Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Synthetic-Data Generation for Enhancing Malware and Phishing Determining Performance

Ms. Thejaswini S
Assistant Professor
Dept. of Computer Application
STG First Grade College Chinakurali,
Mandya, Karnataka, India

Mr. Alfred Vivek Joseph
Assistant Professor
Dept. of English
STG First Grade College
Chinakurali, Mandya, Karnataka, India

***Abstract*—The ML applications like Malware and phishing detection require security datasets, which should be of good quantity, quality, and diversity, but in real-world applications, they may deficit future (zero-day) or avoid variants, are not balanced, and provide privacy issues. Synthetic-Data Generation (SDG) (including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), transformer or large language model (LLM) generation) can be used to expand training corpora as well as simulate obscure variants as well as allow privacy-preserving collaboration. The proposed research model encompasses the literary background, recent developments (2021-2025), an experimental design, guidelines, ethics, and threat assessment, as well as the expected outcomes. Recent studies, such as those by Mal Data Gen, malware benchmarks, phishing synthesis using LLM, and improvements based on GANs, are used to support the affirmation.**

Keywords—phishing, synthetic data, detection models, training, recall, Cybersecurity

I. Introduction

Unethical email and internet-based malware, along with phishing sites, are some of the main causes of identity theft, leakage, and disruption of the supply chain. Although machine learning proves to be effective in determining attack patterns, there are a number of major challenges existing presently. First of all, real harmful or malicious data is insufficient, especially for new and elusive attacks, leading to unbalanced datasets, i.e., benign traffic. Secondly, the tactics used by hackers are constantly changing, and new operations appear, which are not identified in the training data. Thirdly, it is diminished by privacy and legal limitations that can prevent the sharing of original web logs and email datasets between organizations. Synthetic Data Generation (SDG) solves these issues by creating natural artificial samples, including URLs, emails, HTML pages, vectors, and opcode lists. These can be used to raise training data sets and explore uncommon types of attacks, and conduct associated research without exposing the data. The increased number of available SDG tools and datasets in recent years highlights the necessity of a systematic evaluation of SDG to improve web threat identification.

II. Research Questions:

RQ1: To what extent does Systematic Data Generation enhance the phishing and malware supervised detection measures, including recall and F1?

• RQ2: Which families of SDGs—GANs, VAEs, LLMs, or hybrids-work best across different types of data like tabular features, opcode sequences, or text/URLs?

RQ3: Does SDG make models more robust against concept drift or adversarial evasion?

• RQ4: What are the privacy and ethics safeguards for deployments in the wild of SDGs?

III. Recent Developments & Literature Synthesis (2021–2025)

3.1 Generative Models for Malware Augmentation

Recently, a number of works have demonstrated that both GANs and VAEs can generate all kinds of malware-related data-opcode sequences, binary images, behavioral traces-to augment model training and validate their resilience. Joshi et al. (2025) have determined that adding GAN-generated samples to training sets significantly improves models' capability of detecting new, unseen malware. For example, MalDataGen (2025) enables you to generate synthetic, high-quality tabular data that's tailor-made for malware detection. So far, this looks very promising, as long as you will also be able to keep a close eye on realistic generated data and polish it when needed.

3.2 Transformer/LLM Enhancement of Phishing

LLMs and transformers are now being used to whip up phishing emails, fake messages, and social engineering bait. Projects such as PhishEmailLLM (2025) demonstrate that LLMs can already produce convincing phishing samples for training meta-classifiers. On the other hand, some researchers point out that these tools might be leveraged by an attacker to create better lures; thus, there is a real need for safe filtering and controlled generation. The other way to gain traction is a self-identification approach like the PhishSSL family that allows examining and training with a few named data.

3.3 Benchmarks and Dataset Updates

Even now, solid benchmarks matter when evaluating what SDG can do. The results from EMBER2024 - fresh data arriving at KDD/2025 - reflect how crucial large, up-to-date collections have become: they improve generator training and give the results and evaluations a stronger foundation. When it is related to phishing, specially chosen feature packs from 2024 bring together letter patterns, site layouts, and ownership of domain clues, resulting in tests that go far beyond controlled settings into real-world problems.

3.4 Privacy and Data Protection Tools

Compared to instruments like DP-GANs, private VAEs have received notice for their ability to create false data without showing out personal data and securing privacy. When it comes to protection and security, using differentially private generative tools helps in putting a stop to synthetic datasets from sharing users' data and identity. Still, there is always a compromise – tougher protection many times weakens how useful the output turns out to be, so adjusting these models takes real and concerned care.

3.5 Adversarial Considerations

These Generative Models has both Positive and Negative Outputs. On the positive note, creators can use these generative models to create realistic adversarial inputs, such as Malicious inputs, fake attacks, or deceptive patterns, to train and strengthen the security and privacy systems. By applying these models to these difficult cases, systems become better at detecting the threats in real situations. On the negative side, the same models can be utilised by the attackers. They can use these automatically generated persuasive phishing messages, malwares, deepfakes, or create inputs that dodge detection, making attacks more effective and harder to identify.

IV. Theoretical Foundations and Key Concepts

4.1 How SDG enhances Machine Learning in Security

•*What happens if we cutdown how much sample is needed?*

Synthetic (Fake) datas helps in increasing the effective size and diversities in the training datas, helping the algorithms understand uncommon categories faster. This is useful for rare or uncommon attack types that appear only a few times in real datasets. With extra exposure comes a clearer judgement and understand where to draw decision boundaries between malicious and normal behaviour.

- *Discovering More and Smoothing Things Out:*

Generative Models can produce unusual and rare examples, such as clear and sophisticated obfuscation techniques that do not exist in the original datasets. This helps the models learn broader and more general patterns rather than depend on the known examples. This helps them to remain effective even when attacks evolve, as the exposure to various and surprising inputs during the training strengthens the ability to handle unseen or changing threats in real-world scenarios.

- *Strengthening up against adversaries*

Adding synthetic, intentionally misleading example datasets during training works like adversarial training. This guides the models to comprehend how the hackers try to fool it, and each outcome helps it to withstand the misleading inputs more efficiently, so mistakes made before are less likely to happen again. Through repetitive encounters with these kinds of fake scenarios, the system gradually becomes more resilient, thus reducing its susceptibility to false patterns and strengthening its working performance in real adversarial situations.

- *Keeping the Data Private.*

A well-designed synthetic data safeguards the statistical properties of real data without exposing sensitive personal or organizational information. This enables institutions to collaborate and share data on security research while adhering to privacy and data protection regulations.

4.2 Generative Model Types

- GANs/ WGANs-GP/ cGANs: these models are best at creating datasets that look realistic, such as tabular data or images. Conditional GANs can create data for a particular category or label, like a particular attack type. WGAN-GP improves the training stability, so the generated data is less noisy and more accurate.

- Flow Models/ VQ-VAE/VAEs: these focus on understanding a Smooth internal presentation of data. Therefore, these models allow better control when creating new samples and minimise strange or unrealistic outputs. They are especially useful when we need consistency and moderate variation in the generated data.

- LLMs/ Transformers: These models work best with text data and can create realistic phishing fake conversations, emails, and scam messages. When these models are trained with constraints or rules, they can construct compelling patterns of text, such as URLs, one character or word at a time.

- Hybrid-Pipelines: These kinds of hybrid pipeline systems include different models for various types of data. Such as a GAN or VAE data generates number-related features, while an LLM generates practical texts. This makes sure all parts of the data match each other, making the concluding output more practical and more logical.

4.3 Measuring Generative Quality

- Intrinsic metrics:

These examine how related the synthetic data is to the real data by different their statistical properties. Methods like Kullback-Leibler (KL) divergence, Maximum Mean Discrepancy (MMD), and two-sample tests examine whether the total distributions are similar, while variety and novelty measures ensure the Synthetic data is not just duplicating but still diverse.

- Extrinsic Measures (Downstream):

These show how helpful the synthetic data is in practical world tasks. By training these identification models with the artificially generated data and testing on the original data. Systems like recall, F1 score, and AUROC reflect whether implementation improves in identifying rare cases.

- Robustness tests:

These measures how the systems perform under various circumstances, even during exposure to newer samples, intentionally modified data samples, and variants outside the familiar data they were trained on

V. Proposed Research Methodology

5.1 Overview of Experiment

Here, we run a series of controlled experiments to compare different data sources and generation methods. Firstly, the models are trained on real data to construct a baseline. Later, synthetic data is generated using VAE, LLM, GAN, and hybrid approaches are added in different proportions, including Low-label and few-shot settings. Finally, all these models are evaluated using standard test sets and time-based hold-out splits to measure the performance of the models handling new data in the future.

5.2 Datasets (Sources & Preprocessing)

- Malware: In the context of malware analysis, we use the EMBER dataset (including EMBER2024) to learn the static features. We rely on Mallmg and opcode datasets for sequence-based and image-style experiments. First, raw files are converted into meaningful features such as section sizes, imports, entropy, API-call statistics, and sequential representations like opcode or API sizes.
- Phishing: The data is collected from PhishTank, a set of curated phishing pages example- from Kaggle or academic disseminations, and a collection of emails (after a check to ensure it is legal, naturally). For these, we extract features including URL lexical patterns, HTML structure, WHOIS information, and complete email text.
- Safety and Ethics: Safety has to be prioritised; for this, all datasets are cleansed to remove live attacks and active exploits. Ethical approval, including IRB or equivalent review, is acquired wherever required to ensure the responsible and safe practice of research.

5.3 Synthetic Pipelines & Implementation Details

Pipeline 1: WGAN-GP for tabular Malware Features (MalDataGen style):

Train a conditional WGAN-GP on malware samples, conditioned on family labels. We'll post-process to make the output realistic, such as ensuring values remain within intuitive ranges and that integer columns don't behave peculiarly. Baseline settings are used for MalDataGen configurations.

Pipeline 2 => VAE / TVAE for Opcode / Sequence Generation:

In this case, we encode the sequences of op-codes or API calls, sample the decoded space, and decode back to the synthetic sequences. The static program analyzers ensure the produced code is meaningful. Pipeline 3 - Transformer/LLM for Phishing Text & URL Generation: We'll fine-tune a small quantized transformer because we want to keep this on the leash so that it doesn't get out of hand – and for generating phishing templates and URLs. We run all these tools through filters and safety measures before we exploit any of this. We can change attributes for creating more or less sophisticated phishing samples for variety.

Pipeline 3 – Transformer/LLM for Phishing Text and URL Generation:
We will fine-tune a small quantized transformer (to prevent things from going out of control) for creating phishing mail templates and URLs using thoughtfully designed prompts. Everything is filtered using filters based on rules and tested for safety before we use them. To add some variety, we can control attributes for Phelpsian samples that are less or more sophisticated.

Pipeline 4 -Hybrid (LLM + GAN metadata):

It combines LLM-based generated bodies for the email with the use of Host or WHOIS information generated via GAN or VAE, thus creating a phishing sample that resembles a cohesive piece of work.

Privacy Variants — DP-GANs For privacy concerns, we'll train our WGAN/GAN models with differential privacy. In this way, we can share this data safely because we have guarantees on privacy parameters (epsilon and delta) in these models; however, we understand that there is a trade-off between privacy and performance.

5.4 Detection Models and Training Approaches

- Classic Tabular Models: Random Forest Models and XGBoost are put into structured feature vectors extracted from the data, making them effective in handling numerical and categorical attributes.
- Ensembles: results from models trained on various data types are combined into a meta-classifier for improving overall accuracy and robustness.
- Training Setups: There are many ways to set up the training of models. We compare models trained only on real data with those trained on datasets mixed with real and synthetic data (ranging form 10% to completely synthetic), few-shot learning that is enhanced using synthetic samples, and a few models that use differentially private synthetic samples.

5.5 Evaluation Methods and Measures

- Cross-validation and temporal holdouts: 5-fold cross-validation is used to attain reliable performance statistics, besides temporal holdout splits, where the models are evaluated on data collected after the training period. This helps assess how well the models handle concept drift over time.
- Metrics: the entire performance is evaluated and measured using precision, F1-Score, AUROC, false positive rate, recall (with special emphasis on recall for security applications), and detection latency.

- Robustness tests: these tests are evaluated against adversarial samples created with red-team GANs and SpoofBots. We then measure the extent to which recall decreases and the extent to which retraining the model on these difficult-to-classify samples increases or restores the model's performance.
- Intrinsic Quality Checks: the quality of the synthetic data is evaluated using intrinsic metrics like MMD and KL divergence on the key features, plus with a classifier, two-sample tests to determine whether models can differentiate between the real and synthetic datasets
- Statistical Validation: The Wilcoxon signed-rank test is conducted across different and various undirected splits and seeds, and the results are reported with intervals to ensure the accuracy

VI. Threat Model Ethics and Safety Controls

6.1 Dual Use and Responsible Release

Strong protective measures are important as the data generators can be misled to create real scam messages or malware. One way to mitigate risk is to place measures that prevent the surprise or unexpected generation of functional output as executable code, and sensitive datasets should be tightly controlled. Prompt templates or methods that enable the creation of malicious content should never be made available to the public. Following responsible disclosure practices and obtaining ethics/ IRB approval will help ensure research is used only for defensive and academic purposes. Recent studies also stress the careful use of LLMs with proper filtering to prevent misuse.

6.2 Privacy and How Data Stays Anonymous

When synthetic data is developed using data sources, protecting the privacy of the original data is of great importance. Techniques like PATE-GAN, DP-GAN, help prevent models from leaking information about real people or systems. Researchers should clearly report privacy parameters such as epsilon and delta, and explain the balance between privacy and data usefulness. Although improved privacy may lead to a decline in the amount of details present in the dataset, it enables the sharing of data among different institutions and organizations in a much more secure manner.

VII. Expected Results and Assumptions

- H1: Impact of synthetic datasets on Detection performance: according to this hypothesis, using high-quality synthetic datasets produced generated by the GANs or LLMs would lead to better recall in malware and phishing detection models, especially when the dataset is not balanced or datasets that lack sufficient examples of attack. For instance, because there are limited instances of actual attack, models often fail to detect rare threats. These synthetic data increases the number of attack examples, helping the model learn better decision boundaries and detect more malware cases than baseline models trained only on real data.
- H2: Appropriateness of generative models of different data types: this hypothesis states that the efficacy of the synthetic datasets depends on the type of data provided. In the phishing detection LLMs are more effective as the phishing attacks mainly depend on the textual content such as spam messages and emails. GANs and VAEs are more effective for malware identification, as the Malware are numerical and structured in nature. Hence, matching the generative model to the data types guides to better detection performance
- H3: This hypothesis proposes that the synthetic data created with differential privacy can permit secure sharing of data across various institutions without compromising sensitive data and information. Yet, stronger the privacy settings, the noise injected into the data, ultimately leads to reduction in data quality and negatively affects detection performance. This means that while synthetic data maintains data privacy, a balance must be maintained between privacy protection and model accuracy.

VIII. Plan of Action for Implementation, Reproducibility and Artefacts

- Code and Artefacts: the codes will be made available in a modular fashion, including scripts for synthetic data generation, like VAE, WGAN-GP, and LLM fine-tuning, model training scripts, as well as evaluation notebooks. These synthetic datasets will be distributed with proper redactioning and access control in place to ensure they are not misused.
- Compute and Reproducibility: GPUs will be used to instruct generative models and deep learning detectors. The experiment environment may be either a local environment or a cloud. Fixed random seeds, docker containers, and version control be ensure reproducibility of results.
- Timeline (6-9months)

1. Data collection & pre-processing 1-2 months
2. Training synthetic data generators 2 months
3. Training detectors and experiments run 2 months
4. Privacy evaluation and robustness 1-2 months
5. Writing and results sharing 1 month

IX. Limitations & Risks

- Generator Quality: generator with poor build quality might create fake files, which ultimately can reduce model performance. This will be prevented by utilizing quality assurance measures for models (generators), utilising multiple models, and by manually reviewing samples.
- Dual-Use risk: Generative models such as LLMs can be exploited to produce harmful attack materials, hence, certain safety precautions will be in place prior to the release of digital media, and only a small amount of these things will be made available for public consumption.
- Privacy- Utility Trade: data quality lowered by differential privacy protects information, as privacy preference settings affects detection results

X. Contribution & Novelty

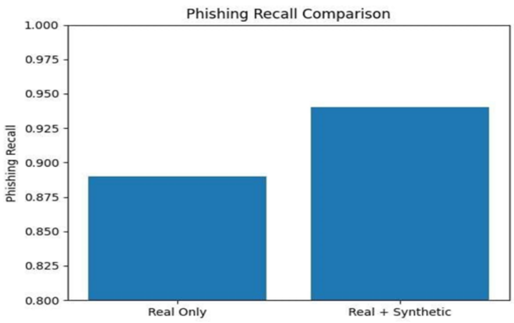
This research makes several key contributions; first, it provides a comprehensive comparison of various synthetic dataset generation methods, such as VAEs, GANs, Large Language Models, and hybrids, for malware and phishing detection. Secondly, it evaluates how efficiently these generators perform under growing malicious activities and avoidance methods like GAN evasion. Third, the work offers a modular and reusable pipeline for synthetic data generation and model training, drawing lessons from MalDataGen. Additionally, the work offers differentially synthetic datasets for safe sharing across institutions. Finally, the research offers real-world guidelines and best practices for applying synthetic data within real-world Cyber security Systems

XI. Project Model Results and how it Works

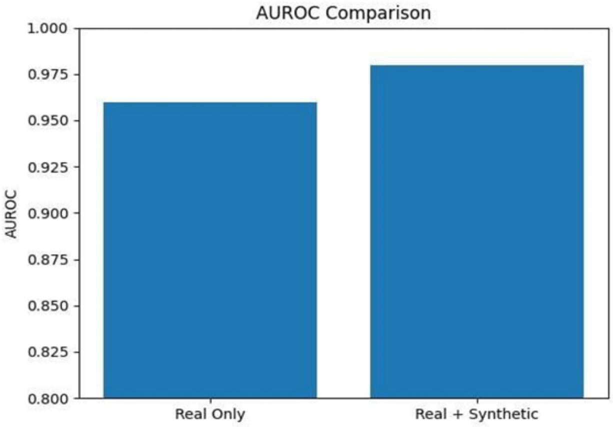
Starting from limited real data, this proposal demonstrates how synthetic datasets, generated using GANs, can enhance the identification of phishing and malware. Instead of relying solely on real attack samples, the inclusion of realistic synthetic data enhances key performances indicators such as F1-score and reacall. Fewer malicious activities go undetected this way. Synthetic data constructively handles two major challenges in cybersecurity research: fewer data availability and class imbalance.

The process starts with gathering actual phishing and malware samples. Noise removal, feature scaling, and feature selection are some of the pre-processing steps applied to them. A generative adversarial network is trained on the cleaned data to construct new synthetic attack samples that nearly resemble real threats while increasing data diversity. After the completion, they mix with real records to create a broader learning base. An XGBoost classifier is trained and instructed on the dataset results. The metrics of the Standard Model Performance include Recall, Precision, and F1-score. When put to the side against variants trained without artificial data, the augmented models show clear performance developments. These succeeding results reflect that the proposed pipeline is suitable for the practical-world security and educational research applications.

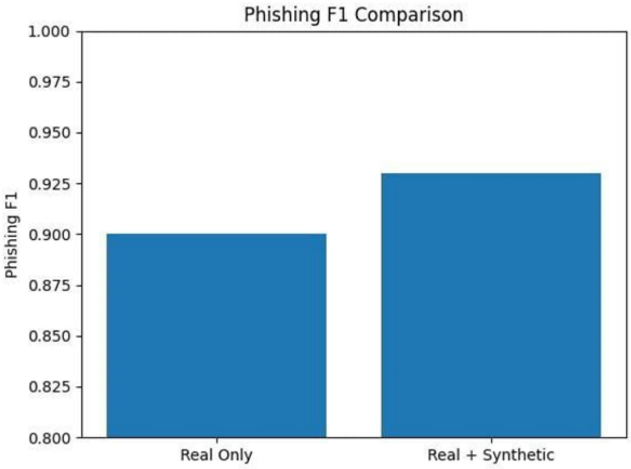
11.1. Graphical Representation of the outputs:



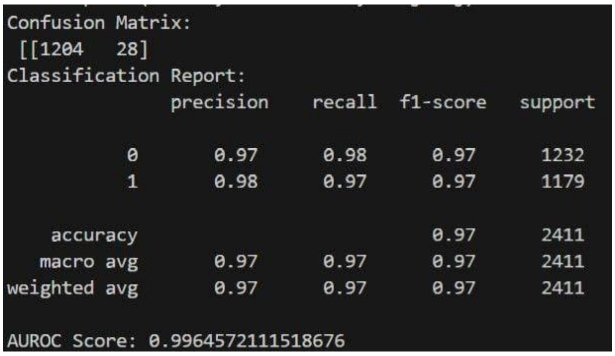
synthetic data to the training dataset improves the performance in phishing detection, increasing the recall from 0.89 to 0.94.



The graph illustrates that including the synthetic data increases the model's AUROC from 0.960 to 0.975.



The graph indicates that models trained on Both Real and synthetic datasets achieve a higher phishing F1 score than models trained only on Real data.



```
Confusion Matrix:
 [[1204   28]
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.97      1232
           1       0.98      0.97      0.97      1179

    accuracy                           0.97      2411
   macro avg       0.97      0.97      0.97      2411
weighted avg       0.97      0.97      0.97      2411

AUROC Score: 0.9964572111518676
```

The Classification Report indicates High and Balanced performance, with the scores of Precision with 0.97, recall with 0.98, and F1 Score with 0.97. finally the AUROC score of 0.996 represents the excellent discrimination capability of the model in identifying the malware

XII. CONCLUSION

The use of artificial datasets increased the identification of malware and spam attacks on internet platforms by machine learning. This paper evaluates these methods in detail, showing under what conditions they work well, where the shortcomings arise, and likewise, establishing the best practices. The code and supporting datasets are made public to enable repeatability and further research. As cyber threats evolve more quickly than ever before, and the use of AI (artificial intelligence) by hackers, data protection systems should also adopt the newly updated modelling tools, while ensuring better protection of privacy and data. Keeping this balance is very important for effective and responsible cybersecurity.

REFERENCES

1. Paim, K. O., Nogueira, A. G. D., Kreutz, D., Cordeiro, W., & Mansilha, R. B. (2025). MalDataGen: A Modular Framework for Synthetic Tabular Data Generation in Malware Detection. arXiv:2511.00361. — modular synthetic tabular framework and evaluation methodology.
2. Joshi, C., et al. (2025). Detection of unseen malware threats using generative models. Scientific Reports / Nature. — empirical evidence that GAN augmentation can improve unseen-variant detection. Nature
3. EMBER2024 (CrowdStrike, 2025 / KDD 2025). EMBER2024: Benchmark for Holistic Malware Classification. — Important recent dataset update for large-scale malware research.
4. PhishEmailLLM / related works (2025). — LLM-assisted phishing generation and detection meta-models. Useful for text augmentation design and safety controls. ACM Digital Library
5. Survey: GANs for Threat Detection (2021–2025). Systematic review (Sept 2025) — synthesizes GAN defenses and generator uses across IDS and malware detection. Useful for adversarial and robustness design.
6. Ho, S., et al. (2021). DP-GAN: Differentially private consecutive data publishing. — foundational DP-GAN methods for privacy-aware synthetic data. ScienceDirect
7. Kaggle & curated phishing feature datasets (2023–2025). e.g., Web-page phishing detection dataset, PhishTank feature datasets — practical sources for phishing experiments and feature engineering.
8. Qi, Q., et al. (2025). SpearBot / LLM adversarial phishing generation. — demonstrates attacker capabilities with LLMs and underscores need for safe generation.