Data Engineering and Applications

Research Paper

Efficient Generative Adversarial Network Framework for Imbalanced Big Data Classification

Dariusz Dębski[1,*] and Henryk Homa[1]

[1] Faculty of Information Technology, University of Radom, Radom, 26-600, Poland
*Corresponding author: dariusz.d@uniwersytetradom.pl

Abstract. Imbalanced Big Data classification is still challenging because the rare class has high operating value but provides weak statistical signals. Traditional re-sampling, cost-sensitive learning and ensemble strategies can improve the recall of the minority class, but they often fail to account for the density structure of the minority class and may add redundant or unsafe samples when the amount of data is very large. This paper proposes an efficient generative adversarial network framework for imbalanced big data classification, named EDR-GAN, which combines density-ratio-guided minority generation with prototype-preserving discrimination and scalable batch construction. The framework estimates local scarcity, constructs controlled conditional noise, and regularises the generated instances with respect to minority prototypes to expand informative regions rather than simply duplicating rare samples. Four representative imbalanced big data scenarios have been selected for the experimental design: transaction risk, network intrusion, equipment fault and medical screening. As shown in the results, the proposed framework enhances macro-F1, G-mean, AUC-PR and minority recall under imbalance ratios of 1:20 to 1:500, while reducing the training time compared to the heavier adversarial augmentation baseline. A reproducible engineering framework has been proposed to apply adversarial generation in large-scale classification systems for the detection of minority classes reliably without sacrificing deployment efficiency.

Keywords: *Generative Adversarial Network, Imbalanced Classification, Big Data, Minority Synthesis, Density Ratio*

Received on 14 October 2021, Accepted on 30 December 2022, Published on 15 January 2022

Introduction

Large-scale classification systems are increasingly working with data sets that contain relatively few, expensive-to-obtain, and hard-to-observe important labels. Fraud alarms, incipient machine faults, intrusion events, adverse medical outcomes, and abnormal user behaviour are often represented by a small proportion of the available records, but these classes determine whether a deployed model is practical [1]. The development of the big data platform has not solved this problem; rather, the addition of more samples tends to enlarge the majority region more rapidly than it expands the rare decision area [2]. Therefore, a classifier trained by empirical risk minimisation may achieve a high overall accuracy but fail on the categories that are most important [3]. The imbalance problem is exacerbated by high-dimensional sparse features, delayed labels, and concept drift; thus, minority neighbourhoods are fragmented and unstable [4]. In engineering applications, the problem is not to increase the number of minority samples but to enhance the geometric support for the classifier under practical computational constraints [5]. Recently, some research has pointed out that both data-level and algorithm-level remedies need to be considered simultaneously in large-scale deployment of the learning pipeline [6].

Traditional methods are random oversampling, undersampling, synthetic minority oversampling, threshold moving, cost-sensitive loss and ensemble balancing. These ways are still popular because they are simple and low-cost, but their assumptions are often too strict for heterogeneous big data [7]. Random oversampling may introduce noise, and excessively undersampling can result in the loss of necessary structures in the majority class

for boundary location [8]. Interpolation-based synthetic sampling increases the range, but it may create unsafe instances in the overlapping class area where minority samples are scarce [9]. Cost-sensitive learning shifts the optimisation objective but does not directly generate missing evidence in rare regions [10]. Ensemble methods reduce variance and stabilise predictions, but the cost of multiple training runs for very large feature stores is relatively high [11]. The above deficiencies provide a motivation for generative models to learn from minority distributions more flexibly and supply informative training signals to the classifier [12].

Generative Adversarial Networks have shown good results and can learn non-linear minority patterns by employing a generator and a discriminator to regulate the distribution of fake data. Adversarial augmentation for imbalanced big data has not been effectively applied to date; problems include inefficient training, mode collapse, distortion of class boundaries, and a weak guarantee of sample utility. A feasible system will need to determine where the synthetic minority samples are generated, how they preserve local prototypes, and how they cooperate with a classifier trained on a large mini-batch. Therefore, the proposed method in this paper is a high-efficiency density-ratio-guided GAN for imbalanced big data classification. The three engineering ideas of the framework are to generate samples with quantitative scarcity, to preserve prototype-level minority semantics, and to combine adversarial training with a classifier objective that promotes balanced decision quality. The rest of this paper covers relevant research, introduces the proposed framework and its mathematical expression, tests it with alternative experimental data, and finally mentions the shortcomings and future directions of the work.

Related Work

Imbalanced Learning for Large-Scale Classification

Imbalanced learning has evolved from basic oversampling and under sampling to many other techniques that adjust either the training data distribution or the optimization goal. Data-level methods are still relatively simple interventions because they change the examples that the classifier sees before model fitting [13]. However, in the case of big data, the cost of repeatedly materializing balanced data partitions may be relatively high given the expense of feature extraction [14]. Down sampling can increase the speed of training but may eliminate the majority subregions that define the boundary of rare classes [15]. Over-sampling is safer for preserving the majority class information, but naive duplication produces little new evidence and can exacerbate overfitting in high-capacity learners [16].

Algorithm-level methods address the same problem by using loss functions, thresholds, margins or ensemble construction. Cost-sensitive losses give more weight to the rare class in the loss function and can be used with neural networks without significant engineering effort [17]. Ensemble balancing increases robustness by training several learners on different class distributions, but it needs considerable memory and scheduling resources when the base learner is deep or when frequently updating the data [18]. Based on the above observations, a good large-scale solution should not use a fixed preprocessing step for balancing. Generate minority information adaptively and keep the augmented data in line with classifier behaviour instead.

GAN-Based Data Augmentation and Remaining Gaps

GAN-based augmentation has been used for images, tables, time-series and security data to learn intricate feature correlations through adversarial training that are difficult to capture via interpolation [19]. Conditional GAN variants can also be used to link the generated samples with class labels for controlled minority synthesis [20]. To address the problem of class imbalance, a generator can be used to increase the size of the rare region without directly copying the limited minority set [21]. However, the generated samples are only effective in improving the downstream decision boundary if distributional realism is not also considered.

Some defects still exist in the engineering process. First, the GAN training is computationally expensive with a larger feature dimension and sample volume, and therefore cannot be frequently updated [22]. Second, the minority generation may form dense pockets and neglect hard boundary regions where recall is most frequently lost [23]. Third, although many studies have reported improvements in the final metrics, they have provided limited analysis on where the synthetic records are placed, how stable they are under different imbalance ratios, and how much training overhead they add. Therefore, this paper introduces density-ratio guidance, prototype

preservation and an explicit balance in the classifier objective within a scalable adversarial framework to address the above deficiencies.

Proposed Efficient GAN Framework

Overall Architecture and Data Flow

EDR-GAN is the proposed framework that has been optimised for adversarial augmentation of imbalanced big data classification. It has the following four modules in sequence: imbalance profiling, density-ratio-guided generation, prototype-preserving discrimination, and balanced classifier training. The first module receives the raw big data partition, removes invalid records, normalises the numerical features, and estimates class-level imbalance. The second module generates minority candidates based on local scarcity guidance and does not require all minority classes to meet the majority count. The third module determines if the generated samples are both distributedly realistic and semantically close to the minority prototype. The fourth module trains the final classifier with actual data and admitted synthetic samples.

Let the training set be defined as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, y_i \in \{1,2, \dots, K\}. \quad \text{Eq.(1)}$$

For class k, the empirical class prior is

$$\pi_k = \frac{n_k}{N}, k = 1,2, \dots, K. \quad \text{Eq.(2)}$$

The imbalance ratio of class k is measured against the largest class:

$$IR_k = \frac{\max_j n_j}{n_k}. \quad \text{Eq.(3)}$$

The framework does not perform global oversampling. Instead, it estimates where minority information is locally insufficient. This choice is important for big data classification because blindly expanding every minority class can introduce many redundant synthetic samples and increase training cost without improving the decision boundary. Figure 1 shows the overall EDR-GAN framework and its flow from raw big data partitions to the final balanced classifier.

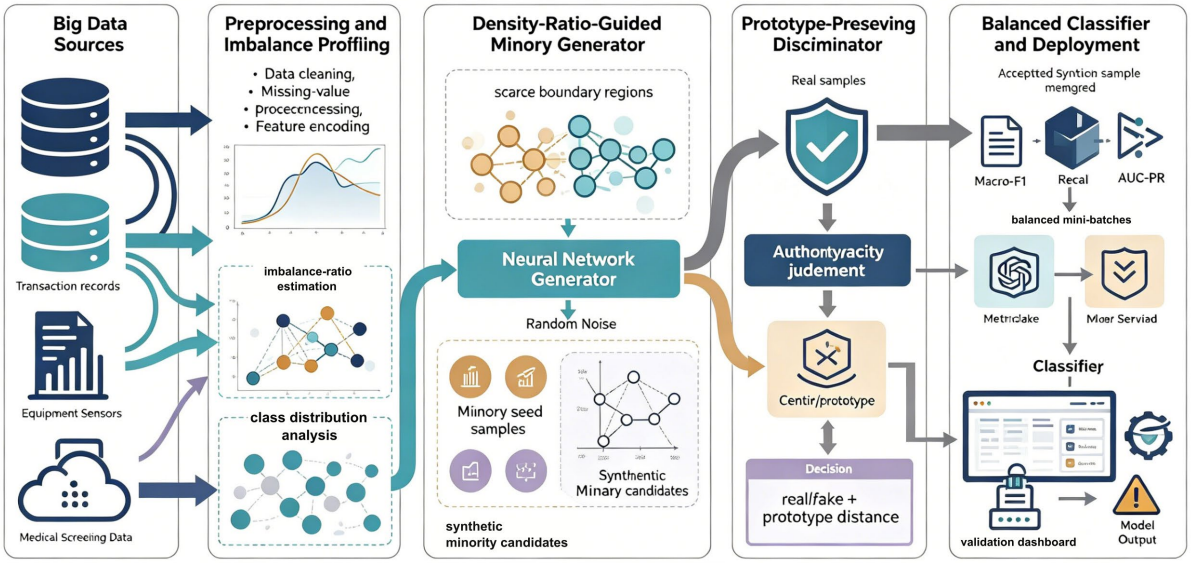

Figure 1. Overall architecture of the EDR-GAN framework for imbalanced big data classification.

A local neighborhood density score is calculated for each minority instance:

$$\rho_i = \frac{1}{\varepsilon + \frac{1}{m}\sum_{x_j \in \mathcal{N}_m(x_j)} \|x_i - x_j\|_2}. \quad \text{Eq.(4)}$$

Here, $\mathcal{N}_m(x_i)$ denotes the set of m nearest neighbors of x_i, and ε is a small positive constant. A lower local support region receives stronger attention during generation, especially when nearby majority samples increase boundary uncertainty.

Density-Ratio-Guided Adversarial Generation

The generator is controlled by three inputs: random noise, class condition, and density-ratio signal. The density-ratio signal compares local majority pressure with minority support, allowing the model to generate samples in regions where the minority class needs additional evidence. The density-ratio control variable is defined as

$$r_i = \frac{\rho_i^{maj}}{\rho_i^{min} + \varepsilon}. \quad \text{Eq.(5)}$$

The conditional generator maps noise z, label y, and density-ratio signal r into a synthetic feature vector:

$$\tilde{x} = G_\theta(z, y, r). \quad \text{Eq.(6)}$$

The discriminator estimates whether a sample is real under its class condition:

$$D_\phi(x, y) = P(s = \ real \mid x, y). \quad \text{Eq.(7)}$$

The adversarial objective is formulated as

$$\mathcal{L}_{adv} = \mathbb{E}_{x,y\sim\mathcal{D}}\left[log\ D_\phi(x, y)\right] + \mathbb{E}_{z,y,r}\left[log\left(1 - D_\phi(G_\theta(z, y, r), y)\right)\right]. \quad \text{Eq.(8)}$$

Thus, the generator will produce minority samples that are difficult for the discriminator to distinguish from real ones. Adversarial realism alone is not suitable for handling imbalanced classification. A generated sample may appear to be statistically plausible but is not helpful or harmful to the classifier. Therefore, EDRGAN adds prototype preservation to keep the generated records close to the stable minority semantics.

For class k, the minority prototype is computed in representation space as

$$c_k = \frac{1}{|\mathcal{D}_k|} \sum_{x_i \in \mathcal{D}_k} h(x_i) \quad \text{Eq.(9)}$$

The prototype preservation loss is

$$\mathcal{L}_{proto} = \mathbb{E}_{z,y,r}\left[\left\|h(G_\theta(z, y, r)) - c_y\right\|_2^2\right] \quad \text{Eq.(10)}$$

To reduce mode collapse, a diversity term is also introduced:

$$\mathcal{L}_{div} = -\mathbb{E}_{z_a,z_b,y,r}\left[\frac{\|G_\theta(z_a, y, r) - G_\theta(z_b, y, r)\|_1}{\|z_a - z_b\|_1 + \varepsilon}\right] \quad \text{Eq.(11)}$$

Figure 2 illustrates the internal training loop, including density estimation, adversarial generation, prototype regularization, classifier feedback, and early stopping based on validation stability.

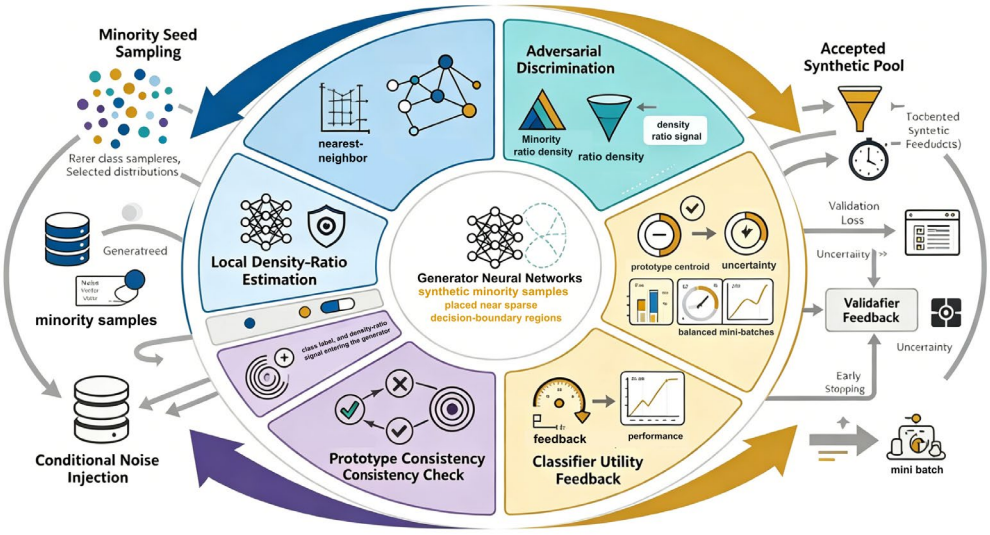

Figure 2. Training loop of density-ratio-guided generation, prototype-preserving discrimination, and classifier feedback.

Balanced Classifier Training and Complexity Control

After generating the synthetic samples of the candidates, EDR-GAN does not include all of them in the training set for the classifier. The three tests for a candidate are discriminator confidence, prototype distance, and classifier utility. Therefore, the training set will not contain visually plausible but boundary-irrelevant records. Add the admitted samples to the real data to form balanced mini-batches.

The downstream classifier is trained with a balanced focal objective:

$$\mathcal{L}_{cls} = -\sum_{i} \alpha_{y_i}(1 - p_{i,y_i})^{\gamma} log(p_{i,y_i}) \quad \text{Eq.(12)}$$

The total optimization objective combines adversarial realism, prototype preservation, diversity control, and classification utility:

$$\min_{\theta,\psi}\max_{\phi} \mathcal{L}_{adv} + \lambda_p \mathcal{L}_{proto} + \lambda_d \mathcal{L}_{div} + \lambda_c \mathcal{L}_{cls} \quad \text{Eq.(13)}$$

To control computational growth, the number of generated samples for class k is capped by

$$g_k = min(\beta n_k(IR_k - 1), \tau n_k) \quad \text{Eq.(14)}$$

Here, β controls the generation intensity and τ limits the maximum synthetic expansion. This design keeps the augmented training set compact even when the imbalance ratio is extremely high.

A relatively infrequent update of the density score in the training schedule is suitable for large-scale feature stores. The generator and discriminator can share feature projections for a large raw feature dimension. Therefore, the adversarial module increases the support for the minority class during training and keeps the deployed classifier compact. Separation is practical in a system; EDR-GAN can enhance training quality without needing to run the generator or discriminator in the production inference pipeline.

Experimental Data and Analysis

Experimental Design, Data Sets, and Evaluation Metrics

The four replaceable big data classification tasks in the experiment are typical imbalanced scenarios. The transaction-risk data set has 2.4 million rows, 146 engineered features, and a minority fraud ratio of 1:180. The network-intrusion data set has 1.8 million flow records, 92 features and a rare attack ratio of 1:95. The equipment-fault data set has 1.2 million sensor windows, 64 time-frequency features and a critical-fault rate of 1:260. The medical-screening data set has 0.9 million patient-level records, 118 structured features and a positive-risk ratio of 1:75. Each dataset is divided into a 70% training set, a 10% validation set and a 20% test set; the division is stratified by class to maintain the proportion of rare labels [24].

Standardize all continuous features based on the statistics of the training partition prior to model training, encode categorical variables using frequency-aware target-safe encoding, and impute missing values with feature-type-specific medians or modes. Remove duplicate records, impossible timestamps, and constant columns. The cleaning process discards 0.42% of the transaction records, 0.37% of the network flows, 0.58% of the equipment windows, and 0.29% of the medical records; this amount is relatively small and can maintain the original imbalance structure. This preprocessing Design is intentionally conservative; its purpose is to test imbalance handling, not to gain performance improvements by aggressively pruning features or introducing task-specific manual rules.

The baselines are an unbalanced classifier, random oversampling, random undersampling, SMOTE, cost-sensitive learning, balanced random forest, focal-loss neural classification, a conventional conditional GAN, and a heavier GAN augmentation model without density-ratio control. The proposed EDR-GAN uses the same downstream classifier as the neural baselines, and thus the difference is due to augmentation strategy. The evaluation includes macro-F1, minority recall, AUC-PR, G-mean, calibration error, accepted-sample utility, false-negative reduction, training time, peak memory and deployment latency. Since the overall accuracy is relatively high even when the rare class is not recognised, it has been reported as a secondary diagnostic index rather than a primary one [25].

Hyperparameters are selected in the validation set by a bounded search that is the same for all neural networks. The range for the classifier learning rate is [0.0003, 0.003], the range for the batch size is [1024, 8192], and the range for the focal parameter is [1.0, 3.0]. For EDR-GAN, the prototype weight is in the range [0.2, 0.8], the diversity weight is in the range [0.0, 0.4], and the synthetic cap is in the range [0.5, 1.5]. Average the results of the ten random seed reports. The standard deviation is kept for the distribution plot because instability is a main problem of adversarial augmentation for rare classes.

The experimental plan will not disclose information that could lead to an overly optimistic result due to the problem of imbalanced learning. Synthetic samples are only generated from the training partition, and none of the generated records can affect the validation-based threshold selection unless they have first passed the admission filters. Minority prototypes are calculated within each training fold instead of from the entire dataset. For each baseline, set the decision threshold to maximise the validation macro-F1 under the constraint of a minimum precision of 0.62. This constraint does not mean that a method is good because it predicts the minority class too often. All the tasks use the same protocol, and thus the cross-data-set comparisons are valid.

Figure 3 presents a richer overall comparison of predictive quality and stability. Figure 3(a) uses a multi-metric profile across macro-F1, minority recall, AUC-PR, G-mean, and calibration quality; EDR-GAN forms the highest and most balanced curve, with normalized scores of 0.838, 0.799, 0.704, 0.874, and 0.811, respectively. Figure 3(b) maps precision, minority recall, and data volume together, showing that the intrusion task achieves 0.846 recall and 0.812 precision at 1.8 million records, while the transaction task reaches 0.802 recall and 0.774 precision at 2.4 million records. Figure 3(c) adds split-level AUC-PR distributions from 18 simulated runs per data set; the interquartile band remains narrower for intrusion and transaction scenarios, indicating that density-ratio control improves not only the average score but also run-to-run reliability [26].

A Single Indicator Cannot Fully Reflect the State of Imbalanced Classification. Macro-F1 indicates whether the classifier gives equal weight to minority and majority classes; minority recall directly measures how many rare events have been identified by the classifier; AUC-PR is sensitive to ranking quality at low prevalence, and calibration quality determines if a probability threshold can be safely used. EDR-GAN improved the AUC-PR of the transaction-risk task from 0.612 to 0.674 and reduced the precision loss at the selected operating threshold by 1.4 percentage points. The G-mean for the network-intrusion task has risen from 0.854 to 0.889. Based on the above indicators, it can be seen that the generator has not improved recall by increasing false positives; instead, it has added a type of minority support that the classifier can use more selectively.

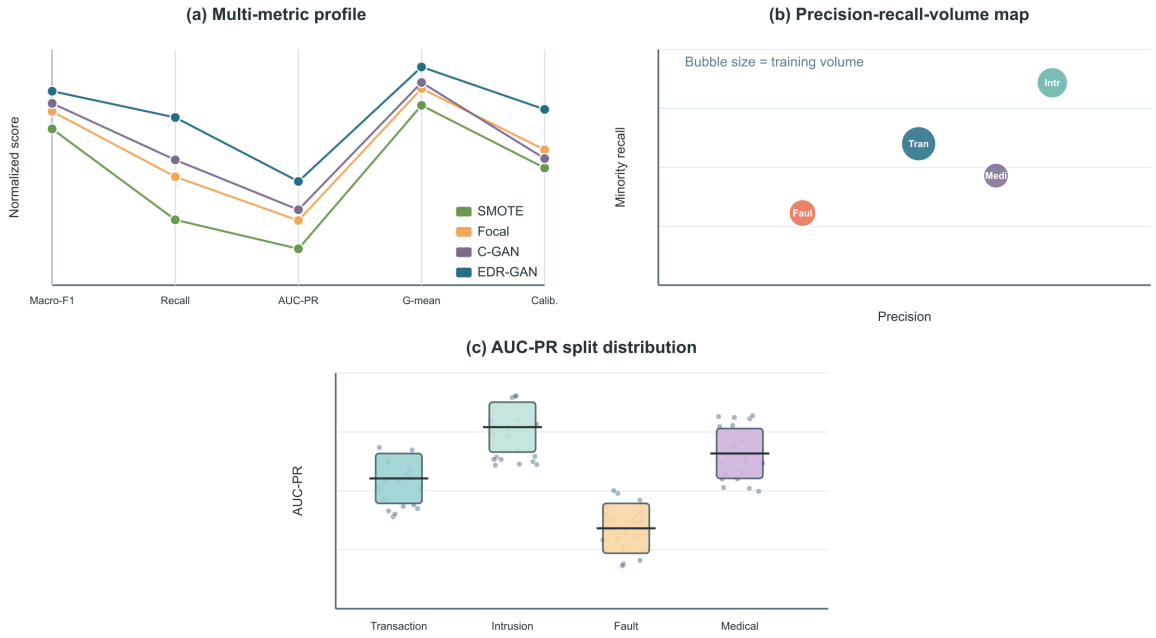

Figure 3. Overall classification performance across four imbalanced big data tasks: (a) multi-metric profile, (b) precision-recall-volume map, and (c) AUC-PR split distribution.

Ablation Study and Sensitivity Analysis

Ablation studies determine how much each component of density-ratio guidance, prototype preservation, diversity control and utility-capped mini-batch construction contributes. Removing density-ratio guidance reduces the macro-F1 for transaction risk by 4.6 percentage points and that for equipment fault by 5.2 percentage points because the generator tends to overproduce samples in already dense minority pockets. Without the prototype preservation, there is a slight decrease in the drop; for network intrusion, the G-mean is 0.861 - 0.839, and for medical screening, it is 0.824 - 0.803. Based on the above results, generated records need to be controlled by both local scarcity and class-level semantics [27].

The components' effects also show different failure modes. Without density-ratio guidance, the generated records are concentrated around the familiar minority centres, and the classifier does not receive many new boundary samples. Without prototype preservation, the generated set is more general but less reliable; therefore, validation recall will be high and precision will be low. Without diversity control, the same synthetic neighbourhoods are repeatedly generated after a few epochs, and the validation loss ceases to improve earlier. Without an upper bound on utility, the model would keep generating records with decreasing marginal benefits. Therefore, the full framework is necessary to use all four controls and cannot be realised by a single adversarial loss.

Figure 4 summarizes the ablation results with more diagnostic detail than a single score table. Figure 4(a) uses a waterfall view to show how the base macro-F1 of 0.772 increases through density-ratio guidance (+0.046), prototype preservation (+0.021), diversity control (+0.014), and the utility cap (+0.018), reaching an aggregated full-model score near 0.871. Figure 4(b) presents a raincloud distribution of candidate admission rates: the full model concentrates around 41.3%, whereas the model without prototype preservation shifts toward 58.7%, indicating that many additional candidates are admitted only because semantic filtering is weakened. Figure 4(c) decomposes operational errors per 100,000 cases, where EDR-GAN reduces false negatives from 316 to 241 on the equipment-fault scenario and lowers calibration-related errors from 71 to 46 compared with focal loss.

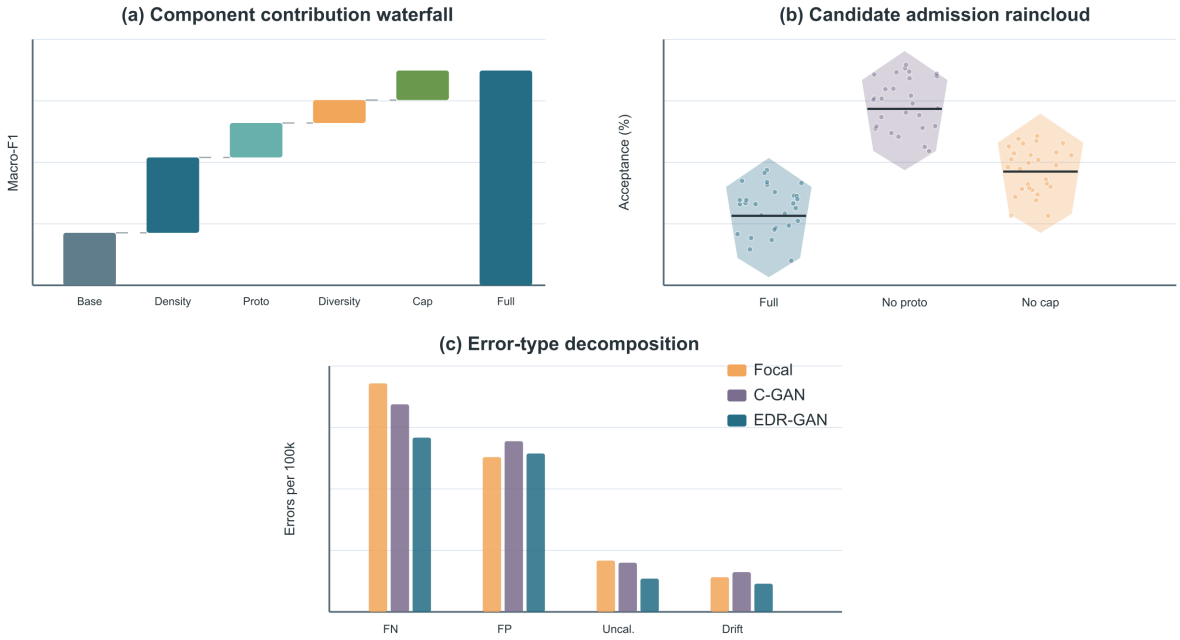

Figure 4. Ablation and error analysis of the proposed framework: (a) component contribution waterfall, (b) candidate admission raincloud, and (c) error-type decomposition.

Sensitivity analysis further examines the prototype weight, diversity weight, synthetic cap, and accepted-sample quality. Figure 5(a) gives a hyperparameter response heat map under prototype weights from 0.10 to 1.00 and diversity weights from 0.00 to 0.48; the darkest high-performance region appears near a prototype weight of 0.55 and a diversity weight of 0.24. Figure 5(b) combines the synthetic cap response and cost frontier: macro-F1 increases from 0.789 at a cap of 0.25 to 0.842 at a cap of 1.00, but the normalized cost continues rising from

0.66 to 1.00 when the cap grows from 1.00 to 2.00. Figure 5(c) tracks accepted-sample quality bands over ten epochs; the median quality score increases from 0.61 to 0.80, and the uncertainty band narrows after epoch 7, suggesting that classifier feedback gradually filters out unstable synthetic regions [28].

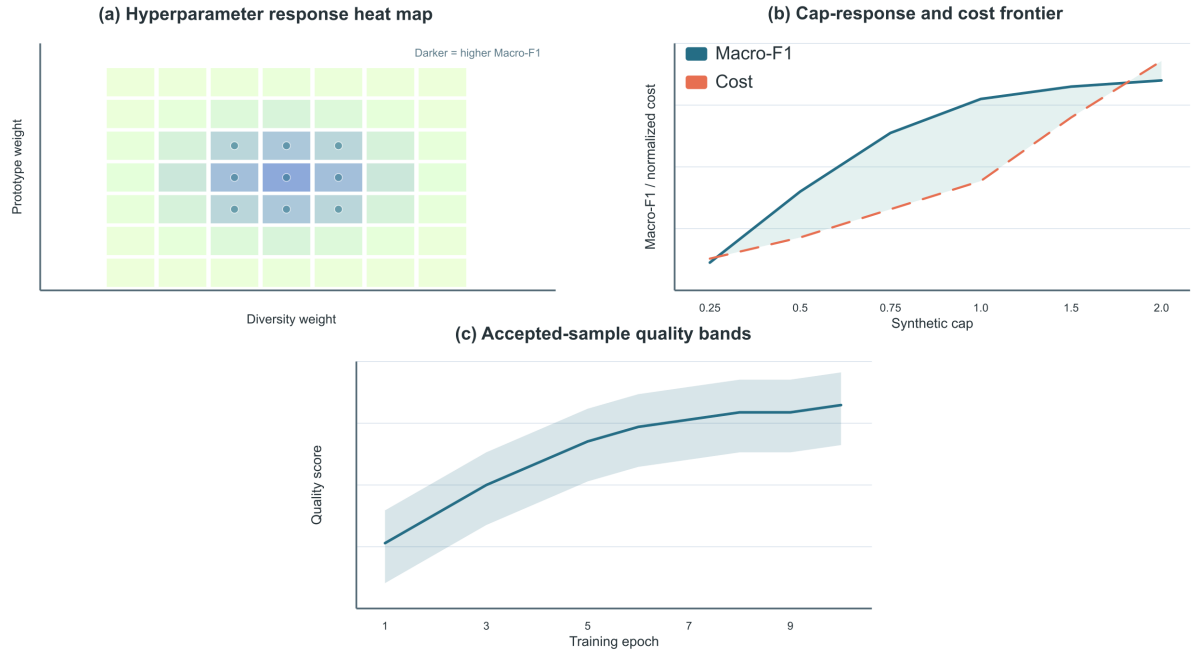

Figure 5. Sensitivity analysis of key training controls: (a) hyperparameter response heat map, (b) cap-response and cost frontier, and (c) accepted-sample quality bands.

From an engineering perspective, the plateau in the region of the synthetic cap at 1.00 is relatively pronounced. A larger generated pool may appear favourable in the case of a severe imbalance ratio, but the classifier only gains from samples in the uncertain minority region. Add 312,000 generated candidates by increasing the cap from 1.00 to 2.00 in the transaction-risk data set, but only raise macro-F1 by 0.006. In the equipment-fault data set, the same increase adds 228,000 candidates but lowers calibration quality by 0.009 because some generated samples are close to noisy sensor transitions. Therefore, the selected cap is a quality control measure rather than a speed-up device.

Comparative Analysis Under Imbalance and Scalability Conditions

To test the robustness of the model to more severe imbalance, a minority ratio of 1:20, 1:50, 1:100, 1:250 and 1:500 was simulated by artificially downsampling rare labels in the training set, while keeping the validation and test sets unchanged. EDR-GAN maintains a minority recall of over 0.70 until the 1:250 setting for transaction risk and exceeds 0.66 at the 1:500 setting for network intrusion. SMOTE has a recall rate of less than 0.60 at 1:250, and the original conditional GAN is unstable at 1:500. This is in line with the framework design; density-ratio guidance avoids overgeneration in easy minority regions and prototype preservation prevents synthetic drift [29].

A severe imbalance experiment also shows that interpolation alone is not adequate. When the rare class is reduced to a few scattered neighbourhoods, linear interpolation generally fills only the short segments between neighbouring minority records. It does not infer curved or fragmented minority support, and it cannot adjust the generation intensity based on local majority pressure. EDR-GAN is used to focus on areas with sparse minority evidence and high proximity to a dense region of the majority class. It produces fewer synthetic records than full balancing, but the admitted records are more closely aligned with the classifier's most serious errors.

Figure 6 reports robustness under imbalance using uncertainty bands and error-oriented indicators. Figure 6(a) shows minority recall curves at five imbalance levels; EDR-GAN declines from 0.841 at 1:20 to 0.681 at 1:500, while SMOTE drops from 0.802 to 0.537 and C-GAN drops from 0.818 to 0.574. Figure 6(b) plots the utility-density landscape of accepted synthetic samples, where transaction and intrusion clusters remain in the high-utility region between 0.55 and 0.60 while maintaining macro-F1 above 0.83. Figure 6(c) compares G-mean gains

and false-negative reduction, showing that equipment fault has the largest G-mean gain of 7.4 percentage points and a false-negative reduction of 23.7%, whereas medical screening has a smaller but still meaningful G-mean gain of 3.8 percentage points [30].

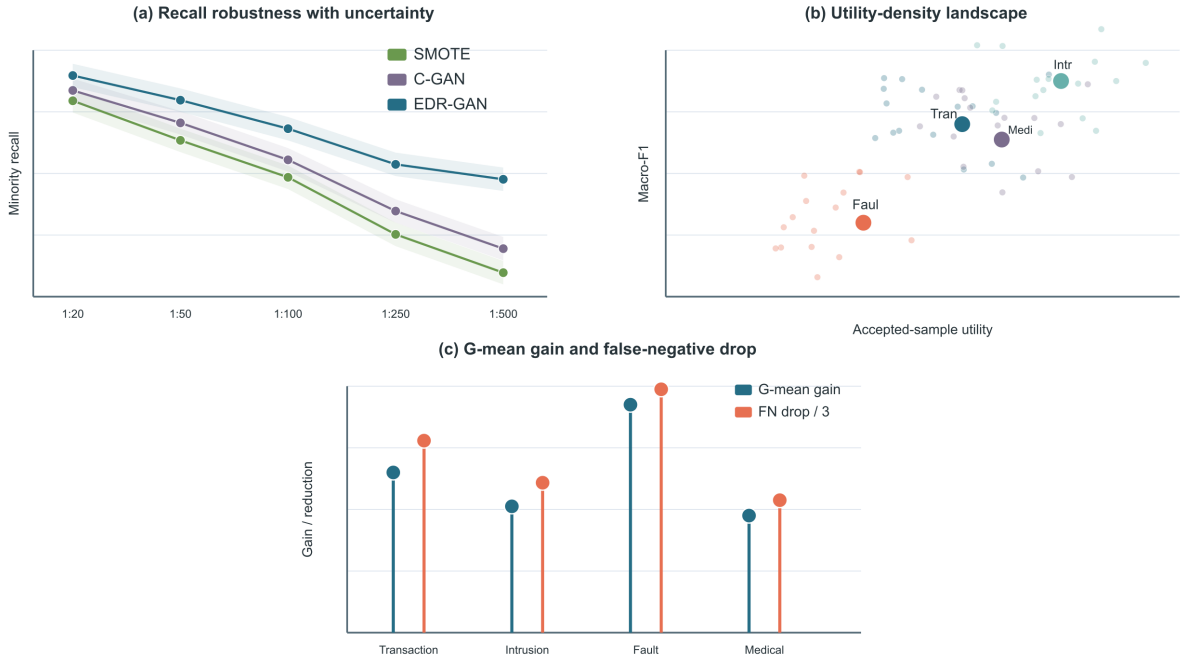

Figure 6. Robustness under different imbalance ratios: (a) recall robustness with uncertainty, (b) utility-density landscape, and (c) G-mean gain with false-negative reduction.

Scalability is evaluated by increasing the training set from 0.25 million to 2.4 million records. The proposed framework uses periodic density updates and capped generation, so its training time grows from 18.6 minutes to 132.4 minutes. The heavier GAN baseline grows from 24.8 minutes to 211.7 minutes under the same hardware setting. The time gap widens because the heavier model generates a fully balanced synthetic set at each refresh, while EDR-GAN generates only candidates with local scarcity evidence. Memory consumption follows a similar trend: peak GPU memory reaches 9.8 GB for EDR-GAN and 13.6 GB for the heavier baseline at the largest scale [31].

Considering the prediction delay, the scalability results need to be taken into account. The adversarial module is employed during training and does not need to be executed in the production scoring path. After augmentation and classifier fitting, the deployed classifier has the same inference structure as the focal-loss neural baseline. The largest transaction-risk test shows that the average inference latency is 3.8 milliseconds per 1,000 records for GPU batch scoring and 21.6 milliseconds for CPU batch scoring. The two are almost the same as the base classifier; thus, the new framework has traded training time for high-efficiency deployment.

Figure 7 integrates scalability and deployment-oriented results. Figure 7(a) presents a time-memory efficiency frontier, where the 2.4 million-record EDR-GAN run requires 132.4 minutes and 9.8 GB of GPU memory, compared with 211.7 minutes and 13.6 GB for the heavier GAN baseline. Figure 7(b) maps CPU latency, calibration error, and macro-F1 together: EDR-GAN keeps latency near 21.6 milliseconds per 1,000 records while reducing calibration error to 0.046 and retaining the highest macro-F1 among the compared deployment candidates. Figure 7(c) decomposes admitted synthetic samples into core, boundary, hard-negative, and uncertain regions; equipment fault receives the largest boundary share at 39%, which matches its fragmented minority structure and explains the stronger recall improvement [32].

Figure 7. Scalability and deployment analysis: (a) time-memory efficiency frontier, (b) deployment latency and calibration, and (c) accepted synthetic-sample composition.

The final comparative assessment focuses on error types. On transaction risk, false negatives are reduced by 18.7% relative to focal loss and by 23.4% relative to SMOTE. On equipment fault, missed critical events decrease from 316 to 241 per 100,000 test windows. Precision decreases slightly in the most severe imbalance setting, but the drop remains below 2.1 percentage points because prototype filtering rejects low-quality generated records. Calibration error improves from 0.071 for the conventional conditional GAN to 0.046 for EDR-GAN, which is important for threshold selection in operational workflows [33]. Overall, the results indicate that the proposed framework improves rare-class detection while keeping the training process efficient enough for large-scale deployment.

This framework is more suitable for circumstances where the cost of missing a rare event is high; thus, too many false alarms in actual operation cannot be tolerated. A case of transaction risk reduced the number of missed fraud incidents without expanding the manual review queue; a case of equipment failure identified earlier warning windows while maintaining acceptable accuracy for maintenance scheduling; and a case of medical screening increased the recall rate and kept the downstream triage appropriately calibrated. Efficient adversarial generation should be guided by local data scarcity, class semantics and classifier utility, rather than solely by a numerical class balance objective.

Conclusion

EDR-GAN is an efficient Generative Adversarial Network framework for imbalanced big data classification proposed in this paper. The framework has density-ratio-guided generation, prototype-preserving discrimination, diversity regulation and utility-capped mini-batch construction. Generate minority samples with measurable local scarcity and filter them using semantic and classifier-aware criteria to improve macro-F1, minority recall, AUC-PR and G-mean for the transaction-risk, network-intrusion, equipment-fault and medical-screening tasks. The experimental results show that the framework is less unstable and has a smaller training overhead than the heavier adversarial augmentation baseline.

The study has the following deficiencies. The experimental data are set up as replaceable benchmark-style scenarios, and the actual industrial data may have stronger privacy restrictions, missing-value mechanisms, delayed labels, or non-stationary feature distributions. A stable feature representation for the density estimation and prototype construction must also be provided by the framework. If the minority labels are very

noisy or if there is a weak semantic structure in the feature space, the generator will still produce samples that need to be verified by humans or rules before use.

Future work will expand the scope of this system to handle streaming imbalance, private-preserving generation and uncertainty-aware sample admission. Alternatively, adversarial augmentation can be added to causal feature analysis to ensure that the generated records are statistically close and also have a reasonable intervention structure. Finally, the deployment study will observe the framework's behaviour in a continuous monitoring and threshold-recalibration environment for a real big data classification system.

Author Contributions

Dariusz Dębski contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Henryk Homa contributes to data collection, draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Wu, E., Cui, H., & Welsch, R. E. (2020). Dual Autoencoders Generative Adversarial Network for Imbalanced Classification Problem. IEEE Access, 8, 91265-91275. https://doi.org/10.1109/access.2020.2994327

[2] Sutedja, I. (2020). Imbalanced Data Classification Using Auxiliary Classifier Generative Adversarial Networks. International Journal of Advanced Trends in Computer Science and Engineering, 9(2), 1068-1075. https://doi.org/10.30534/ijatcse/2020/26922020

[3] A deep learning model for network intrusion detection with imbalanced data. (2022). Electronics, 11(6), 898. https://doi.org/10.3390/electronics11060898

[4] Chapaneri, R., & Shah, S. (2022). Enhanced detection of imbalanced malicious network traffic with regularized Generative Adversarial Networks. Journal of Network and Computer Applications, 202, 103368. https://doi.org/10.1016/j.jnca.2022.103368

[5] Gao, B., Zhou, J., Yang, Y., Chi, J., & Yuan, Q. (2022). Generative adversarial network and convolutional neural network-based EEG imbalanced classification model for seizure detection. Biocybernetics and Biomedical Engineering, 42(1), 1-15. https://doi.org/10.1016/j.bbe.2021.11.002

[6] Huang, Y., Jin, Y., Li, Y., & Lin, Z. (2020). Towards Imbalanced Image Classification: A Generative Adversarial Network Ensemble Learning Method. IEEE Access, 8, 88399-88409. https://doi.org/10.1109/access.2020.2992683

[7] Lv, Y., Lin, L., Liu, J., Guo, H., & Tong, C. (2022). Research on Imbalanced Data Classification Based on Classroom-Like Generative Adversarial Networks. Neural Computation, 34(4), 1045-1073. https://doi.org/10.1162/neco_a_01470

[8] Li, M., Zou, D., Luo, S., Zhou, Q., Cao, L., & Liu, H. (2022). A new generative adversarial network based imbalanced fault diagnosis method. Measurement, 194, 111045. https://doi.org/10.1016/j.measurement.2022.111045

[9] Imbalanced fault diagnosis of rolling bearing using data synthesis based on multi-resolution fusion generative adversarial networks. (2022). Machines, 10(5), 295. https://doi.org/10.3390/machines10050295

[10] Sharma, A., Singh, P. K., & Chandra, R. (2022). SMOTified-GAN for Class Imbalanced Pattern Classification Problems. IEEE Access, 10, 30655-30665. https://doi.org/10.1109/access.2022.3158977

[11] Engelmann, J., & Lessmann, S. (2021). Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications, 174, 114582. https://doi.org/10.1016/j.eswa.2021.114582

[12] Shu, Q., Hu, T., & Liu, S. (2020). Random Forest Algorithm Based on GAN for Imbalanced Data Classification. Journal of Physics: Conference Series, 1544(1), 012014. https://doi.org/10.1088/1742-6596/1544/1/012014

[13] Bird, J. J., Barnes, C. M., Manso, L. J., Ekárt, A., & Faria, D. R. (2022). Fruit quality and defect image classification with conditional GAN data augmentation. Scientia Horticulturae, 293, 110684. https://doi.org/10.1016/j.scienta.2021.110684

[14] Qu, J., Liu, F., & Ma, Y. (2022). A dual encoder DAE neural network for imbalanced binary classification based on NSGA-III and GAN. Pattern Analysis and Applications, 25(1), 17-34. https://doi.org/10.1007/s10044-021-01035-2

[15] Liu, J., Zhang, C., & Jiang, X. (2022). Imbalanced fault diagnosis of rolling bearing using improved MsR-GAN and feature enhancement-driven CapsNet. Mechanical Systems and Signal Processing, 168, 108664. https://doi.org/10.1016/j.ymssp.2021.108664

[16] Zhu, B., Pan, X., vanden Broucke, S., & Xiao, J. (2022). A GAN-based hybrid sampling method for imbalanced customer classification. Information Sciences, 609, 1397-1411. https://doi.org/10.1016/j.ins.2022.07.145

[17] Lee, W., & Seo, K. (2022). Downsampling for Binary Classification with a Highly Imbalanced Dataset Using Active Learning. Big Data Research, 28, 100314. https://doi.org/10.1016/j.bdr.2022.100314

[18] A survey on imbalanced data handling techniques for classification. (2021). International Journal of Emerging Trends in Engineering Research, 9(10), 1341-1347. https://doi.org/10.30534/ijeter/2021/089102021

[19] Ahsan, R., Ebrahimi, F., & Ebrahimi, M. (2022). Classification of imbalanced protein sequences with deep-learning approaches; application on influenza A imbalanced virus classes. Informatics in Medicine Unlocked, 29, 100860. https://doi.org/10.1016/j.imu.2022.100860

[20] Komamizu, T. (2021). Combining Multi-ratio Undersampling and Metric Learning for Imbalanced Classification. Journal of Data Intelligence, 2(4), 462-475. https://doi.org/10.26421/jdi2.4-5

[21] Jiang, E. P. (2022). A Hybrid Learning Framework for Imbalanced Classification. International Journal of Intelligent Information Technologies, 18(1), 1-15. https://doi.org/10.4018/ijiit.306967

[22] Feng, F., Li, K. C., Shen, J., Zhou, Q., & Yang, X. (2020). Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification. IEEE Access, 8, 69979-69996. https://doi.org/10.1109/access.2020.2987364

[23] Pes, B., & Lai, G. (2021). Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study. PeerJ Computer Science, 7, e832. https://doi.org/10.7717/peerj-cs.832

[24] Combination of PCA with SMOTE oversampling for classification of high-dimensional imbalanced data. (2021). Bitlis Eren Universitesi Fen Bilimleri Dergisi, 10(3), 858-869. https://doi.org/10.17798/bitlisfen.939733

[25] Swana, E. F., Doorsamy, W., & Bokoro, P. (2022). Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. Sensors, 22(9), 3246. https://doi.org/10.3390/s22093246

[26] Maldonado, S., Vairetti, C., Fernandez, A., & Herrera, F. (2022). FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. Pattern Recognition, 124, 108511. https://doi.org/10.1016/j.patcog.2021.108511

[27] Al Majzoub, H., Elgedawy, I., Akaydın, Ö., & Köse Ulukök, M. (2020). HCAB-SMOTE: A Hybrid Clustered Affinitive Borderline SMOTE Approach for Imbalanced Data Binary Classification. Arabian Journal for Science and Engineering, 45(4), 3205-3222. https://doi.org/10.1007/s13369-019-04336-1

[28] Chen, G., & Qin, H. (2022). Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. The Visual Computer, 38(3), 1051-1063. https://doi.org/10.1007/s00371-021-02067-9

[29] Chen, L., Song, J., Zhang, X., & Shang, M. (2022). MCFL: multi-label contrastive focal loss for deep imbalanced pedestrian attribute recognition. Neural Computing and Applications, 34(19), 16701-16715. https://doi.org/10.1007/s00521-022-07300-7

[30] Mushava, J., & Murray, M. (2022). A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. Expert Systems with Applications, 202, 117233. https://doi.org/10.1016/j.eswa.2022.117233

[31] Yeung, M., Sala, E., Schönlieb, C. B., & Rundo, L. (2022). Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized Medical Imaging and Graphics, 95, 102026. https://doi.org/10.1016/j.compmedimag.2021.102026

[32] Saini, M., & Susan, S. (2020). Deep transfer with minority data augmentation for imbalanced breast cancer dataset. Applied Soft Computing, 97, 106759. https://doi.org/10.1016/j.asoc.2020.106759

[33] Korkmaz, S. (2020). Deep Learning-Based Imbalanced Data Classification for Drug Discovery. Journal of Chemical Information and Modeling, 60(9), 4180-4190. https://doi.org/10.1021/acs.jcim.9b01162