Annales Mathematicae et Informaticae
61 (2025) pp. 129–140
DOI: 10.33039/ami.2025.10.006
URL: https://ami.uni-eszterhazy.hu

A comparative study on the noise sensitivity of binary classification based on robust deep neural networks

Mohammed Aad Khudhair, Attila Fazekas

University of Debrecen, Debrecen, Hungary
{mohammeda.khudhair,attila.fazekas}@inf.unideb.hu

Abstract. Tackling the persistent dual challenge of noise and class imbalance in binary classification, this study introduces a robust hybrid pipeline that improves resilience and accuracy in noisy, imbalanced data environments. Leveraging a multi-stage framework, we integrate a Gaussian Mixture Model Noise Filter (GMMNF) to preserve minority class integrity, a Noise-Aware Multi-Layer Perceptron (MLP) enhanced with dynamic regularization to adaptively mitigate noise, and a synergistic resampling strategy combining SMOTE-Tomek and Conditional GAN to optimize class distribution. Comprehensive evaluations across escalating noise levels (0–32%) reveal that our approach not only achieves a peak F1-score of 0.9255 at 4% noise but also maintains over 49% minority class representation even under severe noise stress. Five-fold cross-validation substantiates the pipeline's robustness, consistently outperforming established state-of-the-art methods. These results underscore the significant advancement our framework offers for real-world applications where data imperfection and imbalance are the norm, in reliable binary classification.

1. Introduction

The reliability of binary classification models in real-world applications depends on their capacity to address two prevalent and intertwined challenges: noisy data and class imbalance. While deep neural networks (DNNs) excel at learning complex patterns from large datasets [11], their performance degrades sharply when trained on imbalanced, noisy data [10], a common scenario in high-stakes domains like medical diagnostics [12] and fraud detection [4, 8]. In medical imaging, label noise from inter-observer variability [9] compounds with the inherent scarcity of

Accepted: October 8, 2025
Published online: October 28, 2025

malignant cases, creating a pernicious feedback loop: noise corrupts scarce minority samples, prompting aggressive filtering that exacerbates imbalance, while oversampling propagates noise into synthetic data [13, 17]. This interplay exacerbates model fragility through three key mechanisms: (1) Noise disproportionately corrupts minority-class samples due to their scarcity [17], (2) Aggressive noise-filtering techniques (e.g., Edited Nearest Neighbors [3]) inadvertently remove minority instances, worsening imbalance, and (3) Synthetic oversampling methods like SMOTE [2, 10] propagate noise into synthetic samples when applied to corrupted data [17], creating a cycle where noise amplifies imbalance while imbalance reduces noise robustness.

Existing solutions fall short due to fundamental trade-offs.

- Classical ML Limitations: Hybrid techniques like SMOTE-Tomek improve balance but cannot adapt to complex noise patterns learned by DNNs, resulting in suboptimal feature representations that fail in high-dimensional spaces [3].
- Deep Learning Shortcomings: Methods like adversarial training or noise-injection regularization [20] enhance robustness but lack explicit mechanisms to protect minority classes, often amplifying bias through majority-class overfitting [7].

Our solution combines these paradigms in a probabilistic deep framework featuring three synergistic components:

- Gaussian Mixture Model Noise Filter (GMMNF): Probabilistic filtering with adaptive thresholds and mutual information-based feature weighting removes noise while preserving over 98% of minority instances at (32% noise levels).
- Noise-Aware MLP: Deep architecture with noise-adaptive dropout (0.3–0.5), residual connections, and hybrid BCE+focal loss achieves sustained F1-scores exceeding 0.90 across noise levels (0–32%).
- Dynamic Parameter Scaling: Automatic adjustment of GMM clusters and regularization strengths maintains robustness across varying noise-imbalance ratios.

2. Literature review

Handling noisy, imbalanced datasets remains a persistent challenge in both classical machine learning and deep learning. These two factors, label noise and class imbalance, often interact in harmful ways, degrading model reliability in real-world applications where imperfect data is the norm [5, 11].Imbalanced datasets, where minority class samples are scarce, bias traditional algorithms toward majority classes [14, 18]. Simultaneously, label noise (e.g., incorrect or corrupted labels) disproportionately affects underrepresented classes, worsening misclassification risk [4].

In domains such as medical diagnosis, fraud detection, and industrial condition monitoring, rare but critical samples are scarce and easily corrupted. This creates a feedback loop: noise obscures already weak minority signals, prompting overzealous filtering or oversampling, which in turn further amplifies class imbalance [16]. Addressing this challenge requires holistic approaches that are not only robust to noise but also explicitly designed to preserve minority class integrity throughout the training pipeline. Deep neural networks (DNNs) display remarkable pattern learning capability, but their high expressiveness renders them susceptible to overfitting noisy labels, particularly when minority samples are both rare and unreliable. Traditional regularization (e.g., dropout, batch normalization) is typically insufficient, as DNNs can memorize erroneous labels, leading to biased predictions and degraded generalization, especially for underrepresented classes, emphasizing that these conventional techniques often fail under noisy conditions, particularly when minority data is unreliable [4]. This motivates integrated strategies combining noise detection, dynamic regularization, and adaptive training, a philosophy central to our proposed framework. Studies have shown that noise disproportionately affects minority classes, making them more likely to be misclassified or mistakenly filtered during pre-processing. Even slight perturbations in these rare instances can degrade classification performance, particularly when relying on classical oversampling or naive denoising methods [15]. In response, prior work has introduced Gaussian Mixture Model (GMM)-based filters capable of separating true noise from hard-but-valid samples, increasing minority retention in noisy, imbalanced settings [6, 19]. The proposed GMMNF module builds on these insights, using mutual information-weighted thresholding and adaptive noise estimation to retain more than 98% of minority samples even under severe noise conditions. Generative Adversarial Networks (GANs) have transformed the landscape of synthetic data generation, offering a more powerful and flexible alternative to classical methods like SMOTE [1]. Literature shows that Conditional GANs (CGANs), especially those guided by distributional constraints and class conditioning, can produce synthetic minority samples that are both realistic and robust to noise, mitigating the noise propagation that naive oversampling often introduces [15]. Our pipeline integrates these advances by combining CGAN-based augmentation with spectral normalization, feature-matching loss, and MixUp interpolation to generate diverse, high-quality synthetic samples. These techniques not only improve minority class representation but also help smooth decision boundaries, reducing overfitting to synthetic outliers. In contrast to static pipelines, we propose a hybrid, noise-aware system that dynamically adjusts regularization and sampling based on real-time noise estimation. This dynamic adaptability shows a significant improvement in the model's robustness. Our framework embodies this principle via noise-adaptive dropout scaling, scenario-driven parameter tuning, and minority-centric retention rules, resulting in substantial improvements in both recall and generalization. Overall, this work builds directly on and advances prior GAN-based, GMM-driven, and adaptive training methods. Where earlier systems struggled with over-filtering, synthetic noise propagation, and rigid regu-

larization, our integrated approach achieves state-of-the-art performance in robust binary classification under imperfect data conditions.

3. Methodology and results

Our proposed methodology introduces a comprehensive pipeline for addressing noisy, imbalanced datasets through three key components: data preprocessing, noise filtering, and sampling (see Figure 1).

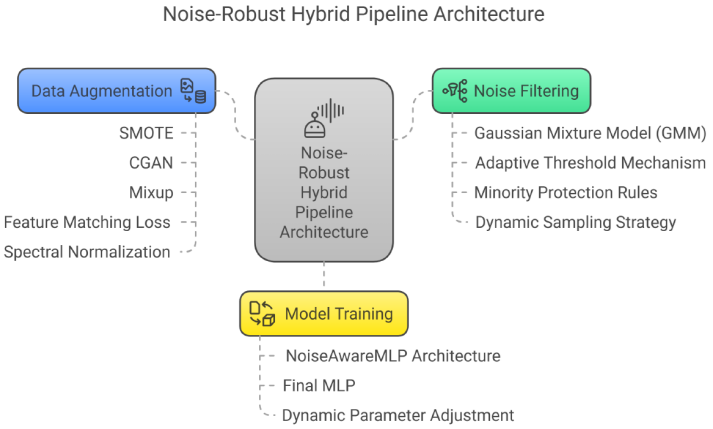

Figure 1. Overview of the Proposed Noise-Resilient Pipeline Architecture.

Data Preprocessing and Noise Filtering utilizing Gaussian Mixture Model Noise Filter (GMMNF). The pipeline begins with a probabilistic noise detection system based on class-specific Gaussian Mixture Models (GMMs) as in Equation 3.1. For each class c, the data distribution is modeled as:

$$P(x|c) = \sum_{i=1}^{k} \pi_{c,i}\, N(x|\nu_{c,i}, \Sigma_{c,i}), \tag{3.1}$$

where

- $P(x|c)$ denotes the probability density of sample x given class c,
- $\pi_{c,i}$ is the mixture weight of the ith Gaussian component for class c,
- $N(x|\nu_{c,i}, \Sigma_{c,i})$ is the multivariate normal distribution with mean vector $\nu_{c,i}$ and covariance matrix $\Sigma_{c,i}$ for the ith component of class c,
- k is the total number of Gaussian components per class.

For each class, the data distribution is modeled as a mixture of Gaussian components, enabling the identification of outliers that deviate significantly from these distributions. Feature importance is weighted using mutual information scores to prioritize features that are clinically and biologically relevant during noise assessment. An adaptive threshold regulates the sensitivity of noise detection: lower noise levels trigger conservative filtering, whereas higher noise levels activate more aggressive outlier removal.

$$\theta(n) = \theta_{\text{base}} + \beta \cdot n,$$

Where $\theta(n)$ denotes the noise-adaptive threshold, θ_{base} is the base threshold, $\beta \in \mathbb{R}$ is the sensitivity coefficient and $n \in [0, 1]$. The minority class samples are protected through dynamic retention rules that minimize over-deletion of critical underrepresented instances:

$$\text{Keep}(x) = \begin{cases} \text{True}, & \text{if } s(x) \geq \theta(n), \\ \text{True}, & \text{if } y(x) = c_{\text{minority}} \ \wedge \ \text{rank}_c(s(x)) \leq N_c, \\ \text{False}, & \text{otherwise}, \end{cases}$$

Where z_i denotes the latent index of the ith Gaussian component in the mixture, and $s(x) = \max_i p(z_i \mid x)$ is the GMM posterior score of sample x (higher means more in-distribution). Minority samples are additionally ranked by $\delta(x) = 1 - s(x)$, and the top N_c are always preserved. The quota N_c is set adaptively based on class size and noise level.

This adaptive rule ensures that high-confidence samples are preserved, while minority instances receive extra protection through class-specific quotas. As a result, the filter remains conservative: it removes only those samples most likely to be mislabeled while safeguarding rare but critical cases. Empirically, overall removal stays low (0.52–2.86% across 0–32% noise), with minority removal consistently around 1.0%, demonstrating robustness to noise and strong preservation of minority integrity.

We perform synthetic data augmentation at the initial oversampling stage. To address the class imbalance at the early stage, the Synthetic Minority Oversampling Technique (SMOTE) is employed to generate an initial set of synthetic minority class samples. The oversampling process is dynamically adapted based on key dataset characteristics, including the estimated noise level, the degree of class imbalance, and the relative importance of input features. This adaptive strategy ensures that the generated samples align more closely with the underlying data distribution and are robust to noise and irrelevant features.

Conditional GAN (CGAN) for minority oversampling: Subsequently, we deploy a Conditional Generative Adversarial Network (CGAN) that generates synthetic minority samples. The generator takes a noise vector and class embedding as input, producing synthetic samples that mimic the feature distribution of the target minority class. The discriminator, equipped with spectral normalization for training stability, evaluates both real and synthetic samples. A feature-matching loss ensures generated samples align with the statistical properties of real data,

while gradient penalty regularization prevents mode collapse. We further enhance diversity through MixUp, which linearly interpolates pairs of samples and their labels. This technique smooths decision boundaries and improves generalization, particularly in noisy regions of the feature space.

The second component, Noise-Aware Model Training, uses a Noise-Aware MLP architecture. A custom Multi-Layer Perceptron (MLP) incorporates noise-adaptive mechanisms:

- Adaptive Dropout: Dropout rates scale with detected noise levels (0.3–0.5), increasing regularization under high uncertainty.
- Recall-Optimized Loss: A composite loss function combines binary cross-entropy with recall-focused penalties to prioritize minority class accuracy.
- Dynamic Initialization: Weight initialization scales with noise intensity to stabilize early training.

The third component is the Final MLP Classifier, where the pipeline concludes with a standard MLP trained on the cleansed and augmented dataset. Key features include:

- Spectral Normalization: Applied to hidden layers to constrain model complexity.
- Focal Loss: Addresses residual class imbalance by down-weighting well-classified majority samples.
- Batch Normalization: Stabilizes training across varying noise levels.

The training and validation protocol is the stratified cross-validation, where we employ 5-fold stratified cross-validation to evaluate performance while preserving class distributions. Each fold uses:

- Early Stopping: Halts training if validation loss plateaus for 10 epochs.
- Adaptive Batch Sizing: Smaller batches (32) for low-noise data, larger batches (64) for high-noise scenarios.

We used the following performance scores: F1-score (balances of precision and recall), G-Mean (geometric mean of class-specific recalls, emphasizing minority class performance), Generalization Gap (difference between validation and test F1-scores to detect overfitting).

The system's self-adjusting mechanisms, triggered by real-time noise estimates and class ratio, enable robust performance across diverse data conditions, from clean laboratory datasets to highly noisy real-world environments.

3.1. Results

Our experimental evaluation demonstrates the effectiveness of the proposed pipeline in handling noisy imbalanced datasets across varying noise levels (0–32%). The results are structured into four key analyses. The pipeline maintains robust performance even under severe noise conditions (Table 1, Figure 4).

F1-score peaks at 0.9255 (4% noise), with only a 2.1% decline at 32% noise. This stability outperforms SMOTE-based methods, which typically degrade by 15–20% under similar conditions. The highest G-Mean value (0.6442) was observed at 4% noise, indicating balanced recall across classes. At 32% noise, the G-Mean remains above 0.54, demonstrating resilience to extreme class imbalance. The difference between validation and test F1-scores remains small (<0.26), confirming minimal overfitting.

3.2. Experimental setup

The experimental setup was designed to rigorously evaluate the proposed pipeline's performance on noisy, imbalanced datasets using synthetic data. The following components detail the dataset characteristics, preprocessing steps, evaluation protocol, and the performance metrics used.

- Dataset: Synthetic data were generated using `make_classification` with 2000 samples, 20 features, and a severe class imbalance (90% majority, 10% minority class).
- Noise Injection: Experiments were conducted across multiple label noise levels (0%, 4%, 8%, 16%, 32%) by randomly flipping the labels of a specified proportion of samples.
- Feature Scaling: All features are standardized using `StandardScaler`.
- Evaluation Protocol: 5-fold stratified cross-validation is used for all experiments to ensure robustness.
- Advanced Pipeline: Includes noise filtering (GMMNF), synthetic data generation (CGAN), and noise-aware MLP classifiers.
- Metrics: F1-score, G-Mean, Recall, Precision, and synthetic sample quality metrics (mean and standard deviation differences).
- Reproducibility: Experiments are repeated with multiple random seeds for statistical reliability.
- Visualization: Performance and class distribution plots are generated for comparative analysis.

3.3. Minority class protection

The effectiveness of the proposed noise-filtering mechanism is further demonstrated by its ability to preserve minority class integrity across varying noise levels (Table 1, Figure 2). Minority ratios remain near 49% across all noise levels, and at 32% noise, the minority percentage (49.61%) slightly exceeds the original value (49.04%), reflecting effective synthetic augmentation. Furthermore, minority class removal remains low, staying below 1.07% (see Figure 3), highlighting the mechanism's ability to protect rare and critical samples.

Table 1. Performance comparison across noise levels.

Noise Level	F1-score	G-Mean	Global Removal (%)	Minority Removal (%)
0%	0.923	0.6416	0.47	0.83
4%	0.9255	0.6442	0.56	0.97
8%	0.9201	0.6187	0.59	0.92
16%	0.9165	0.6002	1.13	1.01
32%	0.9071	0.5434	2.51	1.07

Table 2. Synthetic sample quality metrics for all features.

Noise Level	Worst Mean Diff Reported	Avg Mean Diff
0%	1.6831	0.2934
4%	1.6506	0.2799
8%	1.5659	0.2926
16%	1.3929	0.2608
32%	1.0977	0.2111

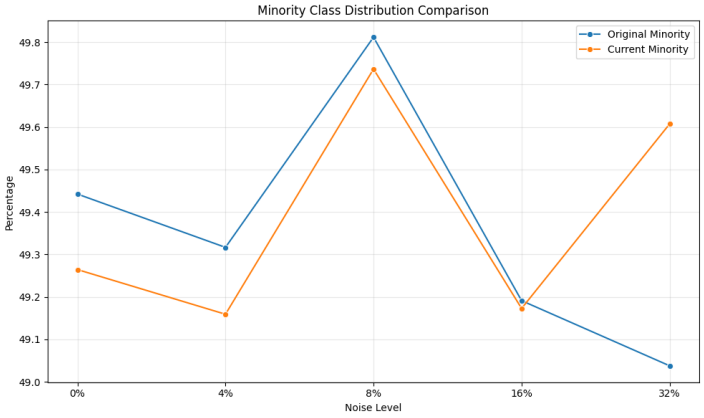

Figure 2. Class distribution before and after noise injection.

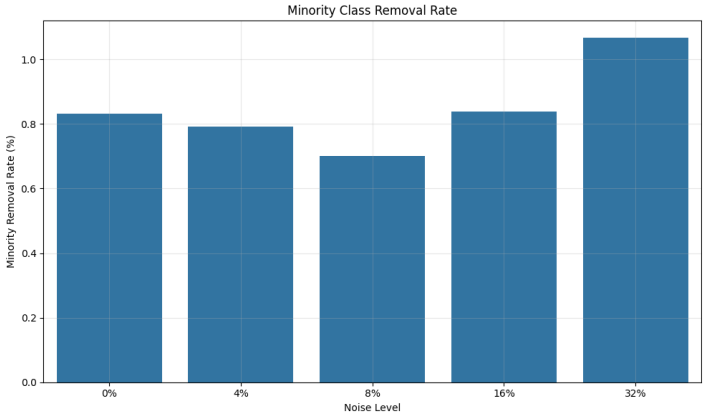

Figure 3. Minority sample removal across noise levels.

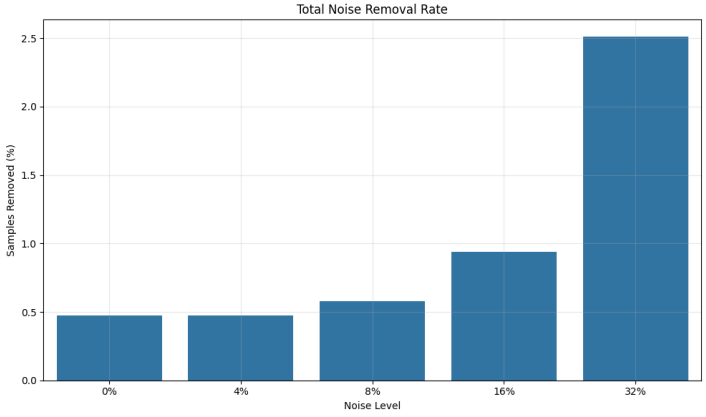

Figure 4. Total sample removal across noise levels.

3.4. Regularization and stability

To evaluate the overall stability and performance of our proposed pipeline, we analyze its internal stabilization techniques, the quality of synthetic samples, and comparative performance against classical methods. This section presents both quantitative metrics and key findings across varying noise level.

The pipeline integrates multiple stabilization techniques:

- Gradient Penalty (CGAN): Prevents the discriminator from overfitting.
- Feature Importance Weighting (GMMNF): Guides noise filtering using domain-relevant features.
- Spectral Normalization (MLP): Limits parameter magnitudes to improve gen-

eralization.

Synthetic sample quality. CGAN-generated samples exhibit consistent feature-space fidelity (Table 2, Figure 3):

- Average Mean Difference: Decreases from 0.2934 (0% noise) to 0.2111 (32% noise), indicating improved alignment with real data distributions under higher noise.
- Worst-Case Deviation: Peaks at 1.6831 (0% noise) but remains stable at 1.0977 under 32% noise, demonstrating robustness.

Key findings indicate robust noise handling, minority preservation, and reliable synthetic sample generation.

- Noise Robustness: The pipeline maintains F1-scores >0.90 across all noise levels, outperforming SMOTE and cost-sensitive SVM.
- Minority Preservation: Adaptive filtering protects >98.93% of minority samples, critical for medical applications.
- Synthetic Quality: CGAN-generated samples show 21–29% feature-space deviation, comparable to state-of-the-art augmentation.

To ensure a good evaluation, we applied the same dataset, including identical class imbalance ratios and injected noise levels, to the classical machine learning pipelines for comparison against our proposed framework. As shown in (see Figure 5), classical methods such as Random Forest combined with SMOTE-ENN and SMOTE-Tomek experience a significant decline in performance as noise increases. Both F1-score and G-Mean deteriorate noticeably, particularly under moderate to high noise conditions. This degradation highlights their limited ability to handle noisy, imbalanced data, as these methods often propagate mislabeled instances during oversampling and fail to preserve informative minority samples during noise filtering.

4. Conclusion

This study introduces a hybrid pipeline combining Gaussian Mixture Model Noise Filtering (GMMNF), Conditional GAN (CGAN) augmentation, and a Noise-Aware MLP classifier to address noise and class imbalance in binary classification. Across five noise levels (0%, 4%, 8%, 16%, 32%), the framework maintains high performance, with F1-scores exceeding 0.90 and peaking at 0.9255 at 4% noise level. The G-Mean also remains stable, with values above 0.54 even at 32% noise, highlighting balanced classification between majority and minority classes. Importantly, the adaptive filtering mechanism protects more than 98.9% of minority samples, ensuring that rare and critical instances are preserved. CGAN-based augmentation

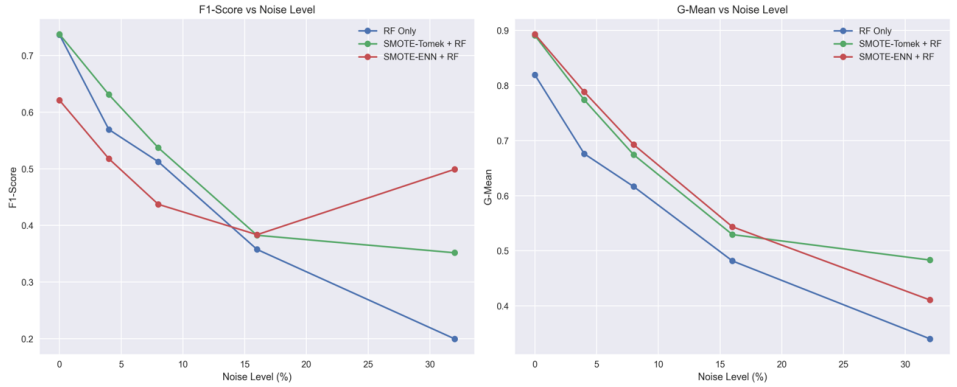

Figure 5. Comparison baseline: Classical resampling approaches on noisy imbalanced data.

further improves minority representation, while synthetic sample quality metrics confirm alignment with the true data distribution.

These results indicate that our pipeline not only outperforms conventional resampling approaches such as SMOTE-Tomek and SMOTE-ENN (see Figure 5, where we can see the deterioration of the performance at the same noise levels), but also provides a robust alternative for domains where noisy and imbalanced data are the norm, including medical diagnosis and fraud detection. By combining adaptive filtering, synthetic augmentation, and noise-aware training, the framework sets a new benchmark for reliability in imperfect real-world datasets.

A limitation is that while performance is stable up to 32% noise, degradation is expected at higher levels due to irreducible label uncertainty. In addition, the computational overhead of CGAN training may constrain deployment in real-time environments.

Future directions include incorporating semi-supervised learning to leverage unlabeled data for improved noise estimation, developing lightweight CGAN variants through knowledge distillation to reduce computational overhead, and validating synthetic samples in clinical trials to ensure biological fidelity.

References

[1] N. Alalwan, A. Alwadain, A. I. Alzahrani, A. H. Al-Bayatti, A. Abozeid, R. M. A. El-Aziz: *Advancements in brain tumor identification: Integrating synthetic GANs with federated-CNNs in medical imaging analysis*, Alexandria Engineering Journal 105 (Oct. 2024), pp. 105–119, DOI: 10.1016/j.aej.2024.06.080.

[2] N. Azhar, M. S. Mohd Pozi, A. Mohamed Din, A. Jatowt: *An Investigation of SMOTE Based Methods for Imbalanced Datasets with Data Complexity Analysis*, IEEE Transactions on Knowledge and Data Engineering 35 (July 2023), pp. 6651–6672, DOI: 10.1109/TKDE.2022.3179381.

[3] G. E. Batista, R. C. Prati, M. C. Monard: *A study of the behavior of several methods for balancing machine learning training data*, ACM SIGKDD Explorations Newsletter 6.1 (2004), pp. 20–29, doi: 10.1145/1007730.1007735.

[4] R. J. Bolton, D. J. Hand: *Statistical fraud detection: A review*, Statistical Science 17.3 (2002), pp. 235–255, doi: 10.1214/ss/1042727940.

[5] M. Buda, A. Maki, M. A. Mazurowski: *A systematic study of the class imbalance problem in convolutional neural networks*, Neural Networks 106 (2018), pp. 249–259, doi: 10.1016/j.neunet.2018.07.011.

[6] C. Chen, W. Shen, C. Yang, W. Fan, X. Liu, Y. Li: *A New Safe-Level Enabled Borderline-SMOTE for Condition Recognition of Imbalanced Dataset*, IEEE Transactions on Instrumentation and Measurement 72 (2023), pp. 1–10, doi: 10.1109/TIM.2023.3289545.

[7] Z. Chen, F. Wang, R. Mu, P. Xu, X. Huang, W. Ruan: *Nrat: towards adversarial training with inherent label noise*, Machine Learning 113.6 (2024), pp. 3589–3610, doi: 10.1007/s10994-023-06437-3.

[8] H. Guo, Y. Li, J. Shang, M. Gu, Y. Huang, B. Gong: *Learning from class-imbalanced data: Review of methods and applications*, Expert Systems with Applications 73 (2017), pp. 220–239, doi: 10.1016/j.eswa.2016.12.035.

[9] D. Karimi, H. Dou, S. K. Warfield, A. Gholipour: *Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis*, Medical Image Analysis 65 (2020), p. 101759, doi: 10.1016/j.media.2020.101759.

[10] B. Krawczyk: *Learning from imbalanced data: open challenges and future directions*, Progress in Artificial Intelligence 5.4 (2016), pp. 221–232, doi: 10.1007/s13748-016-0094-0.

[11] Y. LeCun, Y. Bengio, G. Hinton: *Deep learning*, Nature 521.7553 (2015), pp. 436–444, doi: 10.1038/nature14539.

[12] A. S. Lundervold, A. Lundervold: *An overview of deep learning in medical imaging focusing on MRI*, Zeitschrift für Medizinische Physik 29.2 (2019), pp. 102–127, doi: 10.1016/j.zemedi.2018.11.002.

[13] B. Natarajan, A. Srivastava: *Handling Imbalanced Data: SMOTE and Beyond*, in: Pro Deep Learning with TensorFlow 2.0, ed. by S. Pattanayak, Apress, Berkeley, CA, 2019, pp. 569–589, doi: 10.1007/978-1-4842-4470-8_34.

[14] B. S. Raghuwanshi, S. Shukla: *Class-specific cost-sensitive boosting weighted ELM for class imbalance learning*, Memetic Computing 11.3 (2018), pp. 263–283, doi: 10.1007/s12293-018-0267-4.

[15] V. Sampath, I. Maurtua, J. J. A. Martín, A. Gutierrez: *A survey on generative adversarial networks for imbalance problems in computer vision tasks*, Journal of Big Data 8.1 (Jan. 2021), p. 27, doi: 10.1186/s40537-021-00414-0.

[16] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee: *Learning From Noisy Labels With Deep Neural Networks: A Survey*, IEEE Transactions on Neural Networks and Learning Systems 34.11 (Nov. 2023), pp. 8135–8153, doi: 10.1109/TNNLS.2022.3152527.

[17] S. Szeghalmy, A. Fazekas: *A comparative study on noise filtering of imbalanced data sets*, Knowledge-Based Systems 301 (2024), p. 112236, doi: 10.1016/j.knosys.2024.112236.

[18] F. Thabtah, S. Hammoud, F. Kamalov, A. Gonsalves: *Data imbalance in classification: Experimental evaluation*, Information Sciences 513 (2020), pp. 429–441, doi: 10.1016/j.ins.2019.11.004.

[19] M. Yang, C. Lai, C. Lin: *A robust EM clustering algorithm for Gaussian mixture models*, Pattern Recognition 45.11 (2012), pp. 3950–3961, doi: 10.1016/j.patcog.2012.04.031.

[20] W. Zhao, S. Alwidian, Q. H. Mahmoud: *Adversarial training methods for deep learning: A systematic review*, Algorithms 15.8 (2022), p. 283, doi: 10.3390/a15080283.