ISSN: 2320-5407

International Journal of Advanced Research

Publisher's Name: Jana Publication and Research LLP

www.journalijar.com

REVIEWER'S REPORT

Manuscript No.: IJAR-53868 Date: 18/09/25

Title: Enhancing Pdf Malware Classification Using Ctgan-Based Data Augmentation And Supervised Learning

Recommendation:
Accept as it is .....
Accept after minor revision.....yes.....
Accept after major revision .....
Do not accept (*Reasons below*) .....

Rating	Excel.	Good	Fair	Poor
Originality		y		
Techn. Quality		y		
Clarity		y		
Significance		y		

Reviewer Name:Dr Shaweta Sachdeva **Date:** 18/9/25

Reviewer's Comment for Publication.Accepted with Minor revision

(*To be published with the manuscript in the journal*)

The reviewer is requested to provide a brief comment (3-4 lines) highlighting the significance, strengths, or key insights of the manuscript. This comment will be Displayed in the journal publication alongside with the reviewers name.

Significance of the Manuscript

- Addresses a **critical cybersecurity challenge**: detecting malicious PDF files, which are a common attack vector in real-world digital systems.
- Tackles the **class imbalance problem** in malware datasets by using **Conditional Tabular GAN (CTGAN)**, providing a novel and effective augmentation approach compared to traditional oversampling methods.
- Combines **predictive performance with explainability** by integrating SHAP, which enhances trust and interpretability—important for real-world deployment in sensitive environments (government, education, enterprises).
- Contributes a **deployable framework** with practical relevance, not just theoretical novelty.

ISSN: 2320-5407
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

REVIEWER'S REPORT

✅ Strengths

1. Methodological Innovation:
   - Unique integration of CTGAN-based data augmentation with supervised learning models.
   - Systematic evaluation across six machine learning classifiers ensures robustness.
2. High Performance:
   - XGBoost on augmented data achieved near-perfect metrics (accuracy, F1, AUC, MCC), demonstrating effectiveness of the proposed pipeline.
3. Explainability and Trustworthiness:
   - Use of SHAP for local and global interpretability sets this work apart from many black-box models.
   - Feature insights (e.g., role of JavaScript presence, metadata volume) align well with domain knowledge in malware detection.
4. Thorough Validation:
   - Includes statistical validation (KS tests, PCA analysis, correlation comparison) to show that CTGAN-generated samples maintain structural and statistical fidelity.
   - Comparative performance evaluation across both original and augmented datasets adds credibility.
5. Practical Relevance:
   - Modular, lightweight design compatible with existing Security Operations Center (SOC) tools.
   - Emphasis on operational deployment feasibility, including low computational overhead.

🔍 Key Insights

- CTGAN augmentation significantly improves detection sensitivity for malicious PDFs, especially under class imbalance conditions.
- Ensemble models (Random Forest, XGBoost) outperform traditional classifiers, with XGBoost emerging as the best choice.
- Interpretable features (e.g., JavaScript triggers, metadata size, OpenAction) not only boost accuracy but also provide actionable intelligence for analysts.

ISSN: 2320-5407
International Journal of Advanced Research
Publisher's Name: Jana Publication and Research LLP
www.journalijar.com

REVIEWER'S REPORT

- Demonstrates that **synthetic data, if generated carefully, can meaningfully enhance cybersecurity detection models** without introducing bias.
- Highlights the importance of balancing **predictive accuracy** with **explainability**, especially in high-stakes cybersecurity contexts.

Detailed Reviewer's Report

1. It clearly summarizes the work, but it could be more concise. Consider quantifying improvements (e.g., "XGBoost improved recall by X% compared to Random Forest").
2. Highlight the novelty more explicitly (e.g., why CTGAN + SHAP integration is unique compared to prior works).
3. Add a stronger motivation by including real-world case studies or recent statistics on PDF-based cyberattacks to emphasize urgency.
4. Clarify the research gap earlier: what exactly is missing in existing methods that this paper solves?
5. The workflow is well-explained, but the pseudocode could be simplified for clarity.
6. Provide justification for choosing only six supervised models—why not compare with deep architectures given prior work?
7. Expand on dataset preparation: Were benign vs malicious PDFs sourced from publicly available corpora? If yes, cite them.
8. Current performance metrics show near-perfect results (≈99.8%), which raises concerns of possible overfitting or dataset bias. Recommend validating with an external benchmark dataset.
9. Add confidence intervals or statistical tests (e.g., t-test, Wilcoxon) to prove significance of performance gains.
10. The confusion matrix is useful, but include class-specific precision/recall values in a tabular form for transparency.