hashemi, “To combat multi-class imbalanced problems by means of over-sampling and boosting techniques,” vol. 19, no. 12, pp. 3369–3385, Dec. 2015, doi: 10.1007/s00500-014-1291-z.

[34] Elliot Mbunge *et al.*, “Implementation of ensemble machine learning classifiers to predict diarrhoea with SMOTEENN, SMOTE, and SMOTETomek class imbalance approaches,” Mar. 2023, doi: 10.1109/ictas56421.2023.10082744.

[35] S. SATPATHY, “SMOTE for Imbalanced Classification with Python,” Analytics Vidhya. Accessed: Feb. 24, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/