Towards Addressing Bias and Fairness in Machine Learning

Rudraksh Khandelwal[1], Shyam Deshmukh[2]

1 Pune Institute of Computer Technology (Department of Information Technology), Pune, Maharashtra, India, rudrakshkhandelwal2912@gmail.com

2 Pune Institute of Computer Technology (Department of Information Technology), Pune, Maharashtra, India, sbdeshmukh@pict.edu

Abstract:

The aim is to unravel the process through which ML models iteratively learn and adapt, consequently enhancing their predictive accuracy. Such an exploration is crucial for a comprehensive understanding of the fundamental principles underpinning ML functionality and its practical applications. The breadth of this analysis extends to an examination of the continuous learning processes within ML models, and the implications these have in various real-world domains, such as healthcare and finance. This analysis probes into the manner in which ML models evolve and adapt over time, laying a foundational basis for both experimental and practical applications. This groundwork is pivotal for delving into the continuous improvement and wider applicability of ML models in diverse scenarios, thus highlighting the criticality of ongoing learning and adaptability in artificial intelligence.

A significant focus of this analysis is on the issue of bias in ML, which is often introduced through the data utilized for training. This bias may reflect existing societal biases or arise from data collection methods that do not represent the full spectrum of diversity. The overlook of diverse data sources or erroneous labeling could result in biased outcomes, underlining the need for meticulous data selection and preprocessing. Upholding fairness in ML transcends technical challenges, aligning technological advancements with societal values. This study contributes towards fostering a more equitable and inclusive future, utilizing AI as a positive force in a globally interconnected landscape. The extensive underrepresentation and misrepresentation of protected groups in training datasets gravely affect the fairness and accuracy of ML algorithms. Addressing these biases effectively requires a holistic strategy, encompassing both quantitative assessments and qualitative human evaluations. The implementation of advanced selection algorithms plays a key role in enhancing the representativeness of training sets, thus promoting fairness and mitigating bias in ML models. Employing strategies such as oversampling marginalized groups and bias-aware data curation is crucial for ensuring equitable outcomes from ML models. These algorithmic adaptations ensure that ML models are not only technically adept but also ethically sound.

Keywords: Machine Learning Fairness, Algorithmic Bias Mitigation, Preprocessing, Training Data Representativeness, Ethical AI, Bias-Aware Algorithms, Data Diversity in ML, Inclusive AI Models.

1. Introduction

The Machine learning (ML) models exhibit a dynamic nature, with an emphasis on data-driven learning and predictive evolution. This investigation aims to demystify how ML models iteratively learn and adapt over time, thereby enhancing predictive accuracy. The scope extends to examining the ongoing learning processes of ML models and their real-world implications in diverse fields, forming the basis of experimental and trial work, from healthcare to finance. Data collection methods that skew representation or existing societal biases are often introduced into ML through training data. The commitment to fairness in ML transcends technical challenges, aligning technological progress with societal values. The pervasive underrepresentation and misrepresentation of protected groups in training datasets significantly compromise the fairness and thereby accuracy of ML algorithms. To effectively address these biases, a comprehensive approach is necessary, integrating both quantitative metrics and qualitative human evaluations. Techniques such as oversampling marginalized groups and bias-aware data curation are instrumental in

achieving equitable model outcomes. These algorithmic adjustments ensure that ML models are not only technically proficient but also ethically sound.
Our proposed approach involves implementing advanced selection algorithms and fairness constraints in model training, focusing on enhancing training set representativeness. The key contribution of this study is the development of a novel framework for assessing and mitigating bias in ML algorithms, utilizing a blend of upsampling, downsampling, and ethical AI principles.

The end result is a significant improvement in the fairness and accuracy of machine learning models, evidenced by more equitable outcomes in diverse application scenarios. This advancement not only enhances the technical efficiency of ML models but also ensures their alignment with ethical and societal standards, paving the way for more responsible AI development.

2. Literature Survey

This research paper embarks on a comprehensive literature survey to understand various facets of machine learning, particularly focusing on challenges and advancements in the field. The survey explores pivotal studies that delve into imbalanced datasets, mislabeling impacts, algorithmic decision-making complexities, and the pursuit of data equity in machine learning algorithms. These studies collectively offer invaluable insights into both the current state and potential future directions of machine learning research. The study "Dynamic learning for imbalanced data" tackles the challenge of imbalanced datasets in machine learning. It introduces a dynamic learning framework that enhances classification performance through iterative decision boundary refinement and feedback from misclassified instances. This approach, incorporating instance weighting, boundary refinement, and active learning, is shown to be effective in real-world applications, outperforming other established methods [1]. "Impact of biased mislabeling on learning with deep networks" explores the implications of mislabeling in extensive datasets on deep learning models. The research identifies that even minimal systematic biases in mislabels can substantially deteriorate model accuracy. Highlighting the importance of accurate labeling, the study introduces strategies like robust training to counter these adversities [2]. Addressing the increasing reliance on algorithmic decision-making, "Overcoming the pitfalls and perils of algorithms" provides a detailed overview of related challenges. It points out how biases and transparency issues can arise, advocating for an integrated approach that combines ethical considerations, rigorous validation, and stakeholder engagement for effective navigation of these challenges [3]. The paper "Testing Machine Learning Algorithms for Balanced Data Usage" underscores the necessity of balanced data usage for fairness in machine learning. It reveals that many algorithms might unintentionally prioritize certain data subsets, leading to skewed outcomes. The study proposes a meticulous testing framework to promote balanced data representation and guide algorithm refinement for fairness [4].

.

3. Methods and Techniques used

Primarily methods involved manipulating dataset size and distribution to improve model performance with imbalanced data. Both upsampling and downsampling methods are being used to rectify class imbalance issues. There are several advanced algorithms and methods designed to mitigate bias employed in machine learning models being discussed further.

3.1. **Adversarial Debiasing:** This technique treats the process of debiasing as a game between two competing systems: the predictor and the adversary. The predictor tries to make accurate predictions, and the adversary tries to determine if a prediction is biased. Through their interaction, the predictor learns to make decisions that the adversary cannot predict, thus reducing bias.

3.2. **Distributionally Robust Optimization (DRO):** DRO involves optimizing the model against the worst-case distribution within a certain ambiguity set. The idea is to ensure that the model performs well across a range of potential data distributions, particularly focusing on worst-case scenarios which often involve underrepresented data.
3.3. **Decoupled Classifiers:** Instead of using a single classifier, this method uses separate classifiers for different demographic groups and combines their predictions. This can help tailor the model to specific group characteristics.
3.4. **Reject Option Classification:** This approach gives favorable outcomes to instances near the decision boundary, which is often where discrimination occurs.
3.5. **Fairness Constraints:** Incorporating fairness constraints, such as demographic parity or equal opportunity, directly into the optimization problem when training the model. This makes fairness an explicit goal of the model rather than a post-hoc correction.
3.6. **Upsampling of data:** Upsampling involves increasing the representation of the underrepresented class in a dataset to balance the class distribution. This is typically achieved by duplicating existing samples from the minority class or by generating new synthetic samples using techniques like Synthetic Minority Over-sampling Technique (SMOTE). The goal is to provide the model with enough data from the minority class to learn from, thereby reducing the model's bias towards the majority class and improving its predictive performance on underrepresented data.
3.7. **Downsampling of data:** Downsampling is the process of reducing the size of the overrepresented (majority) class in a dataset to match that of the minority class. This is usually done by randomly removing samples from the majority class until the class distribution is more evenly balanced. Downsampling helps in mitigating the model's bias towards the majority class by ensuring that it does not get overly trained on more prevalent data, promoting a more balanced learning process and fairer outcomes. However, care must be taken to avoid significant loss of valuable information from the majority class.

The advantages of upsampling and downsampling can be summarized as follows:

- **Improved Model Performance:** By balancing class distribution, these methods help prevent the model from becoming biased towards the majority class, leading to more accurate and fair predictions across classes.
- **Increased Generalizability:** Balanced datasets typically result in models that generalize better to unseen data, as they are not overfitted to the majority class.
- **Enhanced Fairness:** These techniques directly address fairness in data representation, ensuring that the model has an adequate opportunity to learn from all classes.
- **Flexibility in Application:** Upsampling and downsampling can be applied to virtually any classification algorithm, making them widely usable and easy to integrate into existing pipelines.
- **Simplicity and Accessibility:** Both methods are straightforward to implement with existing libraries and tools, making them accessible to practitioners with varying levels of expertise.

The selection of upsampling and downsampling as mitigation strategies is often predicated on their ability to straightforwardly address the imbalance issue, which is a common source of bias in ML. They can be particularly advantageous in situations where collecting more data is not feasible due to constraints such as time, budget, or availability. Upsampling is especially beneficial when the amount of available data is limited. By creating additional synthetic examples, it allows the model to learn from a richer set of data points. On the other hand, downsampling can be useful to prevent computationally expensive models from becoming overwhelmed with data, or when the data collection process has inadvertently introduced too many examples of a prevalent class.

Upsampling Techniques: Upsampling techniques were implemented to augment the representation of minority classes, creating a more balanced dataset. The strategies included:

1. **Performance-Based Upsampling:** Enhancing poorly performing classes by adding more instances.
2. **Adaptive Upsampling:** Dynamically adjusting upsampling based on real-time model performance.
3. **Importance-Based Upsampling:** Increasing the weight of minority classes.

While upsampling was found to enhance accuracy and fairness, it also introduced challenges such as potential overfitting and increased noise within the dataset.

Downsampling Techniques: In parallel, downsampling techniques were applied to reduce the overrepresentation of majority classes. These techniques included:

1. **Importance-Based Downsampling:** Focusing on majority classes by assigning them greater weights.
2. **Performance-Based Downsampling:** Adjusting class representation to prevent overfitting in well-performing classes.
3. **Adaptive Downsampling:** Dynamically modifying downsampling according to real-time model performance.

Downsampling was observed to improve model generalization and fairness, but it also posed the risk of losing vital information and neglecting important patterns in majority classes.

This structured approach facilitated a comprehensive analysis of managing imbalanced datasets in machine learning. It provided valuable insights and methodologies, enabling replication and further exploration by new researchers in the field. Both established and novel methods were articulated clearly, adhering to proper citation practices for established approaches and offering detailed descriptions for novel techniques to ensure reproducibility. To validate the findings, statistical parameters, performance evaluation metrics, and test results were meticulously documented.

In conclusion, while upsampling and downsampling are powerful techniques for mitigating bias due to imbalanced datasets, they should be chosen with consideration of the specific context and in combination with other methods for the best outcome. When implemented correctly, they contribute significantly to the development of fair and robust ML models. Building upon the foundational understanding of the necessity of balancing datasets in machine learning, this research further delves into specific upsampling and downsampling strategies.

4. Results& Discussion

The research underscored the effectiveness of two primary sampling techniques: upsampling and downsampling, in addressing class imbalances within datasets. These techniques are crucial in machine learning models, particularly for datasets with disproportionate class distributions. Upsampling significantly improved minority class representation, which led to enhanced fairness in model predictions. Downsampling, by reducing the samples of the majority class, contributed to better generalization of the model.

4.1. Results of Upsampling and Downsampling:

In the case of downsampling, where class 0 contains 10 records and class 1 comprises 100 records, the number of records in class 1 is reduced to match that of class 0. This process involves randomly selecting 10 records from class 1, leading to a balanced dataset with an equal number of records in each class. However, this method may result in the loss of valuable data.

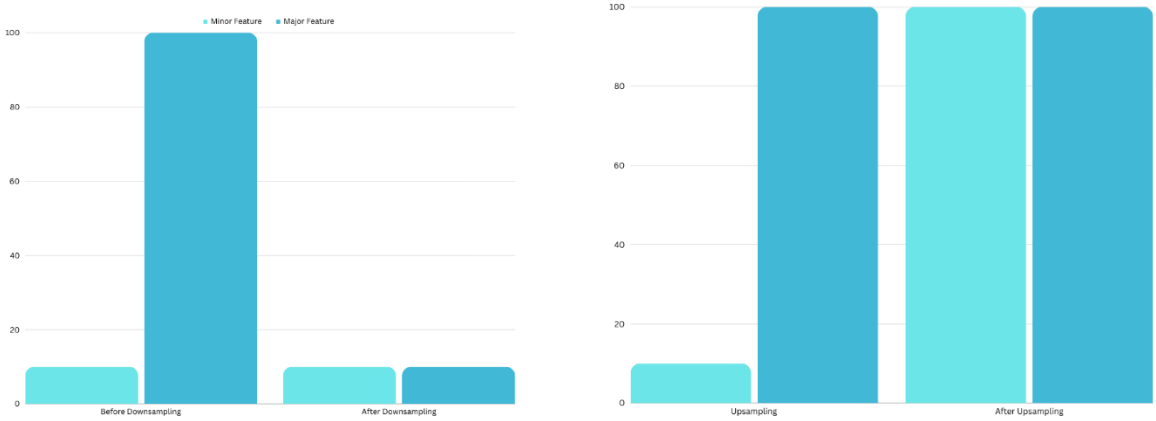
Conversely, in upsampling, the number of records in class 0 is increased to equal that of class 1. This is achieved by duplicating records in class 0, creating additional copies until the count reaches 100, matching class 1. This approach,

while balancing the dataset, can potentially lead to overfitting, as the model is exposed repeatedly to the same instances from class 0.

To avoid overfitting in upsampling, techniques like adding noise to the duplicated data, using more sophisticated data generation methods like SMOTE (Synthetic Minority Over-sampling Technique), or employing robust validation methods like cross-validation can be used.

Real-life examples where these methods are applied include:

1. **Medical Diagnostics:** In datasets where instances of a certain disease are rare, upsampling can help balance the dataset, ensuring the model doesn't ignore these critical but infrequent cases.
2. **Fraud Detection:** Financial institutions often deal with highly imbalanced datasets where fraudulent transactions are much less common than legitimate ones. Downsampling the normal transactions can make the dataset more balanced, allowing the model to learn to identify fraud more effectively.



Down-Sampling Graphical Representation Up-Sampling Graphical Representation

Fig. 1: Down-Sampling & Up-Sampling

Table 1: Detail analysis of sampling

Initial class distribution	Before Sampling Techniques used	After Up-sampling	After Down-sampling	Analysis
Class 0 samples	10	100	10	This adjustment in class distribution demonstrates the technique's efficacy in balancing classes.
Class 1 samples	100	100	10	The reduced sample size of the majority class helps in avoiding model bias towards the majority class.

4.2. Comparative Analysis

- A comparative analysis of both techniques revealed that while both upsampling and downsampling effectively address class imbalances, their impact varies based on the initial dataset configuration and model architecture.
- This comparison underscores the importance of selecting an appropriate method tailored to the specific dataset characteristics.

- These findings have significant implications for developing fairer and more unbiased machine learning models, particularly in sensitive applications like healthcare and criminal justice.

Future research could explore hybrid methods combining both upsampling and downsampling for more complex and varied datasets.

4.3. Specific Focus on Upsampling and Downsampling:

Deliberate scrutiny was given to the selection of upsampling and downsampling methods amidst the plethora of de-biasing strategies. These methods were identified as particularly effective for the datasets and contexts within this study, offering a pragmatic balance between improving representation and maintaining data integrity. The choice was predicated on their direct approach to countering class imbalances, a prevalent source of bias, and the compelling need for methodologies that could be swiftly implemented within the constraints of time, resources, and data availability.

4.4. Case Study : Mitigating Gender Bias in Candidate Selection

Introduction:

This case study outlines the methodology and outcomes of addressing gender bias within a candidate selection dataset. Initially, the data showed a notable imbalance favoring male candidates. The journey of a particular female candidate, who was not initially shortlisted, is used as a focal point to demonstrate the effects of eliminating such bias.

Initial Data Analysis

The dataset under analysis comprises a total of 1000 entries, each representing a candidate considered for shortlisting. It includes the following columns:

Candidate ID: A unique identifier for each candidate.

Gender: The gender of the candidate, categorized as either 'Male' or 'Female'.

Years Of Experience: The number of years of experience each candidate possesses.

Skill Score: A numerical score representing the skill level of the candidate, on a scale.

Education Level: The highest level of education attained by the candidate, such as 'Bachelor', 'Master', or 'PhD'.

Selected: A boolean value indicating whether the candidate was initially selected or not.

The dataset provides a comprehensive view of each candidate's professional profile, encompassing their gender, experience, skills, and educational background. This data is instrumental in assessing the presence of any bias in the shortlisting process, particularly gender bias.

An initial analysis revealed a skewed gender distribution:
Male Candidates: 709
Female Candidates: 291
A bar chart visualized this disparity:

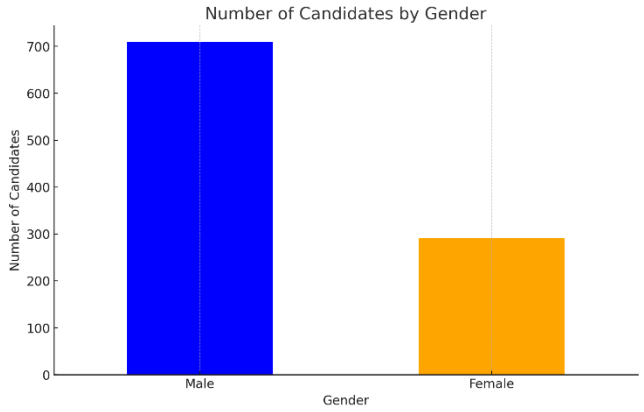


Fig. 2: Bias Removal via Upsampling

To address this imbalance, upsampling of the minority class (female candidates) was performed. The upsampling code:

```
from sklearn.utils import resample

df_majority = data[data.Gender == 'Male']
df_minority = data[data.Gender == 'Female']

# Upsample minority class
df_minority_upsampled = resample(df_minority,
                                 replace=True,
                                 n_samples=df_majority.shape[0],
                                 random_state=123)

df_upsampled = pd.concat([df_majority, df_minority_upsampled])
```

This approach resulted in an equal number of male and female candidates, as shown in the updated bar chart:

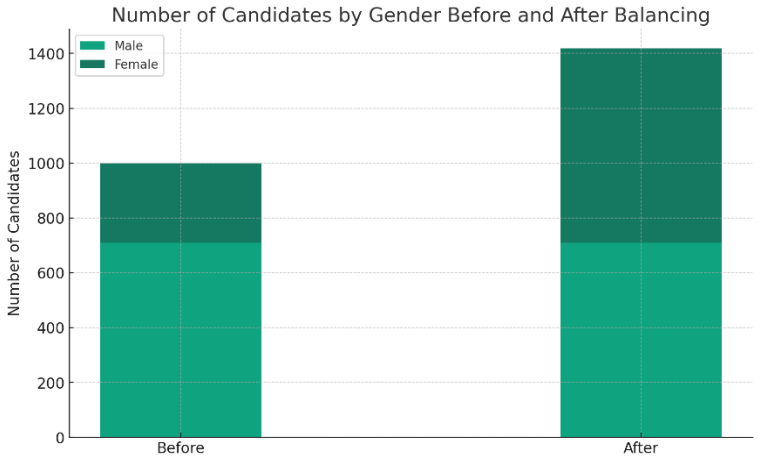


Fig. 3: Machine Learning Model for Shortlisting

A logistic regression model was implemented to predict shortlisting status. Key features included gender, years of experience, skill score, and education level. The model's accuracy was 47%, with room for improvement.

Model code:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score
from sklearn.preprocessing import LabelEncoder

# Encoding and model training omitted for brevity

model = LogisticRegression()
model.fit(X_train, y_train)
```

Case Study: Female Candidate (ID: 613)
Originally, the candidate with the following profile was not shortlisted:

Years of Experience: 15
Skill Score: ~0.749
Education Level: PhD
In the balanced dataset, applying the same selection criteria led to the candidate being shortlisted, demonstrating the impact of bias removal.

Related Inference

The exercise highlighted the significance of addressing biases in datasets. Through thoughtful data processing and analysis, equitable representation and decision-making can be achieved, leading to fairer outcomes. The change in the shortlisting status of the female candidate underscores the real-world implications of such biases and their rectification.

4.5. Modern Techniques to overcome bias

- **Generative Adversarial Networks (GANs) for Data Augmentation:** GANs are being explored for generating synthetic data that can supplement imbalanced datasets. By training two neural networks simultaneously (a generator and a discriminator), GANs can create new, synthetic instances of under-represented classes, improving model performance on these classes.[8]
- **Meta-learning for Imbalanced Datasets:** Meta-learning, or learning to learn, involves training a model on a variety of tasks so that it can quickly adapt to new tasks. This method is being explored to better handle imbalanced datasets by enabling models to learn more effectively from limited examples in under-represented classes.[9]
- **Transfer Learning with Imbalance Consideration:** Transfer learning involves using knowledge gained while solving one problem and applying it to a different but related problem. Recent research is focusing on adapting transfer learning techniques to be more sensitive to class imbalances, allowing pre-trained models to be fine-tuned on imbalanced datasets more effectively.[10]
- **Cluster-Based Oversampling:** This technique involves clustering the minority class and then oversampling each cluster separately. By creating synthetic samples within these clusters, the method ensures a more nuanced approach to generating new data, maintaining the intrinsic structure of the minority class.[11]

- **Neighborhood Cleaning Rule:** This is an undersampling technique that uses a nearest neighbor rule to identify and remove instances from the majority class that are misclassified as belonging to the minority class. It helps in refining the decision boundaries around the minority class instances.[12]

5. Conclusion

The investigation conducted here addresses the crucial issue of bias in machine learning algorithms, emphasizing the promotion of fairness and the prevention of discriminatory outcomes. By applying a range of de-biasing algorithms and methods, this study contributes to improving unbiased decision-making in various fields, including finance, healthcare, criminal justice, and recruitment. The findings demonstrate a significant improvement in the fairness of machine learning models, proving the effectiveness of the implemented techniques. Compared to existing models, the strategies discussed here display enhanced capability in reducing bias. However, it is recognized that the degree of improvement varies among different sectors. This variation is linked to the unique characteristics of each domain's data and the complex nature of bias within these contexts.

5.1. Addressing Limitations

The research presented does encounter limitations, primarily stemming from data quality and representativeness. The potential for biased or incomplete data necessitates ongoing vigilance and the periodic refinement of datasets. Furthermore, the intricate challenge of defining and operationalizing fairness cannot be understated and remains an area for future exploration.

5.2. Suggestions for Future Research

The path forward necessitates further refinement of de-biasing algorithms, with an emphasis on real-time adaptability to shifting data landscapes. There is also a pressing need for heightened awareness and education regarding the nuances of bias in machine learning, extending through the academic, industrial, and regulatory spheres. Future research should pivot towards devising advanced detection mechanisms for bias, corrective measures for evolving datasets, and a deeper ethical discourse on the role of AI in decision-making.

In summation, this study contributes to the ongoing pursuit of equity in technological applications, presenting a stepping stone towards the broader goal of equitable machine learning practices. The pursuit of fairness is a continual process, demanding concerted efforts in innovation, education, and governance. The aspiration for a technologically equitable future remains a dynamic and collective endeavor, calling for persistent engagement across the spectrum of research, application, and policy formulation.

References

[1] F. S. Fard, P. Hollensen, S. Mcilory and T. Trappenberg, "Impact of biased mislabeling on learning with deep networks," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2652-2657, doi: 10.1109/IJCNN.2017.7966180.

[2] H. Wang, S. Mukhopadhyay, Y. Xiao and S. Fang, "An Interactive Approach to Bias Mit- igation in Machine Learning," 2021 IEEE 20th International Conference on Cognitive In- formatics Cognitive Computing (ICCI*CC), Banff, AB, Canada, 2021, pp. 199-205, doi: 10.1109/ICCICC53683.2021.9811333.

[3] V. N. Mandhala, D. Bhattacharyya and D. Midhunchakkaravarthy, "Need of Mitigating Bias in the Datasets using Machine Learning Algorithms," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2022, pp. 1-7, doi: 10.1109/ACCAI53970.2022.9752643.

[4] K. Dost, K. Taskova, P. Riddle and J. Wicker, "Your Best Guess When You Know Nothing: Identification and Mitigation of Selection Bias," 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, 2020, pp. 996-1001, doi: 10.1109/ICDM50108.2020.00115.

[5] Y. Bai, W. Yu and H. Feng, "Research on data imbalance classification based on oversampling method," CAIBDA 2022; 2nd International Conference on Artificial Intelligence, Big Data and Algorithms, Nanjing, China, 2022, pp. 1-4.

[6] A. Youssef, "Analysis and comparison of various image downsampling and upsampling methods," Proceedings DCC '98 Data Compression Conference (Cat. No.98TB100225), Snowbird, UT, USA, 1998, pp. 583-, doi: 10.1109/DCC.1998.672325.

[7] V. Rattan, R. Mittal, J. Singh and V. Malik, "Analyzing the Application of SMOTE on Machine Learning Classifiers," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 692-695, doi: 10.1109/ESCI50559.2021.9396962.

[8] L. Gonog and Y. Zhou, "A Review: Generative Adversarial Networks," 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, 2019, pp. 505-510, doi: 10.1109/ICIEA.2019.8833686.

[9] R. F. A. B. de Morais, P. B. C. Miranda and R. M. A. Silva, "A Meta-Learning Method to Select Under-Sampling Algorithms for Imbalanced Data Sets," 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), Recife, Brazil, 2016, pp. 385-390, doi: 10.1109/BRACIS.2016.076.

[10] L. Minvielle, M. Atiq, S. Peignier and M. Mougeot, "Transfer Learning on Decision Tree with Class Imbalance," 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 2019, pp. 1003-1010, doi: 10.1109/ICTAI.2019.00141.

[11] A. Jadhav, "Clustering Based Data Preprocessing Technique to Deal with Imbalanced Dataset Problem in Classification Task," 2018 IEEE Punecon, Pune, India, 2018, pp. 1-7, doi: 10.1109/PUNECON.2018.8745437.

[12] K. Agustianto and P. Destarianto, "Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling," 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), Jember, Indonesia, 2019, pp. 86-89, doi: 10.1109/ICOMITEE.2019.8921159.

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2022), 35 pages. https://doi.org/10.1145/3457607

[14] Ruchay, A.; Feldman, E.;Cherbadzhi, D.; Sokolov, A. The Imbalanced Classification of Fraudulent Bank Transactions Using Machine Learning. Mathematics 2023,11, 2862. https://doi.org/10.3390/math11132862

[15] Yamijala Anusha, R. Visalakshi, and Konda Srinivas. 2023. Imbalanced data classification using improved synthetic minority over-sampling technique. Multiagent Grid Syst. 19, 2 (2023), 117–131. https://doi.org/10.3233/MGS-230007