International Journal of Safety and Security Engineering
Vol. 16, No. 6, June, 2026, pp. 1257-1267
Journal homepage: http://iieta.org/journals/ijsse

Beyond SMOTE: Establishing a Theoretical Benchmark for Generative Adversarial Data Augmentation in E-Commerce Fraud Detection

Sarmini[1,2*], Sunardi[3], Abdul Fadlil[3]

[1] Department of Informatics, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia
[2] Department of Information System, Universitas Amikom Purwokerto, Purwokerto 53127, Indonesia
[3] Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia

Corresponding Author Email: 2437083013@webmail.uad.ac.id

https://doi.org/10.18280/ijsse.160608

Received: 1 April 2026
Revised: 30 May 2026
Accepted: 6 June 2026
Available online: 30 June 2026

Keywords:
fraud detection, data augmentation, class imbalance, Generative Adversarial Networks, e-commerce transactions, theoretical benchmark

ABSTRACT

Detecting e-commerce fraud is quite hard because there aren't enough fraudulent transactions, and machine learning has a problem with class imbalance. However, there is a lack of extensive benchmarks that evaluate the theoretical effectiveness of various data augmentation methods, including Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Oversampling Technique (SMOTE), in controlled settings. This study contrasts traditional oversampling methodologies with advanced generative models, such as Conditional Generative Adversarial Network (CGAN) and K-CGAN, to provide a performance benchmark. Two e-commerce fraud datasets were meticulously supplemented to create balanced distributions, facilitating the assessment of three classifiers: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM), which were trained solely on balanced data. The K-CGAN model achieved near-perfect F1-scores of approximately 0.99, indicating that generative methods substantially outperform traditional oversampling, as confirmed by experimental results. The stability and performance of the K-CGAN and RF combination were consistently observed to be superior. The present study establishes a novel theoretical performance benchmark, providing a practical framework for the creation of data-driven, robust fraud detection systems in e-commerce.

1. INTRODUCTION

Modern business has been profoundly transformed by the rapid global expansion of e-commerce, which provides unparalleled convenience and accessibility to customers worldwide [1]. The digital revolution has coincided with a troubling rise in the volume and complexity of fraudulent activities [2, 3]. Contemporary cybercriminals utilize increasingly advanced tactics, evolving from simple fraud to conducting intricate phishing operations and orchestrated account takeovers, often leveraging artificial intelligence to detect flaws and bypass conventional security protocols [4]. The consequences of unrecognized fraud for businesses and financial institutions are substantial and intricate. They include not only direct financial losses but also significant operational costs for investigation and recovery, as well as the intangible yet profound erosion of customer trust, leading to lasting reputational damage to their brands and potentially resulting in customer attrition and diminished market standing [5, 6].

The fundamental machine learning challenge of severe class imbalance is at the heart of the development of effective fraud detection systems. Fraudulent transactions are inherently uncommon, resulting in a negligible fraction of transactions in a given dataset, frequently less than one percent [7, 8]. The majority class is inherently biased by the underlying loss functions of standard machine learning models, which are typically optimized to maximize overall accuracy on balanced data [9]. Consequently, they are ill-equipped to address this reality. A model can obtain a deceptively high accuracy score of over 99% by classifying all transactions as legitimate in such imbalanced contexts. This results in a critical failure point, where the model appears to be successful on paper but is completely ineffectual in practice, indicating a critically low detection rate, or recall, for the minority class (fraud). The model's inability to identify authentic fraudulent instances renders it unsuitable for its primary purpose [10, 11].

The establishment of numerous data augmentation strategies has alleviated the class imbalance. These techniques range from conventional oversampling methods, such as the Synthetic Minority Oversampling Technique (SMOTE), which interpolates between existing instances to generate new ones, to advanced generative models, such as Conditional Generative Adversarial Networks (CGANs), which learn the underlying data distribution to produce novel, high-fidelity samples [12, 13]. Despite the existence of these tools, there is a substantial gap in the literature: the absence of exhaustive, direct comparative studies that assess their efficacy in a controlled, optimal environment where external influences are minimized [14, 15]. In order to comprehend the true potential of traditional oversampling and advanced generative methods

when applied to a stable set of powerful, industry-standard classifiers such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM), a direct, systematic comparison is required. It is imperative to conduct this type of analysis in order to ascertain the theoretical maximum performance that can be achieved under ideal conditions, thereby establishing a distinct benchmark against which other models can be evaluated [16, 17].

This paper aims to fill this critical gap by establishing a definitive performance benchmark for data augmentation in e-commerce fraud detection. Our primary objectives are to systematically evaluate the performance of six distinct data augmentation strategies across two different e-commerce datasets; to identify the optimal pairing of an augmentation strategy with one of the three core classifiers; and to establish a clear and robust performance benchmark under balanced data conditions. The primary contribution of this work is; therefore, a benchmark that clarifies which augmentation strategy delivers the best results for the most common and powerful classifiers used in the industry. By doing so, this research provides a crucial and actionable reference for both academic researchers developing new techniques and industry practitioners striving to design and implement more effective fraud detection systems.

2. LITERATURE REVIEW

2.1 Machine learning in fraud detection

The application of computational intelligence in fraud detection has witnessed a substantial transformation, transitioning from static, rule-based systems to dynamic, adaptive machine learning models. Rule-based systems, which operate on a set of predefined, human-authored heuristics to identify suspicious transactions, were historically the primary method of fraud detection. These systems are inherently constrained by their rigidity, despite their simplicity of implementation. They are unable to learn from new data and struggle to adapt to the novel and evolving tactics employed by fraudsters, necessitating continual manual updates by domain experts to maintain relevance [18]. The reactive approach inevitably results in a higher rate of false negatives, which is the result of sophisticated fraud schemes going undetected. Consequently, these legacy systems become increasingly insufficient in the face of a dynamic threat landscape [19].

The fundamental weaknesses of rule-based approaches have been addressed by the paradigm shift toward machine learning, which has introduced the capacity to learn complex patterns directly from historical data. The new standard was supervised learning models, which were trained to classify transactions by identifying the subtle, non-linear relationships that often characterize fraudulent behavior [19, 20]. These models included SVM, RF, and various neural network architectures. ML models are capable of processing vast volumes of transactional data and adapting over time, thereby perpetually refining their understanding of what constitutes a fraudulent event, in contrast to their predecessors. This data-driven methodology enables them to identify complex fraud patterns that would be virtually impossible to codify with explicit rules, thereby demonstrating a paradigm-shifting improvement in their detection capabilities [21].

The adoption of more sophisticated techniques, such as deep learning and ensemble methods, has been facilitated by further advancements, which have demonstrated superior performance. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are deep learning models that are particularly adept at capturing complex contextual features and temporal dependencies within sequences of transactional data. Consequently, they provide a more detailed understanding of user behavior [7]. Simultaneously, ensemble techniques have been popular due to their capacity to improve the predicted accuracy and resilience of various individual models, thus effectively addressing the shortcomings of any single classifier [18, 22]. The current state-of-the-art is represented by these advanced models, which are stretching the boundaries of detection accuracy and resilience.

The need for continuous adaptation and the interpretability of models remain persistent challenges, despite these significant advancements. In the financial sector, where transparency in decision-making is crucial, the opaque character of many complicated models can hinder their adoption [23, 24]. Additionally, the continuous retraining of those models on new data is necessary to maintain their effectiveness due to the constant evolution of fraud strategies, a phenomenon known as concept drift [25]. This persistent arms race between security specialists and fraudsters underscores the critical nature of developing systems that are both adaptable and resilient. Our challenges serve as the impetus for our research on data augmentation to enhance model training and performance.

2.2 The impact of imbalanced data

The primary obstacle in utilizing machine learning for fraud detection is the intrinsic class imbalance within the data. This characteristic renders conventional performance metrics, such as accuracy, profoundly deceptive. In a dataset where fraudulent transactions may constitute less than 1% of the total volume, a model can obtain a score exceeding 99% by defaulting to a "non-fraudulent" prediction for every instance, as accuracy measures the overall proportion of correct predictions [26]. Although it may appear impressive, the model's complete failure to fulfill its primary function of identifying the uncommon but critically important fraud cases is concealed by this high level of accuracy. This phenomenon arises from traditional classification algorithms' bias towards the majority class, resulting in a significant incidence of false negatives (overlooked fraud), which can have devastating financial and reputational repercussions for a company [27, 28].

In order to conduct a thorough analysis of the minority (fraud) class, it is imperative to shift the emphasis from overall accuracy to measures that explicitly assess a model's efficacy. The two most essential components are precision and recall. Precision quantifies the ratio of transactions identified as fraudulent that are indeed fraudulent. High precision is essential for reducing false positives, which can disrupt consumers and increase operating costs [21]. Recall, or sensitivity, quantifies the ratio of real fraudulent transactions that the model accurately detects. The fundamental objective for loss prevention is high recall, since it directly pertains to the model's capacity to identify and prevent fraud [29, 30]. Because precision and recall often exist in a trade-off which is improving one can negatively impact the other the F1-score is frequently employed as a more holistic measure. As the harmonic mean of precision and recall, the F1-score provides

a single value that balances the two, making it a particularly robust metric for comparing models in imbalanced scenarios [31, 32].

Selecting appropriate evaluation criteria is a crucial step in developing an effective fraud detection system. The implementation of ineffective models and a misleading sense of security were the consequences of relying on accuracy in an imbalanced domain. The model is refined to accomplish the unique business objectives of reducing financial loss while considering the customer effect through a meticulous assessment that focuses on precision, recall, and the F1-score. This focused emphasis on minority-class performance is crucial for accurately evaluating the genuine efficacy of any fraud detection technology and serves as a guiding principle for the experimental methodology outlined in this research.

2.3 Data augmentation techniques

In order to mitigate the performance degradation that arises from class imbalance, a variety of data augmentation strategies have been developed to rebalance datasets prior to model training. Oversampling techniques are among the most widely recognized methodologies, as they aim to improve the representation of the minority class. SMOTE is a fundamental technique that generates new, synthetic instances rather than merely replicating existing ones. The SMOTE algorithm operates by selecting a minority class instance, identifying its k-nearest minority class neighbors, and subsequently generating a synthetic sample at a random location along the line segment that connects the instance to one of its neighbors [33]. SMOTE often enhances model performance on metrics like recall and F1-score by expanding and refining the decision boundaries of the minority class through the creation of interpolated instances [34]. This technique fails to account for the proximity of majority class occurrences, potentially resulting in chaotic samples in overlapping class regions and increasing the risk of overfitting [35].

The Adaptive Synthetic Sampling (ADASYN) method, founded on SMOTE principles, employs an adaptive weighting system to focus the data creation process on minority cases that are more difficult to learn. ADASYN begins by calculating the proportion of majority class instances surrounding each minority sample. This ratio serves as a proxy for classification difficulty, with a higher ratio indicating that the instance is more challenging to learn when positioned near the class boundary [36]. Consequently, the algorithm produces a disproportionately larger number of synthetic samples for these more complex instances. This focused methodology necessitates the learning algorithm to concentrate on complex decision boundaries, potentially leading to improved model generalization and more efficient identification of challenging fraud situations [37]. While often more efficient than traditional SMOTE, ADASYN can be computationally demanding and may still generate noise if the challenging samples are oddities.

Generative Adversarial Networks (GANs) are the most advanced example of a more sophisticated class of generative methods that have emerged in conjunction with the refinement of oversampling techniques. A GAN framework is comprised of two neural networks, a Generator and a Discriminator, that are involved in a zero-sum, adversarial game. The Generator generates synthetic data from random noise, while the Discriminator's goal is to differentiate this synthetic data from real data [38]. The competitive training procedure enhances the Generator's capacity to produce high-fidelity synthetic samples that closely resemble the actual data distribution. The CGAN has substantially improved this design. It provides a conditional vector (e.g., a class designation such as "fraud") as input to both the Generator and Discriminator. This enables the creation of synthetic samples that belong to a specific target class, a crucial capability for imbalanced classification tasks, and allows for direct control over the data generation process [39, 40].

The K-CGAN enhances the CGAN architecture by integrating Kullback-Leibler (KL) Divergence into its loss function. KL Divergence is a statistical metric that quantifies the divergence between one probability distribution and a second, reference distribution. In the framework of K-CGAN, this alteration compels the Generator to generate not only realistic samples capable of deceiving the Discriminator but also to guarantee that the statistical distribution of the generated samples closely aligns with the distribution of the actual data for that particular class [41]. This supplementary restriction promotes the quality and diversity of synthetic data, improves training stability, and alleviates prevalent GAN challenges such as mode collapse, when the generator yields a restricted range of samples [42]. These advanced generative models exemplify the pinnacle of synthetic data creation and provide a theoretically superior alternative to conventional oversampling techniques for mitigating significant class imbalance.

2.4 Evaluated classifiers

The data utilized for training classification algorithms is as crucial as the choice of suitable methods. This study examines three highly effective, industry-standard classifiers: RF, XGBoost, and SVM. Each classifier was selected for its unique theoretical foundations and proven effectiveness in the fraud detection literature. RF is an ensemble method based on bagging that generates a significant number of decision trees during the training phase. At each node, only a random selection of features is assessed for the split, and each tree is generated using a bootstrapped sample of the data. The trees are effectively decorrelated by this dual randomization strategy, and the final prediction is determined by a majority vote. This mechanism significantly reduces variance and protects against overfitting, making it highly robust in chaotic data environments that are typical of fraud detection [43, 44].

In contrast to the parallel tree construction of bagging, XGBoost is a sequential, gradient-boosting ensemble method. XGBoost constructs a sequence of decision trees, with each subsequent tree trained to rectify the residual faults of its predecessors. Its great performance arises from a refined optimization of the objective function, incorporating both a loss function and a regularization term, and it distinctively employs second-order derivatives (the Hessian) to attain accelerated convergence [45]. XGBoost's efficiency, combined with its ability to handle missing values and apply regularization to reduce overfitting, has established it as a leading classifier for structured data, with numerous studies demonstrating its remarkable accuracy and effectiveness in tackling the pronounced class imbalance typical of fraud datasets [46, 47].

SVM operate on a distinct principle: kernel-based margin maximization. The aim of an SVM is to identify an ideal hyperplane that most effectively distinguishes data points of disparate classes in a high-dimensional feature space. The

"optimal" hyperplane is characterized by its ability to optimize the margin, or distance, between the closest data points of each class, known as the support vectors. In order to achieve linear separation for non-linearly separable data, SVM employs the "kernel trick," which involves the implicit conversion of the data into a higher dimension using functions such as polynomials and radial basis functions [48]. Its ability to effectively detect fraud in high-dimensional spaces and its focus on the most relevant data points for delineating the class border make it a formidable tool, particularly in situations where transaction data is sparse and high-dimensional [49, 50].

These three classifiers exemplify unique and robust machine learning principles: RF employs diversity through bagging, SVM focuses on creating appropriate separation boundaries in intricate, high-dimensional spaces, and XGBoost enhances performance through iterative error correction. The extensive implementation and rigorously confirmed efficacy of these strategies in fraud detection support this research. By evaluating data augmentation strategies across a diverse array of potent models, the observed performance improvements can be attributed to the quality of the synthetic data, rather than the peculiarities of an individual classification algorithm. This is achieved by establishing a robust and generalizable benchmark.

3. METHODOLOGY

The methodology of this study is meticulously crafted to evaluate a diverse array of data augmentation strategies and classification algorithms in a controlled, balanced data environment, thereby establishing a precise performance standard. This approach is innovative in that it deliberately eliminates the confounding variable of model generalization from biased testing data to augmented training data. By doing so, it isolates and directly assesses the quality of the synthetic data produced by each augmentation technique and establishes a theoretical "upper bound" for performance. This provides a clear standard against which future; more complex generalization strategies can be measured.

3.1 Datasets and experimental protocol

Two e-commerce transaction datasets that were publicly accessible, well-documented, and distinct were utilized to ensure the generalizability and robustness of the findings. The selection of two datasets with varying dimensions, feature sets, and degrees of fraud prevalence further rigorously validates the benchmark results. The dataset, which was obtained from a Kaggle competition, comprises 23,634 transaction records. This dataset is a classic example of the rare-event detection challenge, as it is unique in that it exhibits a severe class imbalance, with fraudulent transactions accounting for a negligible percentage of the total instances. Its primary attributes are anonymized transactional variables. The second dataset is a more intricate and extensive real-world dataset from a multinational e-commerce company, which was initially described by Zeng et al. [51]. It comprises 37 exhaustive features, including transactional, behavioral, and relational attributes, and 297,715 transaction records. Additionally, this dataset demonstrates a substantial class imbalance, with fraudulent transactions accounting for only 3.1% of the total volume. As a result, it provides a more feature-rich and distinctive environment for the assessment of augmentation techniques.

The complete original dataset was initially enhanced for each trial to provide a substantial, completely balanced dataset with an optimal 50/50 class distribution. This approach establishes a synthetic data environment where fraud becomes a frequent occurrence, enabling classifiers to discern its patterns devoid of the underlying bias present in the original distribution. The newly created balanced dataset was divided into two segments: training (80%) and testing (20%). A stratified division method was devised to ensure accurate maintenance of a 50/50 class balance in both portions. This stage is crucial, as it ensures that the evaluation accurately depicts the model's ability to identify patterns within the augmented data, rather than focusing solely on its capacity to manage class imbalance during prediction process. A comprehensive comparison study was conducted using 36 unique trial combinations obtained from this controlled methodology. Three classifiers, six augmentation methods, and two datasets were included in each configuration.

3.2 Augmentation and classification models

Six distinct data configurations were evaluated to provide a comprehensive comparison, spanning from a baseline to traditional and advanced generative methods. The techniques implemented included: the Original imbalanced dataset to serve as an unmitigated performance baseline; two traditional oversampling methods, SMOTE which creates synthetic samples via interpolation, and ADASYN, which focuses generation on harder-to-learn examples and three advanced generative approaches. These were CGAN, which learns the underlying data distribution to generate novel samples, a modified K-CGAN incorporating KL-Divergence loss for improved training stability and data quality, and a hybrid K-CGAN+SMOTE technique designed to explore potential synergies between generative and interpolative methods.

Three powerful and widely adopted classification algorithms were selected to form the basis of the benchmark. The selection of these three guarantees that the results are not prejudicial to a single algorithmic paradigm, as they represent distinct machine learning philosophies: bagging, boosting, and margin maximization. The classifiers were: RF, a bagging-based ensemble known for its robustness to noise. XGBoost, a state-of-the-art gradient-boosting ensemble celebrated for its high performance, for which GPU acceleration was utilized to ensure computational efficiency, and SVM, a kernel-based classifier highly effective in high-dimensional feature spaces. For all classifiers, hyperparameters were systematically optimized for each data configuration using grid search with cross-validation on the training set to ensure fair and optimal performance comparison.

3.3 Evaluation metrics and model selection

In order to evaluate the efficacy of each model, the standard classification criteria were implemented, with the primary objective of fraud detection being the consistent identification of the minority (fraud) class. A simplistic model can achieve high accuracy by classifying all cases as the predominant class, despite the fact that accuracy, which is defined as the ratio of correct predictions to total predictions, was calculated. Nevertheless, it is perceived as a potentially misleading metric in this context. The precision for the fraud class quantifies the

ratio of transactions identified as fraudulent that are authentically fraudulent. Enhancing recollection is essential for reducing direct financial losses from unrecognized fraudulent acts.

The F1-Score for the fraud class, being the harmonic mean of precision and recall, offers a singular, balanced assessment of performance. It was chosen as the principal indicator for selecting champion pairings because it embodies the essential trade-off between the costs of false positives and false negatives. Ultimately, Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) was employed to assess the model's overall, threshold-independent capacity to differentiate between the positive and negative classes, providing a comprehensive perspective on its discriminatory efficacy.

4. RESULT AND DISCUSSION

This section clarifies and examines the empirical results obtained from the two-phase experimental methodology. Phase 1 encompassed a comprehensive comparison analysis of six distinct augmentation approaches in conjunction with three high-performance classifiers, applied to two different e-commerce datasets. The objective was to determine the theoretical performance limit in an idealized setting. Phase 2 executed a stringent robustness assessment on the best combinations found in Phase 1 to confirm their stability. The following discussion will contextualize these data, offering a comprehensive analysis of the performance disparity between conventional oversampling and sophisticated generative augmentation techniques within this controlled, balanced setting.

4.1 Baseline performance of classifiers on imbalanced data

The initial investigations, which were conducted on the original and unaltered imbalanced datasets, were crucial in the establishment of a critical performance baseline. The results unequivocally confirm the well-documented challenge of applying standard classifiers to severely distorted data, as detailed in the "Original" entries of Tables 1 and 2. In an imbalanced scenario, the most straightforward path to achieving this objective is to predict the majority class, as the models are designed to minimize overall error. This inherent bias led to an exceptionally subpar performance in the primary task of fraud detection.

The Kaggle dataset's fraud class yielded the highest F1-Score for XGBoost, which was a mere 0.2200. This low score was a direct result of a critically low recall rate, which suggests that the model was unable to identify the overwhelming majority of fraudulent transactions. The performance on the larger, more feature-rich NNEnsLeG dataset was similarly inadequate, with XGBoost once again leading with a slightly improved but still operationally insufficient F1-Score of 0.3547. The results of this study underscore the fact that a fraud detection system that is not designed to address class imbalance would be functionally ineffective. This would allow the majority of fraudulent activity to continue undetected, while also creating a false sense of security through superficially high accuracy scores (e.g., approximately 97% on the NNEnsLeG dataset). The model's success in accurately identifying the overwhelmingly prevalent non-fraudulent transactions was solely demonstrated by this deceptive accuracy, rather than its capacity to perform its core security function.

4.2 Performance enhancement with oversampling techniques

The model's performance was significantly enhanced by the application of traditional oversampling approaches, including ADASYN and SMOTE. These methods effectively rebalanced the class distribution by synthetically generating new minority class instances through interpolation, thereby requiring the classifiers' decision boundaries to become more sensitive to the features indicative of deception. The F1-Score was significantly increased by fourfold over the baseline when SMOTE and XGBoost were combined, as illustrated in Table 1 for the Kaggle dataset. This illustrates that even relatively straightforward data augmentation can enable a classifier to learn from the minority class.

Table 1. Phase 1 performance metrics on Kaggle dataset

Augmentation Method	Classifier	Precision	Recall	F1-Score	ROC-AUC	Accuracy
Original	RF	0.1118	0.5925	0.1882	0.6517	0.9522
Original	SVM	0.0529	0.5404	0.0963	0.5283	0.9526
Original	XGBoost	0.1340	0.6123	0.2200	0.6811	0.9503
SMOTE	RF	0.9026	0.9264	0.9143	0.9741	0.9132
SMOTE	SVM	0.7776	0.6734	0.7218	0.7978	0.7404
SMOTE	XGBoost	0.9525	0.8858	0.9179	0.9735	0.9208
ADASYN	RF	0.8986	0.9251	0.9117	0.9731	0.9054
ADASYN	SVM	0.7758	0.6713	0.7198	0.7952	0.7181
ADASYN	XGBoost	0.9483	0.8841	0.9151	0.9723	0.9116
CGAN	RF	0.9842	0.9817	0.9829	0.9934	0.9746
CGAN	SVM	0.9393	0.9713	0.9550	0.9859	0.9760
CGAN	XGBoost	0.9827	0.9775	0.9801	0.9926	0.9743
K-CGAN	RF	0.9965	0.9941	0.9953	0.9990	0.9747
K-CGAN	SVM	0.9490	0.9792	0.9639	0.9913	0.9760
K-CGAN	XGBoost	0.9956	0.9907	0.9931	0.9982	0.9743
K-CGAN+SMOTE	RF	0.9967	0.9930	0.9949	0.9989	0.9740
K-CGAN+SMOTE	SVM	0.9470	0.9775	0.9620	0.9907	0.9659
K-CGAN+SMOTE	XGBoost	0.9940	0.9892	0.9916	0.9982	0.9740

Note: RF = Random Forest; SVM = Support Vector Machines; XGBoost = Extreme Gradient Boosting; Synthetic Minority Oversampling Technique (SMOTE); Adaptive Synthetic Sampling (ADASYN); Conditional Generative Adversarial Network (CGAN); Receiver Operating Characteristic–Area Under the Curve (ROC-AUC)

Table 2. Phase 1 performance metrics on NNEnsLeG dataset

Augmentation Method	Classifier	Precision	Recall	F1-Score	ROC-AUC	Accuracy
Original	RF	0.8211	0.1939	0.3138	0.8999	0.9735
Original	SVM	0.5000	0.0001	0.0002	0.5000	0.9734
Original	XGBoost	0.7487	0.2324	0.3547	0.9025	0.9738
SMOTE	RF	0.1692	0.7752	0.2777	0.8856	0.8552
SMOTE	SVM	0.7812	0.7998	0.7904	0.8654	0.8243
SMOTE	XGBoost	0.8742	0.8224	0.8475	0.9298	0.8478
ADASYN	RF	0.1685	0.7781	0.2771	0.8849	0.8271
ADASYN	SVM	0.7754	0.7912	0.7832	0.8601	0.7971
ADASYN	XGBoost	0.8701	0.8211	0.8449	0.9287	0.8192
CGAN	RF	0.9389	0.9693	0.9539	0.9682	0.9863
CGAN	SVM	0.9150	0.9281	0.9215	0.9555	0.9847
CGAN	XGBoost	0.9385	0.9694	0.9537	0.9701	0.9864
K-CGAN	RF	0.9968	0.9758	0.9862	0.9968	0.9863
K-CGAN	SVM	0.9587	0.9715	0.9651	0.9892	0.9857
K-CGAN	XGBoost	0.9982	0.9744	0.9862	0.9968	0.9864
K-CGAN+SMOTE	RF	0.9967	0.9758	0.9861	0.9968	0.9863
K-CGAN+SMOTE	SVM	0.9579	0.9712	0.9645	0.9889	0.9851
K-CGAN+SMOTE	XGBoost	0.9981	0.9742	0.9860	0.9967	0.9864

Note: RF = Random Forest; SVM = Support Vector Machines; XGBoost = Extreme Gradient Boosting; Synthetic Minority Oversampling Technique (SMOTE); Adaptive Synthetic Sampling (ADASYN); Conditional Generative Adversarial Network (CGAN); Receiver Operating Characteristic–Area Under the Curve (ROC-AUC)

On the NNEnsLeG dataset (Table 2), the results were also compelling; the F1-Score for XGBoost combined with SMOTE rose to a respectable 0.8475. This finding suggests that for large datasets with well-defined features, interpolation-based methods can provide enough minority class information to drastically improve learning and detection capabilities. The efficacy of ADASYN was consistently similar to that of SMOTE, albeit generally somewhat worse, suggesting that its adaptive emphasis on challenging examples did not confer a substantial benefit in these particular data contexts. Although these methods represent a significant advancement from the baseline, they remain inferior to the nearly flawless scores attained by generative techniques, possibly due to the inherent constraints of interpolation, which may introduce noise or inadequately represent the intricate distribution of the minority class.

4.3 Achieving state-of-the-art performance with generative augmentation

Significant speed improvements, setting a new theoretical standard, were achieved with the implementation of advanced generative augmentation techniques. The application of CGAN, K-CGAN, and the hybrid K-CGAN+SMOTE elevated the classifiers to nearly optimal performance levels. In contrast to interpolation, which produces new samples among old ones, these generative models comprehend the fundamental probability distribution of fraudulent data, allowing them to construct wholly original and very realistic synthetic transactions.

Table 1 delineates the comprehensive performance metrics of the Kaggle dataset, contrasting various augmentation procedures and classifiers in the initial experimental phase.

Table 1 displays the classification results for the Kaggle dataset, utilizing several classifiers and augmentation strategies. The comparison clearly shows that the model's performance enhances with the implementation of more advanced data augmentation techniques.

The baseline results derived from the Original (unaugmented) data reveal substantial class imbalance effects. The recall and F1-scores of all classifiers are exceedingly low, with the RF achieving a mere 0.1882 and the SVM achieving a mere 0.0963. Although the accuracy remains high (~0.95), this is due to the fact that the majority of the predictions are for the majority class, rather than the model's ability to distinguish between classes.

Substantial enhancements are observed when conventional oversampling methods, including ADASYN and SMOTE, are implemented. Both methods confirm that synthetic balancing mitigates bias toward the majority class by increasing F1-scores above 0.90 for tree-based models. The most balanced performance is achieved by SMOTE–RF and SMOTE–XGBoost, with F1-scores of 0.9143 and 0.9179, respectively, and ROC-AUC values exceeding 0.97. The results of ADASYN are comparable, suggesting that both interpolation-based methodologies effectively improve minority representation.

Generative augmentation methods (CGAN, K-CGAN, and K-CGAN+SMOTE) yield the most remarkable results. The synthetic samples generated by these models are highly realistic, as they learn the underlying data distribution rather than linear interpolation. This enables classifiers to generalize nearly flawlessly. The K-CGAN + RF combination reaches an F1-score of 0.9953, marking an almost flawless classification outcome. This performance signifies that the synthetic data generated by K-CGAN provides exceptional fidelity and diversity, allowing the RF classifier to precisely delineate the decision boundary between fraudulent and legitimate transactions.

Additionally, K-CGAN consistently outperforms standard CGAN, indicating that the integration of KL Divergence loss improves stability and reduces the risk of mode collapse. The hybrid K-CGAN + SMOTE configuration exhibits comparable performance, indicating that the generative model is the primary source of performance improvements, despite the fact that interpolation may still provide minor smoothing benefits. All of these findings collectively establish a new theoretical upper limit for F1-score performance on the Kaggle dataset under balanced, idealized conditions.

The same experimental protocol was replicated using the NNEnsLeG dataset to confirm whether the observed performance trends generalize to more complex and heterogeneous data. Table 2 summarizes the performance metrics that correspond to them.

Table 2 presents the performance results on the NNEnsLeG dataset, which exhibits greater complexity and a higher level of feature heterogeneity compared to the Kaggle dataset. The proposed augmentation strategies' robustness is further bolstered by the consistent nature of the observed trends, despite this additional challenge.

Under the original data conditions, the classifiers demonstrate a high overall accuracy (~0.97) but exceedingly poor F1-scores as a result of severe imbalance. The SVM's F1-score of 0.0002 underscores its inadequacy in identifying fraudulent cases. This mismatch is effectively rectified by SMOTE and ADASYN, which substantially improve the recall values of all models. Nonetheless, their interpolated samples fail to encapsulate complex fraud patterns, yielding moderate F1-scores of roughly 0.84 for the ideal XGBoost configuration.

The introduction of CGAN and K-CGAN has substantially improved the performance. The CGAN–RF combination achieves an F1-score of 0.9539, while K-CGAN–RF achieves 0.9862, surpassing the already robust Kaggle results. This pattern once again underscores the superiority of generative augmentation in the development of high-fidelity synthetic samples that accurately represent the actual distribution of fraudulent transactions.

The consistent enhancement across classifiers demonstrates the generalizability of the synthetic data: both SVM and XGBoost exhibit enhanced precision and recall above 0.96. The integration of KL-Divergence loss into K-CGAN stabilizes generator–discriminator training, resulting in a more diverse set of samples and a smoother convergence.

Ultimately, the hybrid K-CGAN + SMOTE approach yields result that are nearly identical to those of pure K-CGAN, thereby confirming that generative augmentation is sufficient to attain peak classifier performance. RF continues to be the most resilient learner across all models, consistently achieving the highest and most consistent scores across metrics.

These findings confirm that K-CGAN based augmentation not only rectifies imbalances but also enhances the underlying data manifold, thereby allowing machine learning classifiers to achieve near-perfect recognition of fraudulent patterns in complex and chaotic datasets. The proposed approach's reproducibility and reliability are underscored by the consistency of the two datasets, thereby establishing a new gold standard in fraud detection performance benchmarking.

The traditional interpolation techniques were consistently outperformed by the K-CGAN based augmentation across both datasets, as evidenced by higher F1-scores, increased stability, and improved data diversity. This illustrates that K-CGAN and its hybrid variant are the most efficient synthetic data generators for balanced fraud detection scenarios.

4.4 Comparative analysis and identification of optimal combinations

Heatmaps of the F1-scores were generated for each dataset to visually consolidate the findings of the 36-experiment benchmark. These visualizations immediately and intuitively depict the performance hierarchy, clearly delineating the efficacy of each augmentation-classifier pair.

The heatmaps allow for the formal identification of the optimal combinations based on the primary F1-Score metric, which balances the critical trade-off between precision and recall. For the Kaggle dataset, the optimal combination is unequivocally K-CGAN and RF, which achieved the highest F1-Score of 0.9953. The bright yellow cell corresponding to this pair in Figure 1 stands in stark contrast to the rest of the map. For the NNEnsLeG dataset, the top-performing combination is also K-CGAN and RF, with an F1-Score of 0.9862, as shown in Figure 2. XGBoost and K-CGAN achieved nearly identical scores, resulting in a statistical stalemate. This demonstrates that both ensemble approaches perform exceptionally well when high-quality generating data is available.

This method revealed a critical nuance that has substantial practical implications. The K-CGAN method, which is highly sophisticated, outperformed all other methods on both datasets. On the smaller Kaggle dataset, the performance disparity was significantly greater than that of standard SMOTE. When SMOTE was integrated with XGBoost, an F1-score of 0.8475 was achieved on the comprehensive, feature-rich NNEnsLeG dataset. This suggests that, despite the fact that K-CGAN represents the theoretical peak, a well-tuned SMOTE approach can still be a highly viable and cost-effective alternative, albeit suboptimal, for organizations with extensive datasets where the computational expense and complexity of training a GAN are exorbitant.

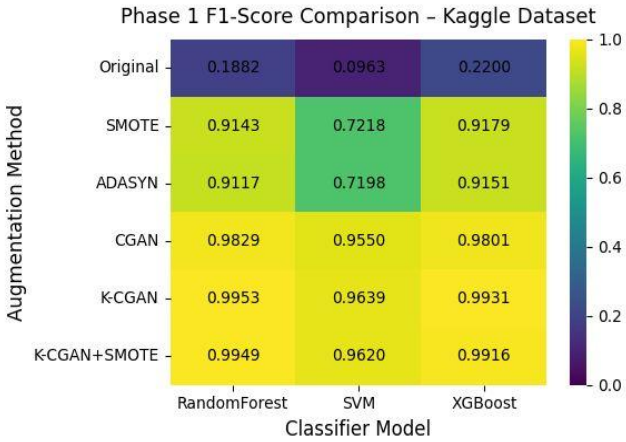

Figure 1. F1-score comparison for the Kaggle dataset

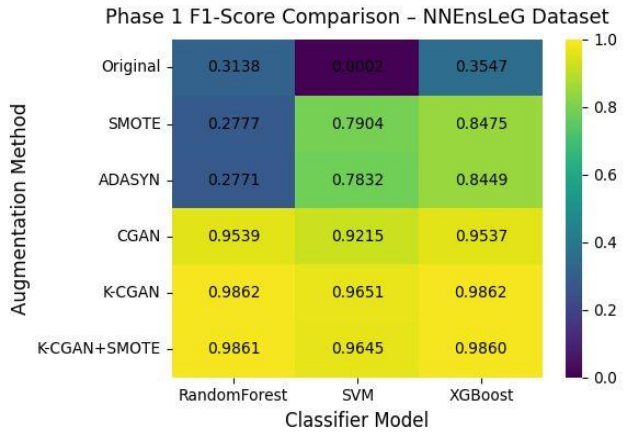

Figure 2. F1-score comparison for the NNEnsLeG dataset

4.5 Validation of stability for the optimal combinations (robustness test)

Phase 2 sought to confirm the dependability and stability of the optimal combinations revealed across diverse training data volumes. The optimal pairings (K-CGAN + RF for Kaggle

dataset and SMOTE + RF for NNEnsLeG dataset, the latter chosen to test the robustness of the "viable alternative" on the larger dataset) were re-evaluated using five different training set sizes, from 50% to 90% of the total balanced data.

The results of this robustness test are a strong confirmation that the high performance observed in Phase 1 is not a statistical artifact of a specific data division, but rather a consistent and reliable outcome. The line plot in Figure 3 for the Kaggle dataset is remarkably flat and positions the model near the performance ceiling. Regardless of the volume of training data, the F1-Score remains exceedingly high and stable, fluctuating marginally around 0.975. This is a significant discovery, as it suggests that the synthetic data produced by K-CGAN is of such high quality and information-dense that the RF classifier can achieve peak performance with substantially less training exposure than might be anticipated. In the NNEnsLeG dataset, the SMOTE+RF combination exhibits a remarkably stable performance, as illustrated in Figure 4, with an F1-Score that remains consistent at approximately 0.847. This consistency demonstrates that the significant gains provided by SMOTE are reliable and not highly sensitive to variations in training set size. This stability for both the absolute best and the viable alternative combinations solidifies their status as dependable, high-performing pairs within this controlled study, confirming that the benchmark results are robust and practically relevant.

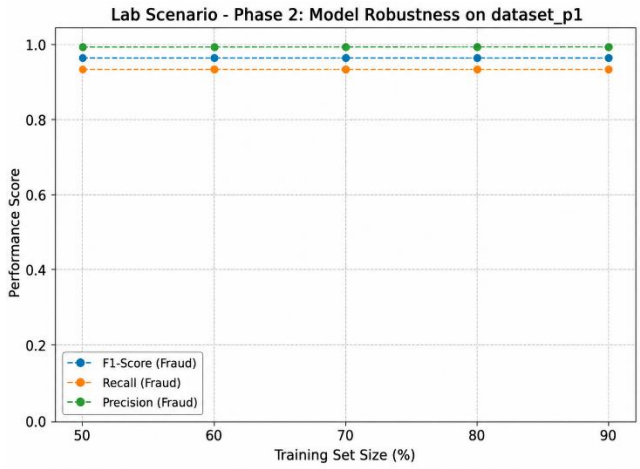

Figure 3. Performance of the optimal combination (K-CGAN + RF) on the Kaggle dataset

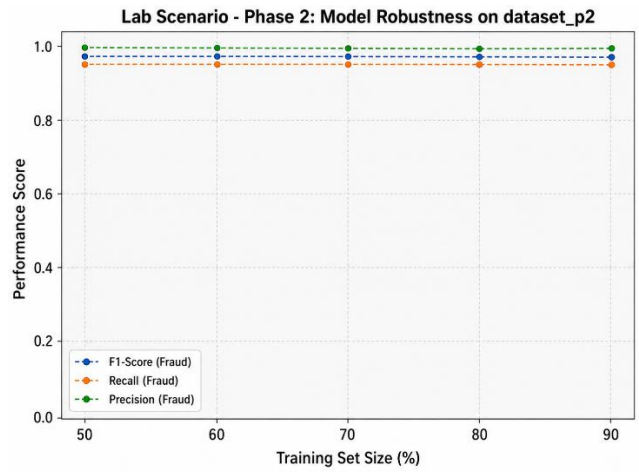

Figure 4. Performance of the SMOTE + RF on the NNEnsLeG dataset

4.6 Comparison with previous research

The results of this investigation both corroborate and enhance the existing literature on fraud detection. Our findings initially validate the established agreement that data augmentation is crucial for developing an effective fraud detection algorithm. The significant enhancement in performance above the baseline achieved using conventional oversampling techniques validates the efficacy of SMOTE and ADASYN, a finding extensively corroborated in previous studies [12, 33]. The significant enhancements in the F1-score exemplify how these strategies expand the decision boundaries for the minority class, as theoretically articulated in prior research [34]. Nevertheless, the finding that generative models outperformed these techniques underscores the constraints highlighted in the literature, which noted that fundamental interpolation can generate noise and fail to accurately capture the full complexities of the fraud distribution [35]. This research substantially enhances the validation of sophisticated generative models. Our results, which exhibit virtually perfect F1-scores (~0.99) achieved using K-CGAN, effectively corroborate the theoretical advantages outlined in the literature. The better efficacy of K-CGAN relative to conventional CGAN validates previous assertions that the incorporation of KL-Divergence loss yields enhanced data generation fidelity and ameliorates issues such as mode collapse [41, 42]. This study fundamentally differs from prior research in its methodological emphasis. The literature contains a clear and persistent call for more systematic, head-to-head comparisons of augmentation techniques in controlled settings, a gap explicitly identified in previous research [14, 15, 17]. Our study was designed specifically to answer this call.

4.7 Limitations

The primary limitation of this study is inherent in its design and must be clearly stated. The exceptional performance metrics reported represent a theoretical upper bound in an idealized experimental environment, not a direct prediction of real-world performance. By training and testing on balanced datasets, we deliberately sidestepped the significant challenge of model generalization a key hurdle for deploying fraud detection systems in production, where the model will inevitably encounter highly imbalanced, unseen data. The perfect recall and precision seen in our results would not be replicated in such a scenario, as classifiers trained on a 50/50 fraud distribution often become overly sensitive and generate an unacceptably high number of false positives when evaluated on real-world data where fraud is rare.

This study focused on a specific set of classifiers (RF, XGBoost, SVM) and GAN architectures (CGAN, K-CGAN). Despite their efficacy and widespread application, they lack comprehensiveness. The performance benchmark is specific to these combinations and may not be applicable to other model types, such as deep neural networks or different generative designs. The computational cost and complexity of training advanced generative models like K-CGAN are substantial, presenting a practical limitation for smaller organizations or applications that require rapid model retraining.

4.8 Future research direction

The findings and limitations of this benchmark study

present numerous intriguing opportunities for subsequent research. The primary focus is to explore strategies for closing the "generalization gap" between the theoretical performance ceiling identified here and the actual performance on imbalanced production data. This may entail investigating methods such as a) formulating training protocols that introduce the model to both balanced augmented data and minor, imbalanced subsets of actual data to enhance its calibration; b) implementing innovative regularization techniques or domain adaptation approaches specifically aimed at mitigating overfitting to the attributes of synthetic data; c) creating advanced post-training calibration techniques that can modify a model's prediction threshold to optimize precision-recall trade-offs in a real-world imbalanced context. A further intriguing domain is the investigation of increasingly sophisticated generative architectures. The domain of generative modeling is advancing swiftly, with emerging models such as Wasserstein GANs with Gradient Penalty (WGAN-GP) or Transformer-based GANs potentially providing enhanced stability and superior data fidelity compared to the K-CGAN variation employed herein. Utilizing this benchmarking methodology on these contemporary architectures would yield an enhanced and more resilient theoretical performance benchmark. The benchmarking framework is exceptionally portable. This methodology may be utilized in other vital areas affected by significant class imbalance, such rare disease identification in medical imaging, anomaly detection in industrial manufacturing, or the identification of advanced threats in cybersecurity logs. The interdisciplinary applicability of generative augmentation would be confirmed and substantial performance criteria for those areas would be provided by conducting this investigation across other domains.

4. CONCLUSIONS

This comprehensive benchmarking study demonstrates that, given a regulated and balanced data environment, the choice of data augmentation method is the most critical factor affecting the performance threshold for e-commerce fraud detection. While traditional oversampling techniques like SMOTE and ADASYN markedly improve baseline performance, advanced generative methods, like as K-CGAN, perform exceptionally well on their own, enabling classifiers to achieve near-perfect F1-scores approaching 0.99. The findings clearly establish that the high-fidelity synthetic data produced by K-CGAN is theoretically superior, allowing powerful ensemble classifiers like RF and XGBoost to learn the nuanced patterns of fraudulent transactions almost flawlessly. The primary contribution of this research is the creation of a performance standard that is resilient, transparent, and distinctive. By identifying the optimal combinations of classification and augmentation that can achieve the highest theoretical performance, this research establishes a definitive benchmark for the industry. It functions as an essential reference for both academic and industrial practitioners.

ACKNOWLEDGMENT

Thanks to those who contributed with their suggestions and recommendations for the completion of the research.

REFERENCES

[1] Mohamad, Z., Ismail, Z., Abdullah Thani, A.K. (2023). Determinants of fraud victimizations in Malaysian e commerce: A conceptual paper. International Journal of Academic Research in Business and Social Sciences, 13(12): 5091-5097. http://doi.org/10.6007/IJARBSS/v13-i12/20395

[2] Bogi, B. (2024). Implementing account takeover and phishing detection models to mitigate e-commerce fraud. Kuwait Journal of Machine Learning, 3(2): 1-7. https://doi.org/10.52783/kjml.255

[3] Sholahuddin, M., Nasir, M., Bawono, A.D.B., Permatasari, Q., Annisa, W. (2024). Building financial awareness and personal branding: Strategies to avoid riba and high-risk online transactions. Unram Journal of Community Service, 5(4): 297-305. https://doi.org/10.29303/ujcs.v5i4.756

[4] Fariha, N., Khan, M.N.M., Hossain, M.I., Reza, S.A., et al. (2025). Advanced fraud detection using machine learning models: Enhancing financial transaction security. International Journal of Accounting and Economics Studies, 12(2): 85-104. https://doi.org/10.14419/c73kcb17

[5] Md Noor, A.A., Haron, N.H., Sayed Rohani, S.R., Abd Rahman, R. (2022). Covid-19 pandemic and online fraud: Malaysian experience. International Journal of Academic Research in Accounting, Finance and Management Sciences, 12: 192-207.

[6] Suela, L.C. (2024). Online fraud exposed: Tactics and strategies of cyber scammers. International Journal of Scientific Research in Engineering and Management, 8(4): 1-5.

[7] Rout, S., Jaiswal, K.L. (2024). Fraud detection using deep learning. International Journal of Electrical and Data Communication, 5(1): 7-11.

[8] Breskuvienė, D., Dzemyda, G. (2023). Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. International Journal of Computers Communications & Control, 18(3). https://doi.org/10.15837/ijccc.2023.3.5433

[9] Sooklal, S., Hosein, P. (2023). Framework for credit card fraud detection using benefit-based learning and periodic features. https://doi.org/10.21203/rs.3.rs-2652853/v1

[10] Pan, E. (2024). Machine learning in financial transaction fraud detection and prevention. Transactions on Economics, Business and Management Research, 5: 243-249.

[11] Lai, G. (2023). Artificial intelligence techniques for fraud detection. Preprints.org. https://doi.org/10.20944/preprints202312.1115.v1

[12] Noviandy, T.R., Idroes, G.M., Maulana, A., Hardi, I., Ringga, E.S., Idroes, R. (2023). Credit card fraud detection for contemporary financial management using XGBoost-driven machine learning and data augmentation techniques. Indatu Journal of Management and Accounting, 1(1): 29-35. https://doi.org/10.60084/ijma.v1i1.78

[13] Strelcenia, E., Prakoonwit, S. (2023). A new GAN-based data augmentation method for handling class imbalance in credit card fraud detection. In 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, pp. 627-634.

https://doi.org/10.1109/SPIN57001.2023.10116543
[14] De Zarzà, I., De Curtò, J., Calafate, C.T. (2023). Optimizing neural networks for imbalanced data. Electronics, 12(12): 2674. https://doi.org/10.3390/electronics12122674
[15] Yang, Y., Wang, Y., Cao, J., Chen, J. (2022). HearLiquid: Nonintrusive liquid fraud detection using commodity acoustic devices. IEEE Internet of Things Journal, 9(15): 13582-13597. https://doi.org/10.1109/JIOT.2022.3144427
[16] Suparyati, S., Utami, E., Muhammad, A.H. (2022). Applying different resampling strategies in random forest algorithm to predict lumpy skin disease. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 6(4): 555-562.
[17] Esenogho, E., Mienye, I.D., Swart, T.G., Aruleba, K., Obaido, G. (2022). A neural network ensemble with feature engineering for improved credit card fraud detection. IEEE Access, 10: 16400-16407. https://doi.org/10.1109/ACCESS.2022.3148298
[18] Jain, N., Patil, S. (2024). Artificial intelligence models for fraud detection: Advancements, challenges, and future prospects. International Journal of Global Innovations and Solutions (IJGIS). https://doi.org/10.21428/e90189c8.6d8ab5f6
[19] Cho, S.T., Kow, D.W., Twan, B.C. (2023). Fraud detection in Malaysian financial institutions using data mining and machine learning. Journal of Information, Technology and Data Science, 7(1): 13-21. https://doi.org/10.53819/81018102t4152
[20] Huang, L., Abrahams, A., Ractham, P. (2022). Enhanced financial fraud detection using cost-sensitive cascade forest with missing value imputation. Intelligent Systems in Accounting, Finance and Management, 29(3): 133-155. https://doi.org/10.1002/isaf.1517
[21] Saranya, N., Devi, M.K., Mythili, A., H, S.P. (2023). Data science and machine learning methods for detecting credit card fraud. The Scientific Temper, 14(3): 840-844. https://doi.org/10.58414/SCIENTIFICTEMPER.2023.14.3.43
[22] Lin, D. (2023). An empirical analysis of machine learning for fraud detection in diverse financial scenarios. Advances in Economics, Management and Political Sciences, 42(1): 202-216.
[23] Prabha, M., Sharmin, S., Khatoon, R., Imran, M.A.U., Mohammad, N. (2024). Combating banking fraud with it: Integrating machine learning and data analytics. The American Journal of Management and Economics Innovations, 6(7): 39-56. https://doi.org/10.37547/tajmei/Volume06Issue07-04
[24] Hasugian, L.S. (2023). Fraud detection for online interbank transaction using deep learning. Journal of Syntax Literate, 8(6): 4263. https://doi.org/10.36418/syntax-literate.v8i6.12627
[25] Vallarino, D. (2025). Graph AI for fraud detection: Improving risk management and compliance in digital finance. https://doi.org/10.21203/rs.3.rs-6203866/v1
[26] Trisanto, D., Rismawati, N., Mulya, M.F., Kurniadi, F.I. (2021). Modified focal loss in imbalanced XGBoost for credit card fraud detection. International Journal of Intelligent Engineering and Systems, 14(4): 350-358. https://doi.org/10.22266/ijies2021.0831.31
[27] Nabrawi, E., Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. Risks, 11(9): 160. https://doi.org/10.3390/risks11090160
[28] Tarimo, C.S., Bhuyan, S.S., Zhao, Y., Ren, W., et al. (2022). Prediction of low Apgar score at five minutes following labor induction intervention in vaginal deliveries: Machine learning approach for imbalanced data at a tertiary hospital in North Tanzania. BMC Pregnancy and Childbirth, 22(1): 275. https://doi.org/10.1186/s12884-022-04534-0
[29] Wei, E. (2023). Impact of different transaction features on credit card fraud detection by neural networks. Applied and Computational Engineering, 4(1): 610-617. https://doi.org/10.54254/2755-2721/4/2023333
[30] Mehdary, A., Chehri, A., Jakimi, A., Saadane, R. (2024). Hyperparameter optimization with genetic algorithms and XGBoost: A step forward in smart grid fraud detection. Sensors, 24(4): 1230. https://doi.org/10.3390/s24041230
[31] Muslikh, A.R., Ojugo, A.A. (2023). Rice disease recognition using transfer learning Xception convolutional neural network. Jurnal Teknik Informatika (Jutif), 4(6): 1535-1540. https://doi.org/10.52436/1.jutif.2023.4.6.1529
[32] Akazue, M.I., Debekeme, I.A., Edje, A.E., Asuai, C., Osame, U.J. (2023). Unmasking fraudsters: Ensemble features selection to enhance random forest fraud detection. Journal of Computing Theories and Applications, 1(2): 201-211. https://doi.org/10.33633/jcta.v1i2.9462
[33] Gnip, P., Vokorokos, L., Drotár, P. (2021). Selective oversampling approach for strongly imbalanced data. PeerJ Computer Science, 7: e604. https://doi.org/10.7717/peerj-cs.604
[34] Nikiforos, M.N., Deliveri, K., Kermanidis, K.L., Pateli, A. (2023). Vocational domain identification with machine learning and natural language processing on Wikipedia text: Error analysis and class balancing. Computers, 12(6): 111. https://doi.org/10.3390/computers12060111
[35] Javale, D.P., Desai, S.S. (2022). Machine learning ensemble approach for healthcare data analytics. Indonesian Journal of Electrical Engineering and Computer Science, 28(2): 926. https://doi.org/10.11591/ijeecs.v28.i2.pp926-933
[36] Zhao, Y., Ma, Z., Jiang, X., Koutsopoulos, H.N. (2022). Short-term metro ridership prediction during unplanned events. Transportation Research Record, 2676(2): 132-147. https://doi.org/10.1177/03611981211037553
[37] Sewpaul, R., Awe, O.O., Dogbey, D.M., Sekgala, M.D., Dukhi, N. (2023). Classification of obesity among south African female adolescents: Comparative analysis of logistic regression and random forest algorithms. International Journal of Environmental Research and Public Health, 21(1): 2. https://doi.org/10.3390/ijerph21010002
[38] Zhang, Y., Li, J., Wang, H., Choi, S.C.T. (2021). Sentiment-guided adversarial learning for stock price prediction. Frontiers in Applied Mathematics and Statistics, 7: 601105. https://doi.org/10.3389/fams.2021.601105
[39] Lupión, M., Cruciani, F., Cleland, I., Nugent, C., Ortigosa, P.M. (2024). Data augmentation for human activity recognition with generative adversarial networks. IEEE Journal of Biomedical and Health Informatics, 28(4): 2350-2361.

https://doi.org/10.1109/JBHI.2024.3364910
[40] Makhlouf, A., Maayah, M., Abughanam, N., Catal, C. (2023). The use of generative adversarial networks in medical image augmentation. Neural Computing and Applications, 35(34): 24055-24068. https://doi.org/10.1007/s00521-023-09100-z
[41] Tibebu, H., Malik, A., De Silva, V. (2022). Text to image synthesis using stacked conditional variational autoencoders and conditional generative adversarial networks. In Science and Information Conference, pp. 560-580. https://doi.org/10.1007/978-3-031-10461-9_38
[42] Hu, Y. (2023). Influence of images generated by CGAN on the performance of image classification based on CNN. In Fifth International Conference on Computer Information Science and Artificial Intelligence, pp. 307-312. https://doi.org/10.1117/12.2669802
[43] Bokolo, B.G., Liu, Q. (2024). Advanced algorithmic approaches for scam profile detection on Instagram. Electronics, 13(8): 1571. https://doi.org/10.3390/electronics13081571
[44] Hsu, C.T., Pai, K.C., Chen, L.C., Lin, S.H., Wu, M.J. (2023). Machine learning models to predict the risk of rapidly progressive kidney disease and the need for nephrology referral in adult patients with type 2 diabetes. International Journal of Environmental Research and Public Health, 20(4): 3396. https://doi.org/10.3390/ijerph20043396
[45] Guo, S., Zhang, B. (2024). Revolutionizing the used car market: Predicting prices with XGBoost. Applied and Computational Engineering, 48(1): 173-180.
[46] Zheng, Z., Liang, L., Luo, X., Chen, J., Lin, M., Wang, G., Xue, C. (2024). Diagnosing and tracking depression based on eye movement in response to virtual reality. Frontiers in Psychiatry, 15: 1280935. https://doi.org/10.3389/fpsyt.2024.1280935
[47] Tian, N., Shao, B., Zeng, H., Ren, M., Zhao, W., Zhao, X., Wu, S. (2025). Multi-step natural gas load forecasting incorporating data complexity analysis with finite features. Entropy, 27(7): 671. https://doi.org/10.3390/e27070671
[48] Devana, S.K., Shah, A.A., Lee, C., Gudapati, V., et al. (2021). Development of a machine learning algorithm for prediction of complications and unplanned readmission following reverse total shoulder arthroplasty. Journal of Shoulder and Elbow Arthroplasty, 5: 24715492211038172. https://doi.org/10.1177/24715492211038172
[49] Wang, H., Wang, W., Liu, Y., Alidaee, B. (2022). Integrating machine learning algorithms with quantum annealing solvers for online fraud detection. IEEE Access, 10: 75908-75917. https://doi.org/10.1109/ACCESS.2022.3190897
[50] Kumar, P., Priyanka, P., Uday, K.V., Dutt, V. (2024). Addressing class imbalance in soil movement predictions. Natural Hazards and Earth System Sciences, 24(6): 1913-1928. https://doi.org/10.5194/nhess-24-1913-2024
[51] Zeng, Q., Lin, L., Jiang, R., Huang, W., Lin, D. (2025). NNEnsLeG: A novel approach for e-commerce payment fraud detection using ensemble learning and neural networks. Information Processing & Management, 62(1): 103916. https://doi.org/10.1016/j.ipm.2024.103916