Communication in Physical Sciences, 2023, 9(4): 968-979

Rare-Event Prediction in Imbalanced Data: A Unified Evaluation and Optimization Framework for High-Risk Systems

Mujeeb Abdulrazaq

Received: **18 March 2023/Accepted: 29 August 2023/Published: 19 September 2023**

Abstract: *Rare events, outcomes that occur infrequently but often carry high stakes, present a major challenge for predictive modeling due to extreme class imbalance. When the majority class vastly outnumbers the minority class, standard machine learning models can achieve deceptively high overall accuracy by simply predicting the common outcome. This imbalance can mask poor performance on the rare event of interest; for example, in a dataset with 0.1% event prevalence, a trivial classifier that predicts "no event" for every case attains ~99.9% accuracy yet fails to detect any true events. To address this, researchers have developed a spectrum of techniques for rare-event prediction, including data-level resampling (oversampling minority cases, undersampling the majority, and synthetic data generation), algorithm-level methods such as cost-sensitive learning and adjusted decision thresholds, and ensemble approaches tailored for imbalance. Evaluating rare-event models also requires special consideration: traditional metrics like overall accuracy are insufficient, and metrics emphasizing the minority class—precision, recall, F1-score, area under the precision-recall curve (AUPRC), and Matthew's correlation coefficient (MCC)—are more informative. This review synthesizes recent advances in rare-event prediction across diverse domains, from healthcare and industrial safety to finance, cybersecurity, and transportation. We discuss the challenges posed by imbalanced safety and health datasets, compare strategies to mitigate class imbalance, examine appropriate evaluation metrics, and highlight case studies in multiple fields. Drawing on best practices from the literature, we propose a unified framework for evaluating rare-event prediction models that can guide machine learning researchers, public health experts, and safety engineers in developing robust, generalizable models for low-frequency yet critical outcomes.*

Keywords: *Rare events, imbalanced data, evaluation metrics, predictive modeling, high-risk systems*

Mujeeb Abdulrazaq
Central South University, School of Traffic and Transportation Engineering, Hunan, People's Republic of China
Email: highyuan7@gmail.com

1.0 Introduction

Rare-event prediction refers to the task of identifying or forecasting events that occur with very low frequency in a dataset (Joshi *et al*., 2002). In an imbalanced dataset, the number of instances of the rare class (such as cases of fraud, highway accidents, and occurrences of diseases) is significantly lesser than that associated with the common class (Joshi *et al*., 2002). Such broader class imbalance can constitute significant challenges for machine learning algorithms because the models may be biased regarding the majority class, with consequences such as sidelining the minority class except special techniques are applied (Johnson & Khoshgoftaar, 2019). However, some exceptional events may be enriched with disproportionate importance despite their infrequency—a rare failure in an industrial system can cause costly downtime, a rare adverse medical event can critically impact patient health, and an uncommon security breach can have outsized consequences. The high stakes and low prevalence of such events mean that standard modeling approaches must

be adapted to avoid misleading results and missed detections.

Numerous terms in the literature overlap with rare-event analysis, including anomaly detection and outlier detection. While related, these concepts have distinctions (Carreño *et al*., 2020). The focus of this review is on prediction of rare events in the context of supervised learning, where rare outcomes are present (albeit scarce) in the data. This challenge has attracted extensive research focus across multiple sectors, ranging from healthcare and public safety to finance, cybersecurity, transportation, and many others. Indeed, researchers have explored specialized solutions at every stage of the machine learning pipeline to tackle rare events, from data preprocessing and augmentation to algorithm design and evaluation metrics.

The rest of this paper presents a detailed survey of techniques and best practices for the prediction of rare events. First, we introduce the challenges caused by class imbalance and why naïve learning strategies often fail on rare events. We then review methods to handle imbalanced data: resampling approaches (oversampling, undersampling, synthetic data generation), cost-sensitive learning, and ensemble methods. Next, we examine evaluation metrics appropriate for rare-event scenarios—highlighting the limitations of accuracy and the use of precision-recall curves, F1 scores, Matthews correlation coefficient, and related measures. We illustrate these concepts with use cases in multiple domains (healthcare, industrial safety, finance, cybersecurity, and transportation), and finally we propose a unified framework for evaluating rare-event prediction models across domains. This review synthesizes recent advances and lessons learned to provide a reference for developing robust models on imbalanced datasets involving critical rare events.

2.0 Challenges with Rare-Event Prediction and Class Imbalance

The class imbalance between the rare class and the majority class creates the fundamental difficulty of modeling rare events. Traditional learning algorithms that optimize for overall error tend to give importance to the abundant class, effectively ignoring the minority class, because those errors contribute only a tiny fraction to the total loss (Leevy *et al*., 2018). In that sense, a model may learn to always predict the majority outcome, still achieving high accuracy, while it fails completely in predicting the events of interest. This leads to deceptively optimistic performance if one naively uses accuracy or other aggregate metrics, as outcome imbalance can distort measures of predictive accuracy in misleading ways (Leevy *et al*., 2018). For instance, standard classifiers often underestimate the probability of rare events (e.g., a logistic regression may output extremely low risk scores for all instances) and yield very low recall on the minority class (Weiss, 2004). In effect, the model performs well on negatives by default but almost never correctly identifies the positives. Such behavior defeats the purpose of rare-event detection.

Data scarcity is another inherent challenge. By definition, there are very few examples of the rare outcome to learn from; hence, it becomes difficult for the model to generalize the underlying patterns. In many realistic situations, rare events are buried in huge volumes of data (He & Cheng, 2021). Consider that a network monitoring system might log millions of events every month but contain just one catastrophic failure or cyber intrusion in that period (He & Cheng, 2021). With this kind of skew, the minority class signals are extremely weak and easily drowned out by the majority. Models can overfit the handful of positive instances or fail to distinguish them at all. Furthermore, the limited information available about the rare class can lead to fundamental issues in model training and validation (Hasanin *et al*., 2020). It becomes hard to validate whether the model truly captures the rare-event pattern or is simply

overfitting noise, since standard cross-validation splits may contain few or no positive instances in some folds.

Another critical challenge is that the consequences of errors are typically asymmetric in rare-event contexts. According to He & Cheng (2021), rare events often align with high-severity outcomes such as severe injury in an accident, critical patient complications, or fraudulent transactions. Therefore, missing a true event (false negative) has far more serious impact than a false alarm (false positive) in many applications (He & Cheng, 2021). Despite this, most machine learning algorithms treat errors equally unless explicitly instructed otherwise. This mismatch can yield models that, while maximizing overall accuracy, perform suboptimally from a decision-making perspective; for example, flagging almost no fraud cases to avoid false alarms at the cost of allowing actual fraud to go undetected. Additionally, rare events may exhibit concept drift and fluctuating patterns over time (e.g., evolving fraud schemes or new failure modes), which further complicates modeling given the paucity of historical examples. All these factors make rare-event prediction a notoriously difficult problem, requiring specialized techniques to ensure that minority cases are properly learned, validated, and detected.

3.0 Methods for Addressing Class Imbalance

3.1 Data-Level Resampling Techniques

One straightforward way to tackle class imbalance is at the data level by resampling the dataset. The aim is to either boost the presence of the minority class (oversampling) or reduce the majority class (undersampling) so that the classifier is trained on a more balanced distribution of events. Random oversampling simply replicates minority instances (Mohammed *et al*., 2020), while random undersampling removes a subset of majority instances. Both basic approaches can improve the balance but have drawbacks: oversampling by duplication can lead to overfitting (the model repeatedly sees the same rare examples and may memorize them), whereas undersampling throws away potentially useful data from the majority class (Seiffert *et al*., 2007).

Researchers have devised advanced sampling techniques to overcome these limitations. A seminal method is the Synthetic Minority Over-sampling Technique (SMOTE), which generates new minority examples by interpolating between existing minority samples (Chawla *et al*., 2002). Instead of simple duplication, SMOTE synthesizes plausible new cases along the line segments joining a minority instance to its nearest neighbors in feature space. This often yields more diverse training examples and has been widely adopted in rare-event contexts; for example, applying SMOTE with logistic regression improved detection of look-alike sound-alike medication errors in imbalanced healthcare data (Zhao *et al*., 2018). Variants of SMOTE target specific situations; Borderline-SMOTE (Han *et al*., 2005), for instance, only oversamples minority instances near the decision boundary. While SMOTE and its variants can bolster the minority class, they are not without issues: the synthetic data may not perfectly reflect the true distribution of minority cases, potentially introducing unrealistic samples or shifting the class decision boundaries.

Another popular extension is Adaptive Synthetic Sampling (ADASYN), which focuses on generating more synthetic examples for minority observations that are harder to learn (He *et al*., 2008). By adaptively concentrating on difficult cases, ADASYN aims to better refine the decision boundary. In practice, oversampling is often paired with data cleaning techniques to mitigate noise. The Edited Nearest Neighbors (ENN) rule removes majority samples that are misclassified by their neighbors, while the Neighborhood Cleaning Method (NCL) extends this by also dropping

certain redundant minority instances (Agustianto & Destarianto, 2019). Hybrid approaches like SMOTE+ENN oversample the minority class and then prune noisy examples from the majority, yielding a balanced and cleaner training set.

Beyond point-wise sampling, more sophisticated strategies consider the data distribution structure. Cluster-based oversampling (CBO) first clusters the data and then performs targeted oversampling within each cluster (Jo & Japkowicz, 2004). This preserves within-class diversity and avoids over-generalizing minority characteristics. In time-series scenarios, time-series subsampling selects representative slices of normal and rare-event sequences to balance the data while maintaining temporal context (Liu *et al*., 2021).

While resampling can dramatically improve a model's ability to learn rare events, it must be applied carefully. Oversampling can amplify minority outliers, whereas undersampling may discard legitimate examples. When used judiciously, resampling is a powerful tool to ensure the model sees enough minority cases during training.

3.2 Cost-Sensitive Learning and Algorithmic Approaches

Instead of altering the data distribution, another approach is to modify the learning algorithm to pay more attention to the minority class. Cost-sensitive learning assigns a higher penalty to misclassifying rare positive instances than common negatives. Most classification algorithms can incorporate class weights or unequal error costs; by giving more importance to the rare class in the cost function, the learner focuses on correctly handling those rare instances (Fernández *et al*., 2018). If a false negative is considered 10 times more costly than a false positive, the algorithm can treat that mistake as 10× worse during optimization, typically improving recall on the minority class.

Machine learning libraries implement class weighting for logistic regression, SVMs, decision trees, and ensembles. King and Zeng's work on rare events in logistic regression provides analytical corrections for probability bias due to class imbalance (King & Zeng, 2001). Specialized loss functions like focal loss dynamically down-weight well-classified examples, forcing the model to concentrate on hard instances, often corresponding to minority samples.

Threshold moving is another technique, where instead of the default 0.5 cutoff, a lower threshold is chosen for the positive class in rare-event settings (Zou *et al*., 2016). This increases sensitivity by capturing more rare events. Threshold selection may be guided by domain requirements. Algorithmic variants also exist: AdaCost increases the weight of misclassified minority examples in boosting, and bagging methods train on balanced subsets. EasyEnsemble trains multiple classifiers on different balanced subsets and then combines predictions (Liu, 2009). These approaches adjust the learning process to reflect that not all errors are equal, improving minority detection without requiring data modification.

3.3 Synthetic Data Generation and Augmentation

When real instances of rare events are limited, synthetic data generation provides a valuable supplement. Rather than resampling existing points, synthetic data explicitly creates new instances through simulation or generative models. Domain-based simulators, such as PaySim for financial fraud (Lopez-Rojas *et al*., 2016), generate realistic datasets with controlled rare-event injection. Autonomous driving systems similarly rely on virtual simulations to create dangerous scenarios that are too rare or unsafe to collect in real life.

Generative Adversarial Networks (GANs) have been used to synthesize minority-class features. Variants such as WGANs and Conditional GANs help generate high-quality targeted examples (Fathy *et al*., 2020). Data augmentation techniques like rotation, cropping, and flipping in imaging, or signal

perturbation in time-series data, help expand small positive datasets. Synthetic data must still be realistic; otherwise, models may learn spurious patterns. Best practices include expert validation and performance testing. When properly implemented, synthetic data generation significantly enhances rare-event learning.

3.4 Ensemble and Specialized Methods

Ensemble learning is particularly effective for rare-event prediction. Combining multiple models exposed to different aspects of the data helps improve minority-class performance. Balanced Random Forests use stratified bootstrap samples with equal rare and common cases. EasyEnsemble trains many weak classifiers on randomly undersampled training subsets and blends their predictions, improving minority detection (Liu, 2009). Boosting methods like SMOTEBoost and RUSBoost integrate sampling within the boosting loop to strengthen weak learners by focusing on minority cases.

Some anomaly detection methods like one-class SVMs and autoencoders model only the normal class and flag deviations as potential rare events. These approaches are useful when positive examples are extremely scarce. However, when labeled rare events are available, supervised methods with imbalance mitigation (resampling, cost sensitivity, ensembles) often perform better (Schapire, 2013). Ensemble techniques remain among the strongest tools for addressing extreme class imbalance.

3.5 Evaluation Metrics for Rare Events

Proper evaluation metrics are crucial when dealing with rare events, as conventional metrics can be misleading. Accuracy – the overall proportion of correct predictions – is notoriously uninformative on imbalanced data. As discussed earlier, a classifier can attain 99% accuracy by simply predicting the majority class in a dataset with 1% rare events, yet such a model has 0% recall on the events of interest (Mortaz, 2020). Thus, accuracy by itself can encourage trivial "all-negative" models that look good on paper but fail in practice. Instead, metrics that emphasize the minority class performance are preferred. Precision and recall are the fundamental rates used to characterize binary classification beyond accuracy. Recall (also called sensitivity or true positive rate) is the fraction of actual rare events that the model successfully identifies. Precision (also called positive predictive value) is the fraction of the model's predicted events that are actually correct. In rare-event contexts, high precision means that when the model flags an event, it is likely to be a true event (few false alarms), and high recall means the model catches most of the actual events (few missed events). There is often a trade-off: one can tune a model to achieve very high recall at the expense of precision (catch all positives but with many false positives), or vice versa. The F1-score combines precision and recall into a single number – it is the harmonic mean of precision and recall. A high F1 requires a balance of both, making it a common metric for imbalanced classification. However, F1 still does not account for the true negatives and may not fully reflect performance on the majority class. It is possible for a classifier to have a decent F1 by doing well on the minority class, yet still issue many false alerts. To address this, more holistic metrics like Matthews Correlation Coefficient (Chicco & Jurman, 2020) or Cohen's kappa are used (Vieira *et al*.). The MCC is essentially a correlation coefficient between predicted and actual labels; it yields a high score only if the model performs well on both the positive and negative classes simultaneously. In other words, MCC will be low if the classifier does poorly on either class, making it very informative for imbalanced scenarios. Indeed, researchers have argued that MCC is a more truthful indicator than accuracy or F1 in skewed data situations (Chicco & Jurman, 2020).

When evaluating rare-event predictors, precision-recall (PR) curves are often more

insightful than the traditional receiver operating characteristic (ROC) curve. An ROC curve plots the true positive rate (recall) against the false positive rate, without regard to class prevalence. While ROC analysis is widely used, its summary statistic (the area under the ROC curve, AUROC) can paint an overly optimistic picture in highly imbalanced settings. This is because the false positive rate (FP/(TN+FP)) may remain very low even when a classifier generates many false alarms, as long as negatives dominate the dataset. By contrast, a PR curve plots precision against recall, directly highlighting performance on the minority class. In extremely skewed data, PR curves tend to give a more faithful view of a model's utility (Juba & Le). For instance, a model might achieve an AUROC of 0.99 by correctly classifying almost all negatives and only a few positives, yet its precision could be abysmal if false positives far outnumber true positives; the AUPRC (area under the PR curve) would reflect this, potentially being very low (since precision is low) despite the high AUROC. In fact, in many low-prevalence scenarios, AUPRC is significantly lower than AUROC, indicating how challenging it is to obtain high precision when events are rare. Consequently, many experts have recommended prioritizing AUPRC over AUROC when comparing models for imbalanced data (Korsunsky *et al*., 2019).

4.0 Cross-Domain Applications and Use Cases

4.1 Healthcare and Public Safety

In healthcare, rare events often correspond to critical but uncommon outcomes – for example, the occurrence of a severe adverse drug reaction, the onset of a rare disease, or an unexpected complication in surgery. Predictive models in medicine frequently confront imbalanced data: an electronic health record database might have thousands of patients without complications for every one patient who experiences a rare event. If not handled properly, the model may simply predict the majority (no complication) for everyone, missing the very cases we care about. Researchers have recognized this and applied the techniques discussed to medical problems. For instance, a study on opioid overdose risk found that standard risk models appeared overly accurate due to class imbalance, with overall accuracy climbing above 95% even as the model's precision on overdose cases plummeted (Cartus *et al*., 2023). The authors highlighted how outcome imbalance can make performance "spuriously high" unless appropriate metrics and rebalancing are used. In another case, to detect look-alike sound-alike (LASA) medication errors – a rare but dangerous mix-up of drug names –(Zhao *et al*., 2018)implemented a rebalancing framework that combined SMOTE oversampling with logistic regression, significantly improving the detection of these incidents in an imbalanced hospital dataset. Across many clinical applications (e.g., predicting rare surgical complications, identifying patients at risk of an infrequent disease flare-up), techniques like oversampling the rare outcomes, using cost-sensitive neural networks, and evaluating with precision-recall curves have become standard practice. The stakes in healthcare are high, so ensuring that predictive models truly capture the minority class (the patients who will have the adverse outcome) is paramount – even if it means accepting more false alarms. By leveraging domain knowledge (for example, augmenting rare case data with expert-derived synthetic examples) and robust evaluation, health informatics researchers are making rare-event prediction tools that clinicians can trust.

Industrial Safety and Engineering

Industrial domains provide many classic examples of rare-event challenges. Predictive maintenance of machinery, for instance, involves forecasting failures or malfunctions before they occur. Yet failures are, by design, infrequent – a factory may record millions of machine readings with only a handful of actual breakdowns. Machine learning models can be

very biased in such scenarios, predicting "no failure" all the time unless guided otherwise. In the power grid or manufacturing settings, it is typical to have data streams where anomalies are "buried in a massive amount of data (He & Cheng, 2021). One may have months of sensor measurements with only one or two fault incidents, making it extremely difficult for a model to learn the signature of a fault. Imbalance countermeasures are routinely employed: engineers use techniques like oversampling historical failure data (or simulating failure conditions) to balance training, or apply one-class classifiers that learn to recognize the normal state and detect any deviation. An example comes from network infrastructure monitoring, where perhaps one critical alarm might occur in one month of logs while all other logs are benign – without resampling or cost weighting, a learned model simply ignores that rare alarm (He & Cheng, 2021). In mechanical systems like engines or turbines, researchers have used GAN-generated synthetic failure signals to augment scarce real failure examples, improving fault detection rates.

Safety incident prediction is another area: consider occupational safety analytics that try to predict workplace accidents based on leading indicators. Accidents are (hopefully) rare compared to normal operation, so such models employ imbalance techniques to avoid always predicting "no accident." They also focus on high-recall solutions – it's better to err on the side of caution (predict a potential accident that doesn't happen) than to miss an impending real accident. Overall, industrial applications of rare-event modeling have embraced the full toolkit: data augmentation (to create virtual examples of failures or defects), cost-sensitive learning (to make "missed detection" errors very costly in the model's loss function), and meticulous evaluation using metrics like recall at a fixed false alarm rate. These approaches have been applied in domains from semiconductor manufacturing (detecting rare product defects) to aviation safety (anticipating rare aircraft component failures) (Dang *et al*., 2022), helping to prevent catastrophic events by learning from relatively few historical incidents.

4.2 Finance and Fraud Detection

Rare-event issues are extremely familiar in finance, particularly in fraud detection and risk modeling. Fraudulent transactions are exceedingly scarce in a sea of legitimate customer activity – often well below 1% of all transactions (Lopez-Rojas *et al*., 2016). For example, a public credit card fraud dataset contains 284,807 transactions of which only 492 are frauds (a mere 0.172%). A model that naively predicts every transaction as "not fraud" will be correct 99.8% of the time, yet fail to catch a single fraud – an unacceptable outcome for a fraud detection system. Accordingly, the financial industry has adopted imbalance learning techniques aggressively. Credit card fraud detection systems commonly use ensemble classifiers trained on balanced samples, or online learning algorithms that adjust to the evolving minority patterns. Oversampling methods (like SMOTE) have been used to synthetically expand the fraud class during training, and cost-sensitive losses ensure the model treats a missed fraud as far worse than a false alarm. The evaluation of such models places heavy emphasis on metrics like precision, recall, and PR curves; an operating point might be chosen to guarantee (for instance) at least 90% recall of frauds while keeping false positive rates within manageable limits to avoid flagging too many legitimate transactions. Beyond credit cards, other financial rare-event cases include insurance claim fraud, where fraudulent claims are a small fraction, and money laundering detection, where illicit transactions are needles in a haystack of banking data. Each presents the same fundamental challenge: the need to detect the few positive cases without being overwhelmed by false positives. Using the tools of imbalanced learning, modern fraud

detection engines have dramatically improved in catching more fraud (sometimes doubling recall) for the same false alarm rate, compared to earlier methods that didn't properly address class imbalance.

4.3 Cybersecurity and Intrusion Detection

Cybersecurity is another field where rare-event detection is critical. Consider an intrusion detection system (IDS) monitoring network traffic or system logs: the vast majority of events are normal behavior, with only a tiny fraction indicating malicious attacks or breaches. The class imbalance here can be extreme – perhaps 1 in 100,000 log entries might be an actual security incident. Without corrective measures, an ML-based IDS will lean toward always predicting "no intrusion", since that yields 99.999% accuracy. To combat this, security researchers utilize techniques like heavy oversampling of known attack instances, synthesizing attack traffic data, and especially cost-sensitive training (to strongly penalize missing an attack). An IDS is typically tuned to achieve high recall of attacks (high detection rate) at a tolerable false positive rate, which mirrors the precision-recall tradeoff discussed earlier. Methods such as one-class anomaly detection are also prevalent: for example, one can train on only normal traffic and then detect outliers as potential intrusions – effectively treating the problem as finding anomalies since labeled attacks may be scarce. This can be useful as an initial line of defense, though it may produce many false positives. In practice, state-of-the-art cybersecurity models combine both supervised imbalanced learning (when known attack examples are available to learn from) and unsupervised anomaly detection for novel threats. The "rare event" nature of serious cyber incidents has even been termed a curse – in the context of autonomous cyber defense or automotive security, the rarity of true attacks makes it hard to validate models, requiring creative solutions (He & Cheng, 2021). Overall, the cybersecurity domain has embraced the imbalanced learning framework: data from past incidents is augmented and weighted, models are ensemble and cost-sensitive, and evaluation focuses on metrics like detection rate (recall) at low false-alarm rates rather than raw accuracy.

4.4 Transportation and Automotive Safety

In transportation safety, rare-event modeling plays a key role in preventing accidents. Most journeys, whether by car, train, or plane, conclude without incident – accidents and critical near-misses are exceedingly rare relative to the enormous number of normal trips or operations. Yet these rare failures are exactly what engineers strive to predict and avoid. An illustrative case is wrong-way driving crashes on highways: they occur very infrequently, but when they do, the consequences are often fatal. A recent study tackled this problem by developing a rare-event model to identify high-risk highway segments for wrong-way crashes, employing data augmentation to compensate for the paucity of historical incidents (Jiang *et al*., 2022). By synthetically boosting the rare crash data and using a specialized modeling framework, the researchers could flag locations with elevated risk, despite only having a handful of observed wrong-way crashes in the dataset. The automotive industry has also encountered the so-called rarity of safety critical events in developing self-driving car technology (Koopman & Wagner, 2017). Autonomous vehicles must be prepared to handle extremely unusual and dangerous scenarios (a child running into the road, a vehicle going the wrong way, etc.), but developers cannot rely on real-world data alone to see enough examples of these events – they are too rare. This has led to extensive use of simulation environments to generate rare traffic situations for training and testing autonomous driving systems, essentially creating synthetic rare events to augment real driving data. The evaluation of advanced driver-assistance and autonomy algorithms similarly focuses on rare high-risk events: metrics like miles per intervention or

collisions avoided are used, and testing protocols deliberately overweight rare hazardous scenarios to ensure the models are truly learning to handle them. In public transportation and air travel, analogous rare-event challenges arise (e.g., predicting a rail derailment or an airplane system failure before it happens). Imbalanced learning techniques have been applied here as well – from using anomaly detection on sensor streams to detect precursors of accidents, to training cost-sensitive classifiers on historical incident databases. By uniting domain-specific knowledge (like physics-based simulations of crashes) with the general principles of rare-event modeling, the transportation field is making strides in proactive safety management.

5.0 Toward a Unified Evaluation Framework

Having surveyed the landscape of rare-event modeling across methods and domains, it is evident that a more unified framework would benefit both research and practice. In the literature, authors have pointed out the lack of standardized approaches for evaluating and comparing rare-event predictors. Too often, studies have employed off-the-shelf metrics and validation procedures suitable for balanced data, which "fail to accurately reflect model performance in rare events" and can lead to misleading conclusions. A unified framework would establish clear guidelines on how to handle class imbalance at each stage – data preprocessing, model training, and performance evaluation – ensuring that critical rare-event signals are not lost in the workflow. Key components of such a framework include: Data Handling: Maintain awareness of class proportions in all steps. Use stratified sampling in cross-validation (so each fold has representative class imbalance) or, if rare events are extremely sparse, use techniques like cross-validation with minority class oversampling to guarantee enough positives in each fold. Encourage the creation and use of benchmark datasets that reflect real-world rarity but are large enough to allow meaningful model development. researchers have noted the need for "standardized datasets" that embody realistic rare-event scenarios (with proper labeling, privacy considerations, and an ideal rarity percentage) to enable fair algorithm comparisons. A unified framework would promote sharing such datasets across domains (e.g., a repository of imbalanced datasets from healthcare, finance, etc.) as well as standardized data augmentation protocols. Model Development: Employ imbalance-mitigation techniques as a default part of the modeling pipeline. This means that for any classification problem with severe class skew, one should automatically consider strategies like oversampling, class weighting, or threshold tuning (Henning *et al*., 2023). Rather than treating imbalance handling as an afterthought, the framework would have it baked in – for example, model development templates could include hooks for resampling the training set or adjusting loss functions. Importantly, the framework would advise choosing methods appropriate to the domain and data size: for small datasets, simple oversampling or cost-sensitive logistic regression might suffice, whereas for large complex data, more advanced methods (like generating synthetic data with GANs or using ensemble learners with adaptive sampling) could be recommended. Recent trends such as ensemble learning and meta-learning for rare events have shown great promise– a unified approach would incorporate these advances, guiding users to combine multiple models or leverage transfer learning to make the most of limited rare-event data. Moreover, the framework would highlight the importance of model calibration under imbalance (ensuring predicted probabilities of rare events are meaningful), as well as techniques for uncertainty quantification so that users know how much to trust a rare-event prediction. Evaluation Protocols: The framework would

mandate evaluation practices that truly measure rare-event performance. This means moving beyond accuracy to a set of metrics that capture different aspects of performance: e.g. always reporting precision, recall, and an aggregate metric like F1 or MCC for the minority class, and comparing models primarily on those measures. It might suggest using AUPRC as a primary metric for model selection in imbalanced cases, given its focus on minority class performance (with the caveat of examining AUROC and other metrics for a full picture). To ensure consistency, the framework can include guidelines like: if class imbalance is above a certain ratio (say 1:100 or more), then avoid using plain accuracy or at least accompany it with precision/recall; or when comparing models, test for statistical significance in differences of recall or other relevant metrics on the rare class. Additionally, a unified approach calls for transparency: publish confusion matrices or detailed error analysis showing how many rare events were missed or falsely flagged, since those counts are crucial in high-impact domains. In short, evaluation should be minority-centric – the framework echoes recent recommendations that new models should be proven on their rare-event detection capability (Cartus *et al*., 2023), and offers "broad guidance" on analytical strategies to mitigate imbalance effects in practice.

Ultimately, the goal of a unified framework is to make rare-event prediction more systematic and comparable across studies. Whether one is predicting failures in an industrial plant or diagnosing a rare disease, the same core principles apply: acknowledge the rarity, compensate for it in model training, and evaluate in a way that reflects real-world costs and benefits. By following a standardized approach, researchers and practitioners can avoid common pitfalls (such as overly optimistic accuracy) and focus on what truly matters – identifying those rare but critical events. As an added benefit, such a framework can facilitate cross-domain learning: techniques proven in one field (say, fraud detection) can be more readily transferred to another field (say, medicine) when a common evaluation language and toolkit is in place. Many open challenges remain, like developing better techniques for extreme rarity (where even augmentation struggles), dealing with concept drift in rare events, and ensuring model interpretability and trust when decisions are based on few examples (Feng *et al*., 2023). However, by converging on a unified set of best practices and evaluation standards, the community can collectively advance the state of rare-event prediction. In the long run, such a framework will help practitioners in safety, health, and other sectors reliably harness machine learning to foresee and prevent the worst-case scenarios – the very instances where these models have the most potential to save lives and resources.

6.0 References

Agustianto, K., & Destarianto, P. (2019). *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),

Carreño, A., Inza, I., & Lozano, J. A. (2020). Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, 53, 5, pp. 3575-3594.

Cartus, A. R., Samuels, E. A., Cerdá, M., & Marshall, B. D. L. (2023). Outcome class imbalance and rare events: An underappreciated complication for overdose risk prediction modeling. Addiction, 118, 6, pp. 1167-1176. https://doi.org/10.1111/add.16133.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling

technique. *Journal of artificial Intelligence Research*, 16, pp. 321-357.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 1, 6, https://doi.org/10.1186/s12864-019-6413-7.

Dang, C., Wei, P., Faes, M. G. R., Valdebenito, M. A., & Beer, M. (2022). Parallel adaptive Bayesian quadrature for rare event estimation. *Reliability Engineering & System Safety*, 225, 108621, doi: 10.1016/j.ress.2022.108621

Fathy, Y., Jaber, M., & Brintrup, A. (2020). Learning with imbalanced data in smart manufacturing: A comparative analysis. IEEE Access, 9, pp. 2734-2757.

Feng, C., Li, L., & Xu, C. (2023). Advancements in predicting and modeling rare event outcomes for enhanced decision-making. *BMC Medical Research Methodology*, *23*, *1*, 243. https://doi.org/10.1186/s12874-023-02060-x

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Cost-sensitive learning. In Learning from imbalanced data sets* (pp. 63-78). Springer.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005, 2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11538059_91.

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2020). Investigating class rarity in big data. *Journal of Big Data*, 7(1), 23, https://doi.org/10.1186/s40537-020-00301-0.

Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN. 2008.4633969.

He, J., & Cheng, M. X. (2021). Weighting methods for rare event identification from imbalanced datasets. *Frontiers in Big Data*, 4, 15 320, https://doi.org/10.3389/fdata.2021.715320.

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). *A survey of methods for addressing class imbalance in deep-learning-based natural language processing*. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 523–540.

Huang, X., Zhang, C.-Z., & Yuan, J. (2020). Predicting extreme financial risks on imbalanced dataset: A combined kernel FCM and kernel SMOTE based SVM classifier. *Computational Economics*, 56, 1, pp. 187-216.

Jiang, L., Xie, Y., Wen, X., & Ren, T. (2022). Modeling highly imbalanced crash severity data by ensemble methods and global sensitivity analysis. *Journal of Transportation Safety & Security*, 14, 4, pp. 562-584.

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, 6, 1, pp. 40-49.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of big data, 6, 1, pp. 1-54.

Joshi, M. V., Agarwal, R. C., & Kumar, V. (2002, 2002). Predicting rare classes: Can boosting make any weak learner strong?. https://sci2s.ugr.es/keel/pdf/specific/congreso/kdd-MaheshJoshi.pdf.

Juba, B., & Le, H. S. (2019). Precision-recall versus accuracy and the role of large data sets. https://doi.org/10.1609/aaai.v33i01.33014039

King, G., & Zeng, L. (2001). Logistic regression in rare events data. Political Analysis, 9, 2, pp. 137-163.

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, 9, 1, pp. 90-96.

Korsunsky, I., Nathan, A., Millard, N., & Raychaudhuri, S. (2019). Presto scales Wilcoxon and auROC analyses to millions of observations. BioRxiv, 653253.

Kumar, V., Lalotra, G. S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., Shorfuzzaman, M., Alsufyani, A., & Uddin, M. (2022). Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. *Healthcare*, *10*(7), 1293. https://doi.org/10.3390/healthcare1007129.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5, 1, pp. 1-30.

Liu, H., Ma, R., Li, D., Yan, L., & Ma, Z. (2021). Machinery fault diagnosis based on deep learning for time series analysis and knowledge graphs. *Journal of Signal Processing Systems*, 93, 12, pp. 1433-1455.

Liu, T.-Y. (2009, 2009). *Easyensemble and feature selection for imbalance data sets*. 2*009* International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, *Shanghai, China, 2009, pp. 517-520, doi: 10.1109/IJCBS.2009.22*.

Lopez-Rojas, E., Elmir, A., & Axelsson, S. (2016, 2016). PaySim: A financial mobile money simulator for fraud detection. Proceedings of the European Modeling and Simulation Symposium, 2016 978-88-97999-76-8; Bruzzone, Jiménez, Longo, Louca and Zhang Eds, pp. 249-255.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020, 2020). Machine learning with oversampling and undersampling techniques: overview study and experimental results. 2020 11th International Conference on Information and Communication Systems (ICICS), *Irbid, Jordan, 2020, pp. 243-248, doi: 10.1109/ICIC S49469.2020.239556.*

Mortaz, E. (2020). Imbalance accuracy metric for model selection in multi-class imbalance classification problems. *Knowledge-Based Systems*, 210, 106490, https://doi.org/10.1016/j.knosys.2020.106490.

Schapire, R. E. (2013). *Explaining adaboost. In Empirical inference: festschrift in honor of vladimir N. Vapnik* (pp. 37-52). Springer.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2007, 2007). Mining data with rare events: a case study. 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007), Patras, Greece, 2007, pp. 132-139, doi: 10.1109/ICTAI.2007.71.

Vieira, S. M., Kaymak, U., & Sousa, J. M. C. (2010). *Cohen's kappa coefficient as a performance measure for feature selection*. International Conference on Fuzzy *Systems*, Barcelona, Spain, 2010, pp. 1-8, doi: 10.1109/FUZZY.2010.5584447.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6, 1, pp. 7-19.

Zhao, Y., Wong, Z. S.-Y., & Tsui, K. L. (2018). A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering*, 2018(1), 6275435, doi: 10.1155/2018/6275435.

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5, pp. 2-8.

Compliance with Ethical Standards

Declarations

The authors declare that they have no conflict of interest.

Data availability

All data used in this study will be readily available to the public.

Consent for publication

Not Applicable.

Availability of data and materials

The publisher has the right to make the data public.

Competing interests

The authors declared no conflict of interest.

Funding

The authors declared no source of funding

Authors' Contributions

The author designed and carried out the entire work