ORIGINAL ARTICLE
Electrica 2026; 26: 1-13

Stability and Error Structure Analysis of Generative Adversarial Network–Based Architectures in Time Series Anomaly Detection

Begüm Çorbacıoğlu[1], Şafak Durukan-Odabaşı[2]

[1]Department of Cybersecurity, Bahçeşehir University, İstanbul, Turkiye
[2]Computer Engineering Department, İstanbul University-Cerrahpaşa, İstanbul, Turkiye

Cite this article as: B. Çorbacıoğlu and Ş. D. Odabaşı, "Stability and error structure analysis of generative adversarial network–based architectures in time series anomaly detection," *Electrica*, 26, 0062, 2026. doi: 10.5152/electrica.2026.26062.

WHAT IS ALREADY KNOWN ON THIS TOPIC?

- *Generative adversarial network (GAN)-based architectures have been widely applied to anomaly detection in time series and network security contexts.*
- *Existing studies often report performance improvements without systematically analyzing stability and error structure.*
- *Threshold selection strategies vary across studies, making fair comparison difficult.*

WHAT THIS STUDY ADDS TO THIS TOPIC?

- *This study provides a controlled comparison of three GAN-based architectures under a unified F1-optimized calibration framework.*
- *It demonstrates that architectural design determines error structure (false positives vs. false negatives).*
- *The results highlight the stability–sensitivity trade-off in adversarial time series anomaly-detection models.*

Corresponding Author:
Şafak Durukan Odabaşı
E-mail:
safak.odabasi@iuc.edu.tr

Received: February 27, 2026

Revision Requested: March 07, 2026

Last Revision Received: March 09, 2026

Accepted: March 14, 2026

Publication Date: March 30, 2026

DOI: 10.5152/electrica.2026.26062

Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ABSTRACT

The rapid expansion of Internet of Things ecosystems and interconnected network infrastructures has intensified the need for robust time series anomaly-detection methods. Generative adversarial network (GAN)-based approaches have gained increasing attention due to their capacity to model complex data distributions and learn representations of normal system behavior without explicit supervision. However, adversarial training may introduce instability and sensitivity issues, particularly in sequential anomaly-detection tasks. This study presents a controlled comparative evaluation of three GAN-based architectures, GANAD, N-GAN, and Net-GAN, under unified experimental conditions. All models are assessed on the Numenta Anomaly Benchmark artificial time series anomaly benchmark using identical preprocessing and an F1-optimized threshold calibration protocol to ensure methodological consistency. Performance is evaluated across ten random seeds using precision, recall, and F1-score to capture both effectiveness and robustness. The results reveal distinct architectural behaviors. The reconstruction-driven GANAD model achieves the highest mean F1-score (approximately 0.32) and maintains stable precision across runs, indicating consistent convergence and controlled false-positive rates. In contrast, the discriminator-centric Net-GAN architecture attains near-complete sensitivity (mean recall ≈ 0.96) but exhibits substantially reduced precision due to elevated false alarm rates. N-GAN demonstrates intermediate performance with greater variability across initializations, reflecting sensitivity to adversarial training dynamics. These findings highlight an intrinsic stability–sensitivity trade-off in GAN-based anomaly-detection frameworks. The study provides empirical evidence that architectural design fundamentally determines error structure and operational suitability, offering practical insights for deployment-oriented anomaly detection in time series environments.

Index Terms— Adversarial learning, anomaly detection, generative adversarial networks (GANs), reconstruction-based models, time series analysis

I. INTRODUCTION

The rapid expansion of Internet of Things (IoT) ecosystems and large-scale network infrastructures has significantly increased the complexity of cybersecurity monitoring processes. As heterogeneous devices continuously generate high-volume traffic streams, detecting malicious or abnormal behavior in a timely and reliable manner has become a critical requirement. Traditional signature-based intrusion detection mechanisms often struggle against previously unseen or evolving threats, which has led to growing interest in anomaly-based detection approaches capable of identifying deviations from learned normal behavior [1, 2].

Recent advances in deep learning have further strengthened anomaly-detection research by enabling the modeling of complex and high-dimensional data distributions. Among these approaches, generative adversarial networks (GANs) have attracted considerable attention due to their ability to approximate the underlying distribution of normal data through adversarial training. By learning the manifold of normal behavior, GAN-based architectures can detect anomalies as instances that significantly deviate from this learned representation.

Several GAN-based models have been proposed for network anomaly detection, including N-GAN [3], Net-GAN [4], and GANAD [5]. These architectures adopt different anomaly scoring mechanisms, ranging from reconstruction-based error measurements to discriminator-driven evaluation strategies. Although adversarial learning offers strong representational capabilities, it may also introduce instability, sensitivity to hyperparameters, and imbalanced detection behavior. Increased detection sensitivity may lead to elevated false-positive rates, which directly affect operational reliability in practical deployments.

Motivated by these considerations, this study presents a comparative evaluation of three representative GAN-based anomaly-detection architectures, GANAD, N-GAN, and Net-GAN, under consistent experimental conditions. The models are assessed on a benchmark time series anomaly dataset using unified evaluation criteria to ensure methodological consistency.

The main contributions of this study can be summarized as follows:

- A systematic comparative evaluation of three GAN-based anomaly-detection architectures (GANAD, N-GAN, and Net-GAN) is conducted under consistent experimental settings.
- A unified F1-optimized threshold selection strategy is implemented to provide a fair and comparable performance assessment framework.
- An empirical analysis of detection sensitivity and false-positive behavior in adversarial anomaly-detection models is presented.
- The study highlights stability differences between reconstruction-based and discriminator-driven anomaly scoring mechanisms in time series environments.

Through this analysis, the study aims to provide a clearer understanding of the stability–sensitivity trade-offs inherent in GAN-based anomaly-detection frameworks.

The paper is structured as follows: Section II reviews related work on GAN-based anomaly detection in time series settings. Section III describes the proposed methodological framework, including problem formulation, dataset preparation, model architectures, and the unified threshold calibration strategy. Section IV presents the experimental results and comparative performance analysis. Finally, Section V concludes the study and discusses implications for practical deployment and future research directions.

II. RELATED WORK

Anomaly detection in networked and IoT environments has been extensively studied over the past decade, with a growing emphasis on deep learning and generative modeling techniques. In particular, the application of GANs to intrusion detection and time series anomaly detection has attracted increasing research attention. Existing studies vary in terms of architectural design, anomaly scoring mechanisms, dataset selection, and evaluation strategies. To provide a structured overview, the related literature can be categorized into deep learning-based intrusion detection approaches, GAN-based anomaly-detection architectures, GAN applications in network and time series security contexts, and studies addressing stability and false-positive behavior in adversarial frameworks.

Anomaly-based intrusion detection has increasingly relied on deep learning techniques to model complex traffic patterns and detect deviations from normal behavior. Deep neural networks have been successfully applied to network intrusion detection systems (IDS) due to their capability to handle high-dimensional data and nonlinear relationships [1, 2]. Learning-based models have demonstrated improved adaptability compared to traditional signature-based systems, especially in dynamic and evolving threat environments.

Recent studies have also emphasized the importance of benchmark datasets and evaluation methodologies in IoT and network security research. Surveys on IoT security datasets and anomaly-detection mechanisms highlight the need for robust and standardized evaluation frameworks to ensure reliable performance comparison across models [6, 7]. These studies underline that evaluation consistency plays a critical role in assessing anomaly-detection performance in cybersecurity contexts.

Generative adversarial networks have emerged as a prominent generative modeling framework for anomaly detection. Early adversarial anomaly-detection approaches, such as efficient GAN-based anomaly detection and adversarially learned anomaly detection, demonstrated that GANs can learn the distribution of normal data and identify anomalies through reconstruction or discriminator-based scoring mechanisms [8, 9]. These studies laid the foundation for adversarial anomaly detection by combining generative modeling with discriminative evaluation.

Subsequent research explored various architectural adaptations to improve anomaly-detection performance. BiGAN-based anomaly-detection approaches introduced bidirectional mapping between latent and data spaces to enhance reconstruction consistency [10]. End-to-end adversarial intrusion detection frameworks extended GAN learning directly to network traffic analysis tasks [11]. Other studies addressed class imbalance problems in intrusion detection using GAN-based augmentation and adversarial training strategies [12, 13].

More recently, dedicated GAN-based architectures specifically designed for network anomaly detection have been proposed. N-GAN introduces an anomaly-based network intrusion detection mechanism that leverages adversarial learning to model traffic distributions [3]. Net-GAN extends GAN modeling to multivariate time series analysis for network anomaly detection [4]. GANAD proposes a reconstruction-driven adversarial framework tailored for network anomaly-detection tasks [5]. These models differ primarily in their anomaly scoring strategies and training dynamics.

The application of GAN-based models has expanded into IoT and industrial network environments. Fog-based and edge-aware adversarial detection systems have been proposed to address distributed intrusion detection challenges in cyber-physical systems [14]. Hybrid architectures integrating GANs with additional learning mechanisms have also been introduced to enhance detection sensitivity in IoT traffic anomaly scenarios [15].

Recent systematic reviews and surveys further demonstrate the growing interest in generative models for cybersecurity applications. Comprehensive analyses of GAN-based IDS identify architectural diversity, dataset variability, and evaluation heterogeneity as major challenges in the field [16, 17]. Contemporary reviews emphasize the necessity of stability analysis and fair comparative evaluation to better understand the operational reliability of adversarial anomaly-detection frameworks [18].

Despite their modeling strength, GAN-based architectures are known to suffer from training instability, hyperparameter sensitivity, and convergence difficulties. In anomaly-detection contexts, these issues may manifest as unstable reconstruction behavior or excessive false-positive rates. Studies addressing imbalanced intrusion detection scenarios highlight how adversarial sensitivity can amplify false alarm generation [12, 13].

Recent works have proposed modified adversarial training schemes, regularization strategies, and hybrid architectures to mitigate instability and improve robustness [15, 18]. However, many studies focus primarily on detection accuracy improvements or dataset-specific optimizations, while fewer works systematically compare different GAN-based anomaly-detection architectures under consistent experimental conditions.

Although numerous GAN-based anomaly-detection models have been proposed, existing literature often evaluates architectures under varying experimental settings, different datasets, or inconsistent thresholding strategies. This heterogeneity makes it difficult to objectively assess the relative stability and sensitivity characteristics of different adversarial frameworks.

This study addresses this gap by conducting a controlled comparative evaluation of three representative GAN-based anomaly-detection architectures—GANAD, N-GAN, and Net-GAN—under consistent experimental conditions and a unified performance assessment framework. By applying a common threshold optimization strategy and standardized evaluation metrics on the same benchmark dataset, the study aims to isolate architectural behavior and provide empirical insight into the trade-off between detection sensitivity and false-positive generation in adversarial anomaly-detection systems. Through this focused comparative analysis, the study contributes to a clearer understanding of the operational stability of GAN-based models in time series anomaly-detection environments.

III. METHODOLOGY

This section presents the methodological framework adopted for the comparative evaluation of GANAD, N-GAN, and Net-GAN. While all three architectures are rooted in adversarial learning, they operationalize anomaly detection through different scoring mechanisms and modeling assumptions. To ensure that the comparison reflects architectural characteristics rather than experimental variability, a unified preprocessing pipeline, consistent training configuration, and common threshold calibration strategy were employed. The following subsections detail the problem definition, the dataset preparation, model formulations, anomaly scoring procedures, and evaluation protocol under which the three approaches were systematically assessed.

A. Problem Definition

Anomaly detection in time series data differs fundamentally from classical supervised classification problems. In practical monitoring environments, anomalous events are rare and typically embedded within long stretches of normal operational behavior. Moreover, anomalies may not appear as isolated outliers but rather as structural or contextual deviations from learned temporal patterns. These characteristics make direct discriminative modeling difficult, particularly under severe class imbalance.

Let the observed univariate time series be defined as:

$$X = \{x_t \in \mathbb{R}\}_{t=1}^{T} \tag{1}$$

where x_t denotes the observed value at time step . Each observation is associated with a binary ground-truth label:

$$y_t \in \{0\ 1\} \tag{2}$$

where $y_t = 0$ represents normal behavior and $y_t = 1$ represents an anomaly.

In benchmark anomaly datasets such as Numenta Anomaly Benchmark (NAB), the proportion of anomalous instances satisfies:

$$P(y_t = 1) \ll P(y_t = 0) \tag{3}$$

Under this imbalance condition, overall accuracy becomes a misleading performance indicator. Instead, anomaly-detection models are typically formulated to produce a continuous anomaly scoring function:

$$S_\theta(x_t) : \mathbb{R} \to \mathbb{R} \tag{4}$$

such that anomalous observations receive higher scores than normal instances. Within this formulation, the objective of learning is not direct class prediction but the estimation of deviations from the learned distribution of normal temporal behavior.

In GAN-based anomaly detection, adversarial learning aims to capture the underlying distributional characteristics of normal temporal behavior. Deviations from this learned representation are subsequently quantified through model-specific scoring mechanisms.

B. Dataset and Preprocessing

The empirical analysis was conducted on the NAB artificial time series dataset, specifically the art_load_balancer_spikes file. This dataset contains a single continuous measurement variable accompanied by timestamp information and annotated anomaly intervals.

The dataset was selected for three principal reasons. First, its anomaly annotations are well-defined and publicly standardized, allowing reproducible evaluation. Second, anomalies appear as clearly localized spike patterns embedded within otherwise stable temporal behavior, making it suitable for analyzing detection sensitivity. Third, its controlled structure minimizes confounding factors introduced by high-dimensional network intrusion datasets, thereby enabling a focused examination of anomaly scoring mechanisms.

1) Data Formatting:

Timestamp values were converted into the datetime format and used as the time index to preserve chronological ordering. Missing observations, if present, were handled using forward-fill imputation to preserve temporal continuity. Only the measurement variable was retained for modeling. Since the study focuses on anomaly scoring behavior rather than feature engineering, no additional derived attributes were introduced.

2) Standardization:

Generative adversarial network–based architectures are sensitive to input scale due to the adversarial optimization dynamics. Therefore, the signal was standardized using z-score normalization defined in (5).

$$z_t = \frac{x_t - \mu}{\sigma} \quad (5)$$

where μ and σ denote the mean and standard deviation of the series.

This transformation ensures numerical stability during training and prevents scale-dominant behavior from biasing reconstruction error or discriminator outputs.

3) Temporal Handling Strategy:
Given the localized spike characteristics of anomalies in the selected NAB series, the data were processed as a standardized univariate time series without applying explicit sliding-window segmentation. In other words, each time step z_t was evaluated directly in its sequential position rather than being embedded into overlapping temporal windows.

This decision was made deliberately. The primary objective of this study is to compare anomaly scoring behaviors across GAN-based architectures, not to investigate temporal feature engineering strategies. By avoiding additional windowing mechanisms, the analysis isolates the scoring logic of each model. As a result, any observed performance differences can be attributed to the internal anomaly scoring dynamics of GANAD, N-GAN, and Net-GAN, rather than to variations in contextual window design.

This point-wise formulation was intentionally adopted to isolate anomaly score dynamics rather than temporal embedding strategies. Introducing sliding-window segmentation could confound architectural comparison through window-size effects and contextual overlap, which are outside the scope of the present structural analysis. Because anomalies in the selected dataset appear as localized spikes, point-wise evaluation of observations may also affect recall by limiting the model's ability to exploit the short temporal context that sliding-window representations could capture.

4) Model Formulations:
Although GANAD, N-GAN, and Net-GAN are all derived from the generative adversarial learning paradigm, they differ substantially in how anomaly information is extracted from the learned representation. The distinction does not lie merely in architectural depth or optimization strategy, but in the operational definition of what constitutes an anomaly. For this reason, the present study places particular emphasis on clarifying how each model translates adversarial training into a measurable anomaly score.

In general, GAN-based anomaly-detection methods attempt to learn the distribution of normal data through adversarial optimization. Once this distribution is approximated, anomalies are interpreted as samples that deviate from the learned manifold. However, the mechanism through which this deviation is quantified varies across architectures.

GANAD follows a reconstruction-centered interpretation of anomaly detection. After adversarial training aligns the generator with the distribution of normal observations, the model produces reconstructed outputs through an encoder–generator pipeline (6):

$$\check{x}_t = G_\theta\left(E_\phi\left(x_t\right)\right) \quad (6)$$

Here, E denotes the encoder mapping the input to a latent representation, and G_θ denotes the generator acting as a decoder.

The anomaly score is defined as the reconstruction residual in (7):

$$S\left(x_t\right) = x_t - \check{x}_t^{\,2} \quad (7)$$

This formulation assumes that normal observations lie close to the learned data manifold and can therefore be reconstructed with low error. In contrast, anomalous samples are expected to produce larger reconstruction deviations because they are poorly represented by the learned latent structure.

A key advantage of this approach is interpretability. The anomaly score directly corresponds to reconstruction discrepancy. However, its effectiveness depends on the stability of the encoder–generator alignment and the model's ability to avoid trivial identity mapping.

N-GAN extends the classical adversarial framework by incorporating an explicit feature transformation stage. Instead of relying solely on raw reconstruction error, the model evaluates deviations within learned latent representations before mapping back to the input space.

Let h_t denote the extracted latent feature vector:

$$h_t = F_\phi\left(x_t\right) \quad (8)$$

where F represents the feature extraction component.

The anomaly score is then computed as:

$$S\left(x_t\right) = \| x_t - G_\theta\left(h_t\right) \|^2 \quad (9)$$

While the mathematical expression resembles reconstruction error, the conceptual difference lies in the intermediate feature learning stage. By introducing structured latent representations, N-GAN attempts to capture higher-level regularities of normal behavior. Consequently, anomaly sensitivity may increase for structural deviations rather than purely amplitude-based discrepancies. However, this additional representation layer may also introduce variability in scoring, particularly if adversarial convergence is unstable.

In contrast to reconstruction-based paradigms, Net-GAN derives anomaly likelihood directly from discriminator behavior. Instead of measuring reconstruction deviation, the model interprets discriminator confidence as a proxy for normality.

Let $D_\psi\left(x_t\right) \in [0,1]$ denote the discriminator's estimated probability that the observation belongs to the learned normal distribution. The anomaly score is defined as:

$$S\left(x_t\right) = 1 - D_\psi\left(x_t\right) \quad (10)$$

Under this formulation, samples assigned low discriminator confidence are considered anomalous.

This approach eliminates the need for explicit reconstruction and may enhance sensitivity to subtle distributional shifts. However, discriminator outputs are known to be sensitive to instability in adversarial training and to mode collapse. As a result, discriminator-driven scoring can achieve higher recall but may also increase false-positive rates, particularly in imbalanced settings.

From a methodological perspective, the three architectures represent distinct anomaly scoring paradigms: reconstruction residual

(GANAD), feature-mediated deviation (N-GAN), and discriminator-driven confidence (Net-GAN). By evaluating these scoring mechanisms under identical preprocessing and calibration conditions, the study isolates how differences in anomaly score formulation translate into measurable detection behavior.

Rather than treating GAN-based anomaly detection as a single homogeneous category, this formulation enables a controlled comparison of how architectural design choices influence sensitivity, stability, and false-positive characteristics.

5) Experimental Configuration:
To ensure that the comparative analysis reflects architectural characteristics rather than implementation variability, all three models were trained under a controlled and unified experimental configuration. Except for architecture-specific design elements, hyperparameters and optimization strategies were maintained as consistently as possible across models.

The dataset, input structure, training duration, and threshold calibration procedure were aligned to minimize experimental confounders. Where architectural differences necessitated parameter adjustments (e.g., batch size or learning rate), such variations were explicitly preserved to respect the intrinsic modeling assumptions of each approach. In particular, the GANAD implementation follows the original TensorFlow-based configuration reported in the reference implementation, which uses the default Adam learning rate of 0.001. In contrast, the PyTorch implementations of N-GAN and Net-GAN employ a learning rate of 0.0002, consistent with their published experimental settings.

As the evaluated architectures follow an unsupervised anomaly-detection paradigm, model training was performed on the standardized time series without predefined training–testing splits. This design choice reflects realistic deployment conditions, in which anomaly-detection systems operate over continuous data streams with sparse anomaly occurrences.

Ground-truth annotations were used solely for performance evaluation and threshold calibration, and were not involved in parameter optimization. Threshold calibration was performed after model training was completed and did not influence parameter learning. The ground-truth labels were used exclusively to determine the F1-optimized operating point for each trained model. Because the primary objective of this study is architectural comparison rather than deployment-oriented generalization estimation, the identical calibration protocol was uniformly applied across all models to ensure fair structural evaluation without introducing model-specific bias.

To further ensure fair comparison, identical data preprocessing, anomaly score computation pipelines, and evaluation metrics were applied across all models. No architecture-specific post-processing or score normalization techniques were introduced, so that performance differences reflect intrinsic model behavior rather than auxiliary tuning strategies.

The unified experimental configuration applied to GANAD, N-GAN, and Net-GAN is summarized in Table I. For robustness, results were additionally averaged over ten random seeds (42–51 inclusive) using the same configuration.

6) Unified Threshold Calibration:
A critical but often underemphasized component of anomaly-detection systems is the decision threshold. While anomaly scores provide a continuous indication of deviation, practical deployment requires a binary decision boundary. In many studies, threshold values are determined heuristically. Although such rules may be convenient, they introduce an additional source of variability when comparing architectures that produce anomaly scores with fundamentally different statistical distributions.

TABLE I. EXPERIMENTAL CONFIGURATION OF GENERATIVE ADVERSARIAL NETWORK–BASED MODELS

Component	GANAD	N-GAN	Net-GAN
Dataset	NAB	NAB	NAB
Input type	Univariate	Univariate	Univariate
Temporal handling	No windowing (point-wise)	No windowing (point-wise)	No windowing (point-wise)
Epochs	300	300	300
Batch size	256	64	64
Optimizer	Adam	Adam	Adam
Learning rate	Adam default (0.001) (TensorFlow)	0.0002	0.0002
Loss function	MSE (reconstruction)	MSE (reconstruction) + 0.1·BCE (adv term) (plus BCE for discriminator training)	BCE (adversarial)
Anomaly score	Reconstruction MSE	Reconstruction MSE	−log(D(x)) (discriminator-based)
Thresholding	F1-optimized via linear grid search	F1-optimized via linear grid search	F1-optimized via linear grid search
Random seeds	Ten seeds (for stability reporting)	Ten seeds (for stability reporting)	Ten seeds (for stability reporting)

GAN, generative adversarial network; NAB, Numenta Anomaly Benchmark.

Because GANAD, N-GAN, and Net-GAN generate anomaly scores with different scaling behaviors and dispersion characteristics, using model-specific heuristic thresholds would risk conflating architectural differences with threshold selection bias. To prevent this, a unified, metric-driven calibration strategy was employed.

Let $S(x_t)$ denote the anomaly score sequence produced by a given model. For each candidate threshold τ, binary predictions are generated as (11):

$$\breve{y}_t(\tau) = \begin{cases} 1, & \text{if } S(x_t) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Instead of fixing τ heuristically, thresholds were evaluated over the empirical score distribution. The optimal threshold was selected as:

$$\tau^* = \arg\max_{\tau} F1\left(y, \breve{y}(\tau)\right) \tag{12}$$

This selection criterion directly optimizes the harmonic balance between precision and recall, which is particularly appropriate under severe class imbalance. Unlike accuracy, which can be artificially inflated by dominant normal samples, the F1-score penalizes both false positives and false negatives in a balanced manner.

Importantly, this threshold calibration procedure was applied independently but identically to all three models. The sweep was conducted over candidate thresholds derived from the observed anomaly score range of each model, ensuring that the decision boundary adapts to score distribution characteristics without privileging any architecture. In practice, candidate thresholds were generated via a linear grid search spanning the full anomaly score range (from minimum to maximum), and the threshold maximizing the F1-score was retained for each model to ensure consistent and reproducible calibration.

This unified calibration strategy ensures that performance comparisons are not influenced by arbitrary or architecture-specific decision heuristics. Instead, differences in detection outcomes are primarily attributable to the intrinsic properties of the anomaly scoring mechanisms themselves. By standardizing the threshold selection procedure, the analysis focuses on scoring stability and sensitivity rather than post-hoc tuning effects.

IV. RESULTS AND DISCUSSION

A. Overall Performance Comparison Under Unified Calibration

Table II presents the aggregated anomaly-detection performance of GANAD, N-GAN, and Net-GAN over ten independent random seeds under a unified F1-optimized threshold calibration protocol. The reported threshold values correspond to the raw anomaly score distributions produced by each model after input standardization, and therefore may differ in magnitude depending on the scoring formulation of the architecture. For completeness, the median optimal threshold across seeds is also reported to provide a scale reference for the anomaly score distributions. This value serves descriptive purposes only and does not influence comparative ranking or interpretation. The reported values correspond to mean ± standard deviation, thereby reflecting not only central tendency but also the stability of each architecture across different initialization conditions.

Among the evaluated models, GANAD demonstrates the most stable behavior. Its performance metrics remain effectively invariant across seeds, yielding a mean precision of 0.5222, a recall of 0.2333, and an F1-score of 0.3225 with negligible variance. This indicates that the reconstruction-based formulation of GANAD converges to a consistent decision boundary under the standardized experimental setting. While its recall is relatively modest, the model achieves the highest mean F1-score overall, reflecting a balanced trade-off between false positives and missed anomalies. N-GAN exhibits intermediate performance with noticeable variability across seeds. Its mean F1-score reaches 0.2677 with a standard deviation of 0.0463, suggesting sensitivity to random initialization. The model achieves a higher average recall than GANAD (0.2938), but this improvement is accompanied by substantial fluctuations in predicted anomaly counts, as reflected in the large standard deviation of predicted anomaly points. This behavior indicates that N-GAN may converge to either relatively conservative or overly aggressive detection regimes depending on adversarial training dynamics. In contrast, Net-GAN consistently favors high sensitivity at the expense of precision. Across seeds, it achieves a mean recall of 0.9635, substantially higher than the other models, but with a low mean precision of 0.0969. The resulting mean F1-score of 0.1761 confirms that the model tends to classify a large proportion of the time series as anomalous. Although recall remains high and relatively stable, the excessive number of predicted anomaly points indicates a strong bias toward over-detection.

The comparative trends summarized numerically in Table II are visually illustrated in Fig. 1. While GANAD consistently achieves the highest precision with negligible variance, Net-GAN prioritizes recall, resulting in substantially higher sensitivity but reduced selectivity. N-GAN exhibits intermediate behavior, with moderate precision and recall values accompanied by greater variability across seeds. The error bars in Fig. 1 highlight that architectural design choices influence not only detection performance but also robustness under repeated initialization.

The stability of overall detection quality is further reflected in Fig. 2, which reports mean F1-scores across seeds. GANAD demonstrates

TABLE II. COMPARATIVE ANOMALY-DETECTION PERFORMANCE OF GANAD, N-GAN, AND NET-GAN

Model	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)	Optimal Threshold (Median)	Predicted Anomaly Points (Mean ± SD)
GANAD	0.5222 ± 0.0000	0.2333 ± 0.0000	0.3225 ± 0.0000	1.954e+00	180 ± 0
N-GAN	0.3728 ± 0.1364	0.2938 ± 0.2490	0.2677 ± 0.0463	3.291e-07	608 ± 1204
Net-GAN	0.0969 ± 0.0050	0.9635 ± 0.0595	0.1761 ± 0.0093	1.197e-01	4006 ± 41

GAN, generative adversarial network.

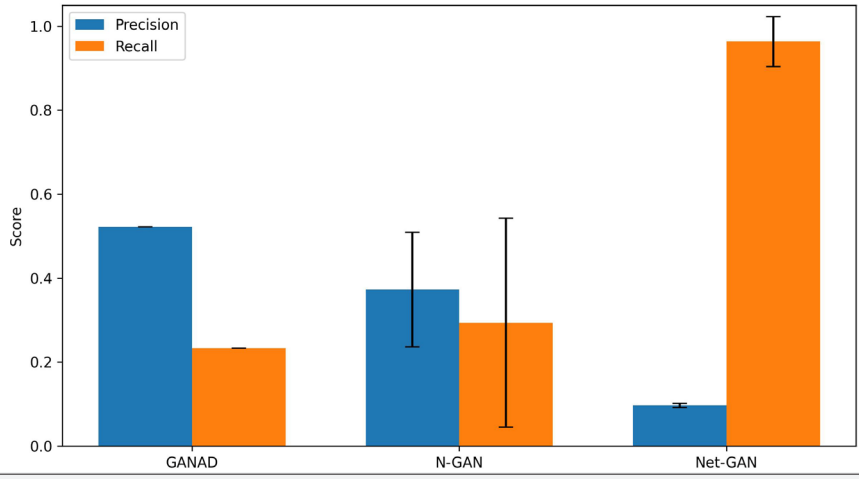

Fig. 1. Mean precision and recall across ten random seeds for the evaluated architectures.

virtually zero variance, indicating deterministic convergence under the present configuration. In contrast, N-GAN shows moderate dispersion, suggesting sensitivity to adversarial training dynamics. Net-GAN maintains relatively stable but consistently lower F1-scores, reflecting its strong bias toward high-recall detection.

Taken together, the results indicate three distinct detection profiles: a conservative and stable reconstruction-based approach (GANAD), a seed-sensitive hybrid adversarial-reconstruction architecture (N-GAN), and an aggressively recall-oriented discriminator-driven model (Net-GAN). These differences underscore that GAN-based anomaly-detection frameworks exhibit not only performance trade-offs but also markedly different robustness characteristics under identical experimental conditions.

B. Precision Characteristics of the Evaluated Architectures

Precision values reported in Table II reveal substantial differences in how the evaluated architectures control false positives under identical experimental conditions. Unlike recall-oriented behavior, which reflects anomaly sensitivity, precision directly captures the selectivity of each model and its tolerance for over-detection.

Among the three approaches, GANAD consistently achieves the highest precision, with a mean value of 0.5222 and virtually zero variance across seeds. This stability indicates that the reconstruction-based decision mechanism produces a well-defined and reproducible separation between nominal and anomalous observations. In practical terms, GANAD tends to generate fewer false alarms, reflecting a conservative detection strategy grounded in reconstruction deviation rather than adversarial discrimination.

In contrast, Net-GAN exhibits the lowest precision, averaging 0.0969 across seeds. This result is not due to instability, but rather to a systematic bias toward high recall. As previously observed, the discriminator-based anomaly score encourages widespread anomaly labeling once the threshold is optimized for F1. Consequently, although most true anomalies are detected, a substantial portion of nominal points is also flagged, leading to precision degradation.

N-GAN demonstrates intermediate precision behavior, with a mean value of 0.3728 but noticeably higher standard deviation. This variability reflects the sensitivity of hybrid adversarial-reconstruction training dynamics to initialization conditions. While certain seeds yield precision values comparable to GANAD, others converge to

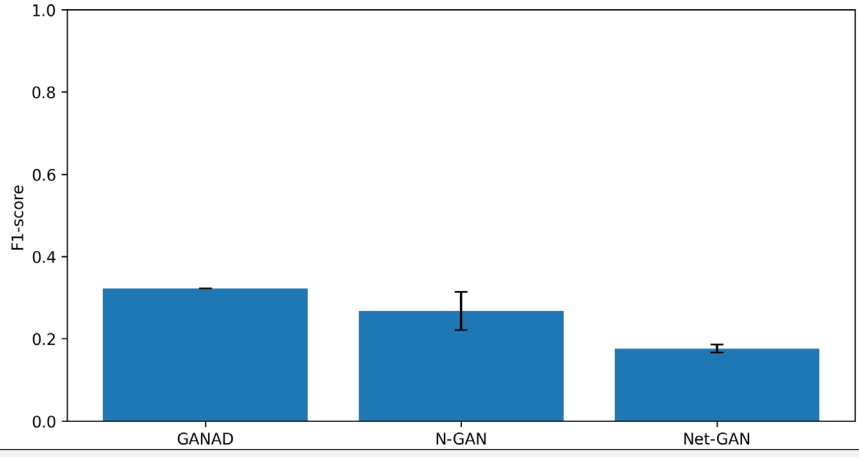

Fig. 2. Mean F1-score of each model across ten random seeds.

more aggressive regimes. This dispersion indicates that adversarial components introduce additional degrees of freedom in decision boundary formation.

The precision trends are visually confirmed in Fig. 2, where GANAD maintains the highest and most stable selectivity profile, N-GAN exhibits moderate but variable performance, and Net-GAN remains consistently low in precision despite its strong recall performance.

Overall, these findings suggest that architectural emphasis on reconstruction fidelity promotes precision stability, whereas discriminator-driven anomaly scoring prioritizes sensitivity at the expense of false alarm control.

C. Sensitivity Behavior and False-Positive Trade-offs

Sensitivity, measured through recall, reflects the ability of a model to capture true anomalous events within the time series. While high recall is generally desirable in safety-critical monitoring settings, it often comes at the cost of increased false-positive rates. The results summarized in Table II illustrate that the evaluated architectures adopt markedly different trade-off strategies in this respect.

Across ten random seeds, Net-GAN exhibits the highest sensitivity, with a mean recall of 0.9635. This indicates that nearly all anomaly points are detected under F1-optimized threshold calibration. However, this aggressive detection behavior leads to extensive over-labeling of nominal observations, as reflected in the consistently high number of predicted anomaly points. The low precision reported in Table II confirms that Net-GAN prioritizes coverage over selectivity. In operational environments, such behavior may induce alarm fatigue and reduce the practical usability of the detector despite its high sensitivity.

In contrast, GANAD demonstrates a substantially more conservative sensitivity profile, achieving a mean recall of 0.2333. Although this indicates that a considerable portion of anomaly windows remains partially undetected, the model maintains significantly higher precision. This behavior suggests that reconstruction-based anomaly scoring imposes a stricter deviation requirement before labeling a point as anomalous. Consequently, GANAD sacrifices detection breadth in exchange for improved reliability of flagged events.

N-GAN occupies an intermediate position, with a mean recall of 0.2938 but a notably large standard deviation. This variability reflects seed-dependent convergence patterns in adversarial training. In certain runs, N-GAN behaves conservatively, whereas in others it approaches the aggressive regime observed in Net-GAN. The dispersion in predicted anomaly counts further indicates instability in the learned decision boundary when adversarial components influence reconstruction dynamics.

The trade-offs described above are visually apparent in Fig. 1, where Net-GAN's recall dominance contrasts sharply with its limited precision. This inverse relationship highlights a fundamental tension in GAN-based anomaly detection: architectures emphasizing discriminator-driven scoring tend to maximize anomaly coverage, while reconstruction-focused models constrain detection to stronger deviations.

The trade-off between sensitivity and selectivity is more clearly observed in Fig. 3, which maps each model in the precision–recall space using mean values across ten seeds. Net-GAN occupies the high-recall, low-precision region, reflecting its tendency toward aggressive anomaly labeling. GANAD, by contrast, is positioned in the higher-precision, lower-recall region, consistent with its conservative reconstruction-based strategy. N-GAN appears between these two extremes, although its relative dispersion across seeds indicates reduced stability. The spatial separation shown in Fig. 3 confirms that architectural design fundamentally determines the balance between anomaly coverage and false-positive control.

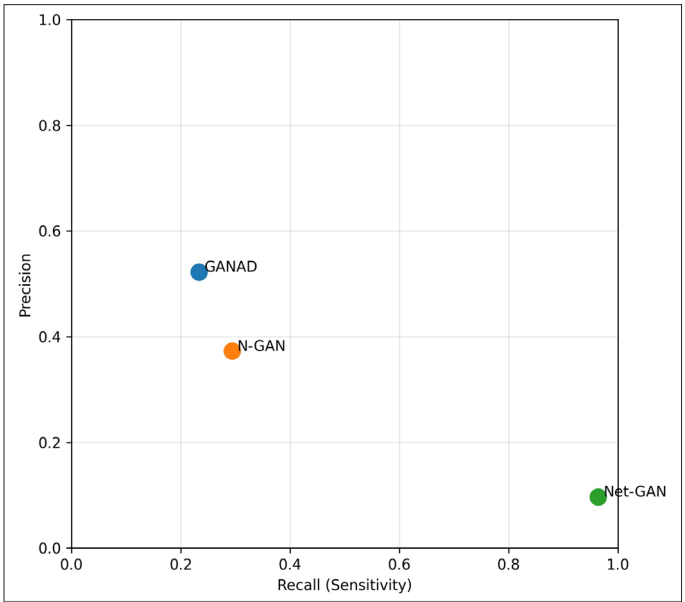

Fig. 3. Precision–recall positioning of the evaluated architectures (mean over ten seeds).

It is important to emphasize that this precision–recall positioning is not the result of model-specific threshold tuning, but rather emerges from the unified F1-optimized calibration protocol applied consistently across architectures. Because the same optimization objective and threshold search strategy were enforced for all models, the observed separation in the precision–recall space reflects intrinsic anomaly score dynamics rather than post-hoc decision bias.

The results suggest that sensitivity must be interpreted jointly with false-positive behavior. A high recall value, in isolation, does not guarantee operational effectiveness. Instead, robustness, selectivity, and anomaly prevalence alignment must be considered simultaneously when evaluating practical deployment suitability.

D. Overall Detection Balance Under a Unified Decision Criterion

To examine how the evaluated architectures behave when the same decision rule is imposed, they are compared under a unified F1-optimized threshold calibration. This setting is intentionally strict: it removes ad-hoc, model-specific threshold heuristics and forces each method to operate under an identical optimization target. In practice, this makes the comparison less about "how well a threshold can be tuned" and more about how each scoring mechanism naturally separates anomalous and normal regions in the score distribution.

The results summarized in Table II indicate that the reconstruction-driven approach (GANAD) tends to deliver a more selective detection profile, typically producing fewer alarms and achieving higher

precision, albeit at the cost of missing portions of the labeled anomaly windows. In contrast, discriminator-driven scoring (Net-GAN) consistently prioritizes sensitivity: recall increases substantially, but this is accompanied by a marked inflation in false positives, which compresses precision and limits the attainable F1 under the same calibration objective. The feature-mediated architecture (N-GAN) often sits between these extremes in principle; however, its stability across runs is a critical factor, since small shifts in the learned representation can change the tail behavior of the anomaly score distribution and, consequently, the calibrated operating point.

This balance is shown in Fig. 4, where the mean precision, recall, and F1 values across multiple random seeds are plotted together with their variability. The figure highlights that the "best" model under a unified criterion is not merely the one with the highest sensitivity, but the one whose score dynamics produce a usable separation between normal behavior and anomaly windows without triggering excessive false alarms. In deployment-oriented anomaly detection, this distinction is central: a detector that saturates the alert channel may look strong on recall, yet remain operationally impractical.

E. Error Structure and Operational Implications

While summary scores (precision, recall, F1) are convenient for ranking models, they can hide how each architecture fails. For operational monitoring, the structure of errors matters at least as much as the aggregate metric: frequent false alarms lead to alert fatigue and desensitization, whereas missed anomalies (false negatives) may translate into delayed incident response or undetected degradation. In the current setting, where thresholds are calibrated under the same F1-optimized protocol, the remaining differences largely reflect the intrinsic behavior of each anomaly score, rather than post-hoc decision tuning.

Across the evaluated models, the error profiles align with the scoring philosophy. Reconstruction-driven scoring (GANAD) typically concentrates errors on the miss side: it tends to avoid flooding the timeline with alarms, but it can fail to flag parts of anomaly windows when spikes are short-lived, weakly reconstructed, or partially aligned with the learned manifold. Discriminator-driven scoring (Net-GAN), by contrast, often shifts the burden toward false positives: it is sensitive to distributional shifts and local deviations, yet that sensitivity may also elevate normal fluctuations into alarms, particularly when the discriminator output is not sharply separated for normal versus anomalous regions. Feature-mediated approaches (N-GAN) can behave closer to either side depending on training stability; when the learned representation compresses normal variability effectively, false alarms may be contained, but if the score distribution develops a heavy tail, the calibrated threshold can move toward an aggressive operating point.

The aggregated confusion matrices presented in Fig. 5 reveal that the dominant error mode differs substantially across architectures, with reconstruction-based scoring primarily increasing false negatives, whereas discriminator-driven scoring amplifies false positives under identical calibration constraints.

The imbalance in false-positive behavior becomes more apparent when normalized rates are examined. As illustrated in Fig. 6, the false-positive rate differs substantially across architectures under identical calibration constraints. Discriminator-driven scoring produces the highest false-positive burden, reflecting its aggressive sensitivity profile, whereas reconstruction-based scoring maintains a markedly lower rate of spurious alarms.

Conversely, the false-negative rate presented in Fig. 7 highlights the complementary trade-off. Models that constrain false positives tend to incur a higher proportion of missed anomalies, whereas architectures optimized for maximum coverage reduce false negatives at the expense of increased alarm volume. This structural asymmetry reinforces that model selection should be aligned with the operational cost of missed events versus excessive alerts.

Taken together, Figs. 5–7 demonstrate that performance differences are not merely quantitative but structural. The dominant error mode, whether false alarms or missed detections, emerges as an intrinsic property of the anomaly scoring mechanism rather than a threshold artifact. Consequently, deployment decisions should consider not only F1-score rankings but also the operational tolerance for specific error types.

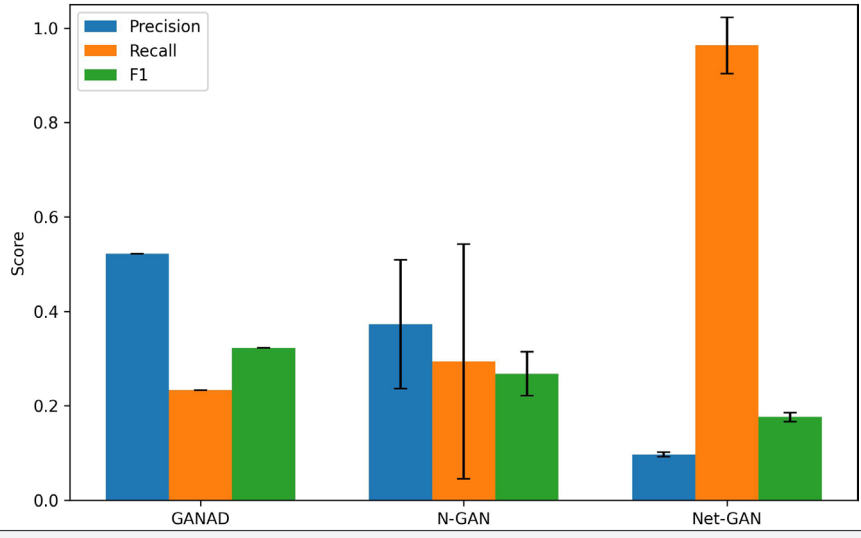

Fig. 4. Detection balance under F1-optimized thresholding (mean ± standard deviation).

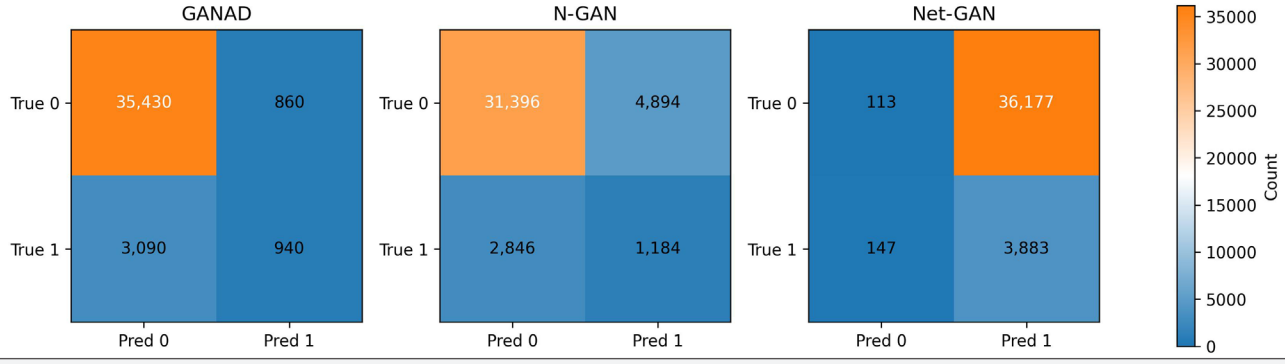

Fig. 5. Aggregated confusion matrices across ten random seeds.

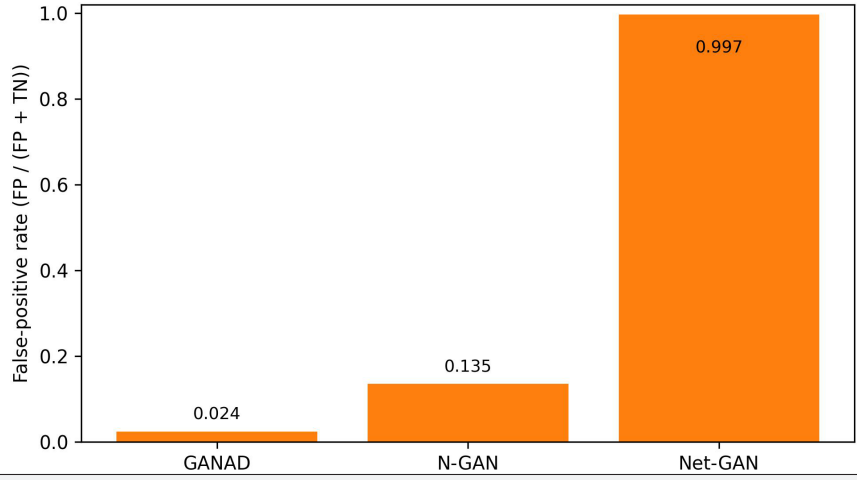

Fig. 6. False-positive rates across models derived from aggregated confusion matrices.

V. CONCLUSION

This study conducted a controlled comparison of three GAN-based anomaly-detection architectures, GANAD, N-GAN, and Net-GAN, under a unified F1-optimized threshold calibration protocol on the NAB artificial spike dataset. By enforcing identical preprocessing and decision criteria, the analysis isolated architectural effects from threshold tuning artifacts.

The findings indicate that architectural emphasis directly shapes detection balance. GANAD achieved the highest mean F1-score, approximately 0.32 across ten seeds, while maintaining stable

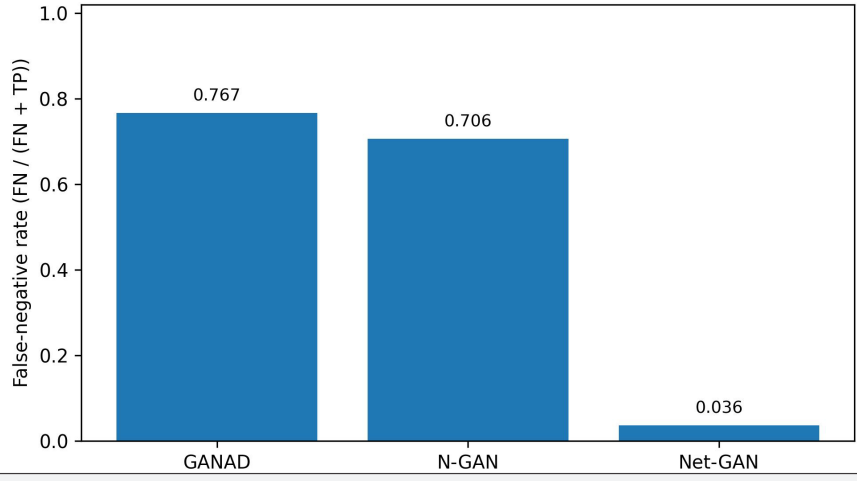

Fig. 7. False-negative rates across models derived from aggregated confusion matrices.

precision around 0.52. In contrast, Net-GAN reached the highest sensitivity, with a mean recall approaching 0.96, but at the cost of substantially reduced precision. N-GAN exhibited intermediate performance yet demonstrated greater variability across runs, highlighting sensitivity to adversarial training dynamics.

Structural error analysis further showed that reconstruction-driven scoring concentrates errors in missed anomaly segments, whereas discriminator-driven scoring shifts the burden toward false positives. These differences are not incidental but intrinsic to the anomaly scoring formulation.

Overall, the results demonstrate that GAN-based anomaly detection cannot be evaluated solely through recall or F1 metrics in isolation. Stability across seeds, dominant error mode, and operational tolerance to false alarms are equally decisive. Future research may extend this framework to multivariate settings, adaptive calibration schemes, and cost-sensitive optimization strategies that explicitly integrate deployment constraints into the learning objective. It should also be noted that the present evaluation is based on a single univariate benchmark series from the NAB dataset, and therefore, the observed detection behaviors may not fully generalize to multivariate or more complex temporal environments.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Artificial Intelligence Usage Statement: The authors declared that no Artificial Intelligence Tool was used in the preparation of the manuscript.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – B.Ç., Ş.D.O.; Design – B.Ç., Ş.D.O.; Supervision – B.Ç., Ş.D.O.; Resources – B.Ç., Ş.D.O.; Materials – B.Ç., Ş.D.O.; Data Collection and/or Processing – B.Ç., Ş.D.O.; Analysis and/or Interpretation – B.Ç., Ş.D.O.; Literature Search – B.Ç., Ş.D.O.; Writing – B.Ç., Ş.D.O.; Critical Review – B.Ç., Ş.D.O.

Declaration of Interests: The authors have no conflict of interest to declare.

Funding: The authors declared that this study has received no financial support.

REFERENCES

1. R. Mills, A. K. Marnerides, M. Broadbent, and N. Race, "Practical intrusion detection of emerging threats," *IEEE Trans. Netw. Serv. Manag.*, vol. 19, no. 1, pp. 582–600, 2022. [CrossRef]
2. D. Yuan et al., "Intrusion detection for smart home security based on data augmentation with edge computing," In *Proceedings of the IEEE International Conference on Communications (ICC)*. New York: IEEE, 2020, pp. 1–6. [CrossRef]
3. A. S. Iliyasu, and H. Deng, "N-GAN: A novel anomaly-based network intrusion detection with generative adversarial networks," *Int. J. Inf. Technol.*, vol. 14, no. 7, pp. 3365–3375, 2022. [CrossRef]
4. G. G. Gonzalez, P. Casas, A. Fernández, and G. Gomez, "Network anomaly detection with Net-GAN: A generative adversarial network for analysis of multivariate time series," In *Proceedings of the SIGCOMM 2020 Poster and Demo Sessions*. New York, NY, USA: ACM, 2020, pp. 62–64. [CrossRef]
5. J. Fu, L. Wang, J. Ke, K. Yang, and R. Yu, "GANAD: A GAN-based method for network anomaly detection," *World Wide Web*, vol. 26, No. 5, pp. 2727–2748, 2023. [CrossRef]
6. H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, Efficient GAN-based anomaly detection. arXiv preprint arXiv:1802.06222. 2018, Available: https://arxiv.org/abs/1802.06222.
7. H. Zenati, M. Romain, C. S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially learned anomaly detection," In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. New York: IEEE, 2018, pp. 727–736. [CrossRef]
8. M. O. Kaplan, and S. E. Alptekin, "An improved BiGAN-based approach for anomaly detection," *Procedia Comput. Sci.*, vol. 176, pp. 185–194, 2020. [CrossRef]
9. B. Mohammadi, and M. Sabokrou, "End-to-end adversarial learning for intrusion detection in computer networks," In *Proceedings of the IEEE Conference on Local Computer Networks (LCN)*. New York: IEEE, 2019, pp. 270–273. [CrossRef]
10. S. Huang, and K. Lei, "IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection systems in ad-hoc networks," *Ad Hoc Netw.*, vol. 105, 102177, 2020. [CrossRef]
11. J. H. Lee, and K. H. Park, "GAN-based imbalanced data intrusion detection system," *Pers. Ubiquitous Comput.*, vol. 25, no. 1, pp. 121–128, 2021. [CrossRef]
12. P. Freitas de Araujo-Filho, G. Kaddoum, D. R. Campelo, A. Gondim Santos, D. Macedo, and C. Zanchettin, "Intrusion detection for cyber-physical systems using generative adversarial networks in fog environments," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6247–6256, 2021. [CrossRef]
13. C. Alex, G. Creado, W. Almobaideen, O. A. Alghanam, and M. Saadeh, "A comprehensive survey for IoT security datasets taxonomy, classification and machine learning mechanisms," *Comput. Secur.*, vol. 132, 103283, 2023. [CrossRef]
14. S. H. Mekala, Z. Baig, A. Anwar, and S. Zeadally, "Cybersecurity for Industrial IoT (IIoT): Threats, countermeasures, challenges and future directions," *Comput. Commun.*, vol. 208, pp. 294–320, 2023. [CrossRef]
15. M. Arafah, I. Phillips, A. Adnane, W. Hadi, M. Alauthman, and A. K. Al-Banna, "Anomaly-based network intrusion detection using denoising autoencoder and Wasserstein GAN synthetic attacks," *Appl. Soft Comput.*, vol. 168, 112455, 2025. [CrossRef]
16. M. Alauthman, N. Aslam, A. Al-Qerem, A. Aldweesh, and P. Sureephong, "Generative adversarial networks for intrusion detection systems: A comprehensive survey of applications, challenges, and research directions," *Arab. J. Sci. Eng.*, vol. 51, No. 1, pp. 179–203, 2026. [CrossRef]
17. T. Ndayipfukamiye, J. Ding, D. S. Sarwatt, A. G. Philipo, and H. Ning, 2025. *Adversarial Defense in Cybersecurity: A Systematic Review of GANs for Threat Detection and Mitigation*. arXiv preprint arXiv:2509.20411.
18. F. Wang, Y. Jiang, R. Zhang, A. Wei, J. Xie, and X. Pang, "A survey of deep anomaly detection in multivariate time series: Taxonomy, applications, and directions," *Sensors (Basel)*, vol. 25, no. 1, 190, 2025. [CrossRef]



Begüm Çorbacioğlu received her B.Sc. degree in Mathematics from Istanbul University in 2020. She completed her M.Sc. in Cybersecurity at Bahçeşehir University in 2024. She has over four years of industry experience in IT security. Her academic interests include cybersecurity, IoT, and cloud computing.



Şafak Durukan-Odabaşi received the B.S., M.S., and Ph.D. degrees in Computer Engineering from Istanbul University, Istanbul, Türkiye, in 2005, 2008, and 2013. She worked as a Research Assistant and Assistant Professor of Computer Engineering at Istanbul University and Istanbul University-Cerrahpaşa between 2005-2019. She was a Visiting Assistant Professor at Illinois Wesleyan University, USA, from 2019 to 2021. She is currently working as an Assistant Professor at Istanbul University – Cerrahpaşa. Her research areas are next-generation networks, IoT, and cybersecurity.