https://doi.org/10.31891/2219-9365-2026-86-14
UDC 004.852:519.217:621.391.1

ZHYVYLO Yevhen
National University «Yuri Kondratyuk Poltava Polytechnic»
https://orcid.org/0000-0003-4077-7853
e-mail: zhivilka@i.ua
CHORNYI Anton
Limited Liability Company Private Higher Education Institution «University of Modern Technologies»
https://orcid.org/0000-0001-9857-2151
e-mail: chornyiav@gmail.com
ROMASHKO Ihor
National University «Yuri Kondratyuk Poltava Polytechnic»
https://orcid.org/0000-0002-1287-2864
e-mail: riwukr27@gmail.com
KALASHNIKOVA Yuliia
National University «Yuri Kondratyuk Poltava Polytechnic»
https://orcid.org/0000-0001-9899-4784
e-mail: kalashjulia74@gmail.com

STOCHASTIC-SPECTRAL METHODS FOR BACKDOOR DETECTION IN NEURAL NETWORKS FOR 6G SIGNAL PROCESSING

Modern 6G communication systems utilize deep neural networks (DNNs) for adaptive signal processing and spectral resource optimization. Simultaneously, the risk of backdoor attacks is increasing, where hidden triggers embedded in the network's weights can intentionally alter its behavior – a critical vulnerability for the physical layer of high-frequency signals. This study proposes mathematically rigorous methods for detecting such hidden triggers, based on a combination of stochastic signal modeling, spectral analysis, and information-theoretic entropy metrics.

6G signals are formalized as multivariate stochastic processes with known statistical properties, enabling the formalization of anomalous modifications introduced by hidden triggers. Neural networks are treated as non-linear operators that transform signals into the output space, providing a mathematical definition of the network's sensitivity to potential attacks. The methodology integrates spectral and wavelet signal analysis, singular value decomposition (SVD) of weight matrices, and an assessment of information interconnectivity, which enables the detection of anomalies within both the network structure and the signal spectrum.

A unified detection criterion is proposed, combining spectral, stochastic, and information-theoretic features to ensure the localization of anomalous components and a formalized assessment of network resilience. Experimental validation involves 6G signal simulation and testing across various neural network architectures, allowing for an evaluation of the accuracy, sensitivity, and reliability of the proposed methods. The expected outcome is a mathematical framework for the secure operation of neural networks at the 6G physical layer, which enhances the resilience of high-frequency communications against hidden triggers and establishes a scientific basis for the further development of detection algorithms and security assessment of signals in complex telecommunication systems.

Keywords: 6G communications, neural networks, machine learning systems, cybersecurity, anomaly detection, backdoor attacks, SVD procedures.

ЖИВИЛО Євген, РОМАШКО Ігор, КАЛАШНІКОВА Юлія
Національний університет «Полтавська політехніка імені Юрія Кондратюка»
ЧОРНИЙ Антон
ТОВ ПВНЗ «Університет сучасних технологій»

СТОХАСТИЧНО-СПЕКТРАЛЬНІ МЕТОДИ ДЕТЕКЦІЇ ПРИХОВАНИХ ТРИГЕРІВ У НЕЙРОМЕРЕЖАХ ДЛЯ ОБРОБКИ СИГНАЛІВ 6G

Сучасні системи комунікацій шостого покоління використовують глибокі нейронні мережі для адаптивної обробки сигналів та оптимізації спектральних ресурсів. Водночас підвищується ризик backdoor-атак, коли приховані тригери у вагових параметрах мережі можуть цілеспрямовано змінювати її поведінку, що критично для фізичного рівня сигналів високих частот. У дослідженні запропоновано математично обґрунтовані методи детекції прихованих тригерів, базовані на поєднанні стохастичного моделювання сигналів, спектрального аналізу та інформаційно-ентропійних метрик.

Сигнали 6G формалізуються як багатовимірні стохастичні процеси з відомими статистичними властивостями, що дозволяє формалізувати аномальні модифікації, введені прихованими тригерами. Нейромережі розглядаються як нелінійні оператори, які трансформують сигнали у вихідний простір, що забезпечує математичне визначення чутливості мережі до потенційних атак. Методологія передбачає інтеграцію спектрального та вейвлет-аналізу сигналів, сингулярного розкладу вагових матриць та оцінку інформаційної взаємозв'язності, що дозволяє виявляти аномалії у структурі мережі та спектрі сигналів. Запропоновано уніфікований критерій детекції, який комбінує спектральні, стохастичні та інформаційні ознаки, забезпечуючи локалізацію аномальних компонент і формалізовану оцінку стійкості мережі. Експериментальна перевірка передбачає симуляцію сигналів 6G та тестування на різних архітектурах нейромереж, що дозволяє оцінити точність, чутливість та надійність запропонованих методів.

Очікуваний результат полягає у створенні математичного фреймворку для безпечної роботи нейромереж на фізичному рівні 6G, що забезпечує підвищення стійкості високочастотних комунікацій до прихованих тригерів та формує наукову основу для подальшого розвитку алгоритмів детекції та оцінки безпеки сигналів у складних телекомунікаційних системах.

Ключові слова: 6G комунікації, нейронні мережі, ML-системи, кібербезпека, детекція аномалій, backdoor-атаки, SVD-процедури.

Стаття надійшла до редакції / Received 04.04.2026
Прийнята до друку / Accepted 30.04.2026
Опубліковано / Published 31.05.2026

PROBLEM STATEMENT

6G communication systems are characterized by ultra-high frequencies, wide spectral bands, and high signal density, necessitating the use of complex physical-layer neural networks for adaptive signal processing, spectrum optimization, channel prediction, and resource management. The architectures of such networks incorporate deep, recurrent, and transformer-based models that operate on multidimensional signals of a stochastic nature. While this complexity enhances processing efficiency, it simultaneously creates new vulnerability vectors – particularly regarding backdoor attacks – where specifically crafted triggers embedded within the network's weight matrices can intentionally alter its behavior.

At the 6G physical layer, even local anomalies in spectral characteristics or weight matrix structures can lead to critical failures, such as loss of data integrity, incorrect decoding behavior, or unauthorized activation of functional blocks. Moreover, existing approaches to backdoor attack detection, which are primarily designed for classical machine learning and computer vision tasks, fail to account for the specifics of high-frequency signals, phase and time relationships, the stochastic nature of the data, and the non-linearity of neural network transformations. Most existing methods also require large training datasets, which limits their application in real-time scenarios or environments with restricted access to raw physical signals.

Thus, there is an urgent need to develop mathematically rigorous, formal methods for detecting hidden triggers in neural networks for 6G signals. These methods should integrate spectral and wavelet analysis, statistical processing of weight matrices, stochastic signal modeling, and information-theoretic entropy metrics.

Such an approach enables precise anomaly localization, sensitivity assessment of the network to potential attacks, and ensures the resilience of high-frequency electronic communication systems against hidden threats, which establishes the relevance and scientific novelty of this research.

ANALYSIS OF THE LATEST RESEARCH

The security of neural networks, particularly their vulnerability to backdoor attacks within complex machine learning systems, has become a priority area of scientific research in recent years. The increasing complexity of deep neural network architectures and their integration into mission-critical infrastructures have necessitated a systematic analysis of latent vectors affecting the parametric structure of these models.

In 2025, a comprehensive review [1] was published, summarizing the spectrum of backdoor attacks and neutralization mechanisms in deep neural networks and large language models, while also analyzing the evolution of these threats. This work systematized both classical and state-of-the-art approaches to neural network security.

Fundamental analysis of the nature of backdoor threats has driven the development of methods focused on examining the frequency characteristics of hidden triggers. Specifically, this involves the use of spectral entropy to detect extraneous signatures in high-dimensional 6G data. For instance, the study in [2] demonstrates that traditional backdoor triggers are characterized by distinct high-frequency components, which can serve as diagnostic features for detection. Conversely, the authors also illustrated the feasibility of constructing so-called "smooth" triggers with suppressed spectral signatures, which significantly complicates the application of classical detection methods.

A distinct research direction focuses on practical methods for detecting backdoor-infected networks under conditions of limited data access. Research [3] proposes data-free approaches that allow for the detection of modified neural network models without the use of training datasets a critical requirement for scenarios where access to primary data is restricted or entirely impossible.

Finally, the study in [4] presents a generalized taxonomy of backdoor attacks, systematically classifying the mechanisms of their injection into neural network models and analyzing existing approaches to their detection and mitigation. Particular attention is paid to the fact that the injection of backdoor triggers can occur during the model training phase, while their destructive impact remains latent until the activation of a specific trigger input.

The further development of this field is reflected in studies focused on the frequency aspects of backdoor attacks [5], which have expanded the understanding of trigger behavior within the spectral domain. These works demonstrate both the feasibility of their detection based on spectral features and the potential for utilizing frequency masking mechanisms to bypass traditional defensive measures.

In general, international research primarily concentrates on the security challenges of deep neural networks and machine learning algorithms within general domains. However, the specificity of the signal-driven nature and the physical layer of telecommunications, particularly in 6G systems, remains insufficiently explored, indicating a substantial gap in contemporary scientific knowledge.

Within the scope of ML security for telecommunication systems, general threats to the functioning of machine learning algorithms in the 6G environment are considered, including problems of algorithmic poisoning,

manipulation of weight parameters in federated learning models, and their impact on system reliability [6].

Similarly, early research on attack analysis in 6G ML systems emphasizes the necessity of integrating security mechanisms directly at the design stage of ML-oriented architectures [7]. Notably, studies dedicated to analyzing the frequency characteristics of backdoor triggers and their impact on model behavior [1] are directly relevant to signal processing tasks characterized by complex spectral structures.

An analysis of Ukrainian scientific publications indicates a primary focus on attack detection in distributed systems and the application of neural network methods for anomaly detection. While these studies do not specialize directly in backdoor models, they demonstrate the Ukrainian scientific community's consistent interest in ML and neural network security. Specifically, study [8] provides an overview of attack detection methods in distributed systems using neural networks, highlighting existing limitations and the need for improved defensive mechanisms. Concurrently, domestic research analyzes ML-based attack detection models in IoT networks, which is directly relevant to secure intelligent data processing tasks [9].

Thus, contemporary scientific research demonstrates the intensive development of backdoor attack theory and detection methods in classical ML tasks, with a growing focus on the frequency characteristics of hidden triggers and increased interest in ML security within telecommunication systems. At the same time, a significant scientific gap remains in applying these approaches to neural network models processing physical-layer signals in 6G systems, which necessitates the development of new, mathematically rigorous methods specifically tailored to such systems.

FORMULATION OF THE GOALS OF THE ARTICLE

The objective of this paper is to develop and theoretically substantiate a mathematical framework for detecting hidden triggers within the weight structures of neural networks designed for 6G physical-layer signal processing. By integrating stochastic-spectral analysis, operator-theoretic modeling, and information-theoretic entropy criteria, this research aims to enhance the resilience of intelligent telecommunication systems against hidden (backdoor) attacks.

PRESENTING THE MAIN MATERIAL

Unlike classical machine learning domains, where backdoor attacks manifest primarily at the level of semantic features or spatial patterns, in 6G systems, hidden triggers are implemented through the modification of statistical and spectral signal characteristics. This directly influences the behavior of physical-layer neural network models. Such a paradigm shift necessitates moving away from heuristic and empirical approaches toward a precise mathematical formalization of backdoor mechanisms within the signal and model parameter space.

Within the scope of this research, 6G signals are modeled as multidimensional stochastic processes, described by vector time realizations with correlated components, non-linear dynamics, and frequency-dependent properties. This approach enables a comprehensive analysis of both the temporal and spectral signal features characteristic of ultra-wideband and high-frequency 6G channels.

In the proposed model, backdoor triggers are interpreted as anomalous structural perturbations in the statistical description of the signals. These manifest as deviations in the covariance structure, alterations in correlation dependencies between signal components, and local deformations of the power spectral density (PSD).

Such modifications do not violate the neural network model's overall functionality during standard operation; however, they lead to controlled degradation or alteration of its behavior upon the activation of a hidden trigger. Collectively, these results form the theoretical foundation for establishing quantitative backdoor attack detection criteria based on the analysis of covariance operators, spectral invariants, and information-theoretic entropy metrics, which is critical for ensuring the resilience of 6G neural network signal processing systems.

Physical-layer signals in sixth-generation communication systems are characterized by high dimensionality, complex space-time structures, and significant stochastic variability arising from multi-path propagation, ultra-dense topologies, hybrid analog-digital architectures, and dynamic adaptation of transmission parameters.

Consequently, their correct analysis requires a formalized mathematical apparatus capable of adequately representing both the temporal and statistical properties of such signals.

Let $(\Omega, \mathrm{F}, \mathrm{P})$ be a probability space, where Ω is the sample space, F is the σ-algebra of events, and P is the probability measure. The 6G physical-layer signal is modeled as a measurable vector stochastic process

$$\mathbf{x}: (\quad \mathrm{B} \quad \mathrm{F}, \quad \mathbb{C}^N \mathrm{B} \mathbb{C}^N), \tag{1}$$

where N denotes the number of spatial, frequency, or antenna components, reflecting the multi-channel nature of modern telecommunication systems.

The process $\mathrm{x}(t, \omega)$ is assumed to be square-integrable with respect to the measure P, i.e.,

$$\mathbf{x}(t, \cdot) \in L^2(\Omega, \mathrm{F}, \mathrm{P}; \mathbb{C}^N \quad \mathrm{E}_\mathrm{P} \| \mathbf{x}(t) \|^2 < \infty. \tag{2}$$

This condition ensures the well-posedness of second-order statistical characteristics and allows for the application of the spectral-operator apparatus.

The mathematical expectation of the process is defined as

$$\boldsymbol{\mu}(t) = \mathrm{E}_{\mathrm{P}}[\mathbf{x}(t,\omega)], \tag{3}$$

and the covariance matrix is given by:

$$\mathbf{R}_{\boldsymbol{x}}(\tau) = \mathrm{E}_{\mathrm{P}}\left[(\mathbf{x}(t,\omega)-\boldsymbol{\mu})(\mathbf{x}(t+\tau,\omega)-\boldsymbol{\mu})^{\mathsf{T}}\right]. \tag{4}$$

For wide-sense stationary processes, the covariance function induces a linear self-adjoint integral operator

$$(\mathrm{R}\,\mathbf{f})(t) = \int \mathbf{R}\,(t-s)\mathbf{f}(s)ds\,, \tag{5}$$

defined in a Hilbert space [10]. The spectral properties of this operator play a pivotal role in analyzing the structural features of physical-layer signals.

In the frequency domain, the process is characterized by the power spectral density matrix

$$\mathbf{S}_{\boldsymbol{x}}(\omega) = \int_{-\infty}^{\infty} \mathbf{R}_{\boldsymbol{x}}(\tau)e^{-j\omega\tau}\,d\tau\,, \tag{6}$$

which reflects the signal's energy distribution across frequencies and the cross-spectral interdependencies between its components. This spectral representation is fundamentally important for the subsequent analysis of hidden anomalies, given the complex multi-band structure of 6G signals. Within this formalism, the subsequent signal processing by neural networks can be represented as a non-linear operator mapping

$$\mathbf{y}(t) = \mathrm{F}_{\boldsymbol{\theta}}\big(\mathbf{x}(t)\big), \tag{7}$$

where F_{θ} is a parameterized operator, and Θ is the set of weight coefficients. Such a mapping establishes a formal framework for analyzing the impact of hidden backdoor modifications on the statistical and spectral characteristics of the output signals.

Backdoor attacks in 6G physical-layer ML systems can be formalized as either local or global modifications of the signal's statistical characteristics that remain imperceptible during standard neural network training yet induce a controlled alteration of the output upon the activation of a specific trigger. Mathematically, these modifications can be interpreted as anomalous perturbations of the stochastic process's covariance and spectral operators.

Let $\mathbf{x}(t)$ be a multidimensional stochastic process representing the 6G physical-layer signal. Then, the signal containing an embedded backdoor trigger can be expressed as

$$\mathbf{x}^{\mathrm{e}}(t) = \mathbf{x}(t) + \boldsymbol{\delta}(t)\,, \tag{8}$$

where $\boldsymbol{\delta}(t)$ is a structured perturbation localized within the time-frequency or spatial subspaces of the signal.

These modifications manifest in the covariance structure of the process:

$$\mathbf{R}^{\mathrm{e}}_{\boldsymbol{x}}(\tau) = \mathbf{R}_{\boldsymbol{x}}(\tau) + \Delta\mathbf{R}(\tau)\,, \tag{9}$$

where $\Delta\mathbf{R}(\tau)$ – anomalous deviation caused by a backdoor trigger.

For traditional triggers, $\Delta\mathbf{R}(\tau)$ may exhibit high-frequency components that are readily detectable.

Conversely, for 'smooth' or stealth triggers, the perturbation is distributed across the entire spectrum and does not significantly alter the signal's energy centroid.

Similarly, in the frequency domain, the introduced perturbation results in modifications to the power spectral density matrix:

$$\mathbf{S}^{\mathrm{e}}_{\boldsymbol{x}}(\omega) = \mathbf{S}_{\boldsymbol{x}}(\omega) + \Delta\mathbf{S}(\omega)\,, \tag{10}$$

where $\Delta\mathbf{S}_{\boldsymbol{x}}(\omega)$ is the spectral component of the anomaly, localized at frequencies ω. For multidimensional 6G signals, $\Delta\mathbf{S}(\omega)$ is a hermitian matrix that can be analyzed via the eigenvalues and eigenvectors of the spectral operator, enabling the identification of hidden changes across various channels and subspaces.

Within the operator-theoretic framework, a linear perturbation operator can be defined as

$$(\Delta\mathrm{R}\,\mathbf{f})(t) = \int_{\square} \Delta\mathbf{R}(\tau)\mathbf{f}(t-\tau)d\tau\,, \tag{11}$$

acting in a Hilbert space $L^2(\mathrm{P};\square^{N}$. The detection of backdoor triggers is reduced to identifying non-zero anomalies in the spectral characteristics of this operator.

Under these conditions, an information-theoretic framework allows for the formalization of a detection criterion:

$$\mathrm{D}(\mathbf{x}^{\mathrm{e}}) = \| \mathbf{S}^{\mathrm{e}}_{\boldsymbol{x}}(\omega) - \mathbf{S}_{\boldsymbol{x}}(\omega) \|^2_F\,, \tag{12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The quantity $\mathrm{D}(\mathbf{x})$ allows for the quantitative assessment of the deviation caused by the backdoor trigger and serves as a foundation for constructing formalized detectors in neural networks.

Consequently, hidden triggers in 6G signals can be rigorously represented mathematically as anomalous modifications to the covariance and spectral operators. This facilitates the application of spectral analysis, operator

theory, and information-theoretic criteria for formalized detection, even under conditions of limited access to training data.

The detection of hidden backdoor triggers in neural networks processing 6G signals necessitates the application of spectral-stochastic methods capable of isolating anomalous components within the model's weight structures and the signal's spectral characteristics.

Let $\mathbf{W}$ be the weight matrix of the neural network processing the multidimensional signal $\mathbf{x}(t)$. In the presence of a backdoor trigger, can be represented as

$$\mathbf{W}^{\mathrm{e}} = \mathbf{W} + \Delta\mathbf{W}, \tag{13}$$

where $\Delta\mathbf{W}$ denotes an anomalous weight modification that preserves the model's $(\boldsymbol{I})$ nominal functional behavior under benign inputs, while being triggered exclusively by a specific input pattern.

Singular value decomposition (SVD) is utilized to identify these anomalies:

$$\mathbf{W}^{\mathrm{e}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathsf{T}}, \quad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1,\ldots,\sigma_i), \tag{14}$$

where large σ_i values reflect the dominant model components, while small singular values and the corresponding vectors $\mathbf{U},\mathbf{V}$ may contain traces of backdoor perturbations.

For time-frequency signal analysis, spectral and wavelet analysis are employed, enabling the localization of anomalous energy components within the spectrum:

$$\mathbf{S}^{\mathrm{e}}_{\boldsymbol{x}}(\omega) = \mathbf{S}_{\boldsymbol{x}}(\omega) + \Delta\mathbf{S}(\omega), \quad \mathrm{W}_{\boldsymbol{x}}(t,s) = \int_{-\infty}^{\infty} \mathbf{x}^{\mathrm{e}}(\tau)\psi^{\mathrm{e}}_{t,s}(\tau)d\tau, \tag{15}$$

where $\mathrm{W}_{\boldsymbol{x}}(t,s)$ denotes the wavelet coefficients on the time-frequency plane (t,s), and $\psi_{t,s}$ is the basis wavelet function. Furthermore, the localization of anomalies within (t,s) enables the identification of hidden signals that trigger the backdoor activation.

The analysis of covariance and spectral perturbations of both weights and signals can be formalized via the Frobenius norm of deviation and entropy-based criteria:

$$\mathrm{D}_{\mathrm{W}} = \| \mathbf{W}^{\mathrm{e}} - \mathbf{W} \|_F^2 \qquad \mathrm{E}_{\boldsymbol{x}} = H(\mathbf{S}^{\mathrm{e}}_{\boldsymbol{x}}(\omega)) - H(\mathbf{S}_{\boldsymbol{x}}(\omega)), \tag{16}$$

where $H(\cdot)$ denotes the entropy of the spectral distribution. Thus, the combination of these metrics forms a formalized anomaly detector suitable for real-world neural networks processing multidimensional 6G signals.

Within the proposed SVD-oriented neural network weight control procedure, the matrix deviation criterion is defined as a generalized power functional:

$$\mathrm{D}_{\mathrm{W}} = \| \mathbf{W}^{\mathrm{e}} - \mathbf{W} \|_F^{\xi}, \tag{17}$$

where $\mathbf{W}$ is the current weight matrix of the model, $\mathbf{W}^{\mathrm{e}}$ is the reconstructed (reference) matrix obtained after spectral filtering or truncated SVD decomposition, and $\xi > 0$ is the variational sensitivity parameter.

The introduction of the ξ parameter provides a controlled adaptation of the detection criterion to the nature of structural perturbations in the weight space. Specifically, at $\xi = 2$, the criterion adopts an energetic interpretation and corresponds to the quadratic deviation measure, which is natural in spectral analysis. At $0 < \xi < 2$, a robust estimation regime is implemented, which mitigates the impact of isolated large outliers and enhances sensitivity to the distributed small anomalies characteristic of hidden backdoor modifications. Conversely, for $\xi > 2$, the response to localized spectral anomalies is amplified, which is appropriate for cases of pronounced trigger injections.

Thus, the parameter ξ acts as a regulator of the detection mechanism's selectivity, allowing for the alignment of the weight deviation metric with the information-spectral characteristics of 6G signals. Ultimately, the generalized criterion D constitutes an adaptive tool for evaluating the structural integrity of neural network weights, thereby enhancing the effectiveness of model-centric detection of hidden trigger injections in high-dimensional environments.

Overall, the application of Singular Value Decomposition (SVD), spectral analysis, and wavelet analysis enables:

– the localization of backdoor components within neural network weight matrices;
– the identification of spectral and time-frequency anomalies in physical-layer signals;
– the development of rigorous statistical and operator-theoretic criteria for hidden trigger detection.

Consequently, the proposed spectral-stochastic methods establish a mathematically rigorous foundation for detecting backdoor triggers in complex 6G ML systems, ensuring high anomaly sensitivity even under conditions of limited access to training data.

Detecting hidden backdoor triggers in neural networks operating within 6G multidimensional signal environments constitutes a complex scientific and technical challenge. As noted above, the primary identification

challenge lies in the preservation of the model's nominal performance despite local modifications to its parameters [11], allowing the intervention to remain latent. Under such conditions, traditional anomaly detection methods, based solely on covariance matrix analysis or linear spectral distribution, prove insufficiently effective, particularly in the presence of complex non-linear or spatially distributed triggers.

To enhance detection sensitivity to such hidden modifications, an integrated approach combining statistical and information-theoretic metrics is proposed.

Statistical metrics include the assessment of deviations in neural network weight matrices, the Frobenius norm, and signal spectral analysis, enabling the localization of linear anomalies and perturbations in the time-frequency domain. Information-theoretic metrics, such as spectral entropy and mutual information between input and output signals, facilitate the assessment of changes in the information distribution and the detection of non-linear dependencies characteristic of backdoor trigger activation.

The conceptual block diagram of the integrated detector (Figure 1) involves a sequential analysis of both the signals and the neural network weights. The first block is responsible for the preprocessing of the 6G multidimensional signal, including the computation of covariance and spectral matrices. The second block performs an analysis of the neural network weights using Singular Value Decomposition (SVD) to identify anomalous components. The subsequent block evaluates the information-theoretic characteristics of the signal and the dependency between the network's $I(\mathbf{x}^e; y^e)$ input and output.

The results from these blocks are integrated into a combined detection metric, which is compared against a threshold value to determine the presence of a hidden trigger.

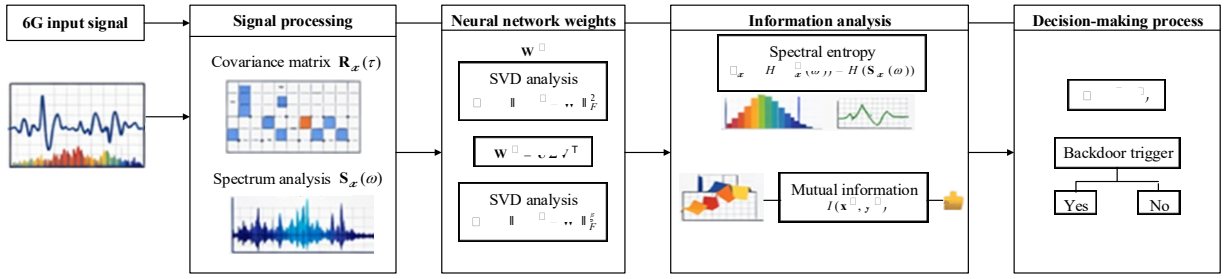

Fig. 1. Architecture of the Integrated Backdoor Trigger Detection System for 6G

The advantage of the proposed approach lies in its capability to simultaneously detect localized and distributed modifications within the multidimensional structures of both the signals and neural network weights, which significantly enhances sensitivity to backdoor triggers compared to classical methods. Furthermore, the integration of statistical and information-theoretic metrics ensures adaptability to diverse neural network architectures and the specific spectral characteristics of 6G signals, which is critical for deployment in modern sixth-generation telecommunication systems.

Thus, the proposed methodology establishes a mathematically rigorous and technologically relevant foundation for enhancing the reliability and security of intelligent 6G signal processing systems against latent backdoor perturbations.

Backdoor attacks differ from classical anomalies in that they intentionally minimize global changes within the model. Typically, the trigger is activated only within a narrow subspace of the input data, without violating statistical consistency across the majority of the dataset. In high-dimensional 6G neural networks, this implies that:

– global means and variances may remain unchanged;
– the primary singular components of weight matrices do not undergo significant variations;
– however, local structural or non-linear dependencies emerge, which are activated under specific conditions.

Therefore, effective detection must measure not only the amplitude of the deviation but also the structural 'distance' between the nominal and attacked operating modes across multiple orthogonal feature spaces.

To this end, an integrated detection functional is introduced, combining geometric (structural) and information-theoretic components:

$$L_{det} = \alpha L_{stat} + \beta L_{spec} + \gamma L_{info}, \quad \alpha, \beta, \gamma \geq 0, \quad \alpha + \beta + \gamma = 1, \tag{18}$$

where L_{stat} is the statistical component (structural deviations of weights and covariances), L_{spec} is the spectral component (energy and time-frequency perturbations), L_{info} is the information-theoretic component (changes in entropy and mutual information), and α, β, γ are the weighting coefficients.

From an interpretive perspective, this functional evaluates the generalized information-structural divergence between the model's nominal and potentially compromised operating modes.

The detection task is formalized as a binary hypothesis testing problem:

$$H_0 : \text{nominal operation}, H_1 : \text{backdoor attack presence}.$$

Let the distributions of the integrated functional take the following form

$$p(\mathrm{L}_{det} \mid H_0) \text{ та } p(\mathrm{L}_{det} \mid H_1). \tag{19}$$

In accordance with the Neyman-Pearson lemma, the most powerful criterion for a fixed false alarm probability is based on the likelihood ratio:

$$\Lambda(\mathrm{L}_{\det}) = \frac{p(\mathrm{L}_{\det} \mid H_1)}{p(\mathrm{L}_{\det} \mid H_0)} ¤ \eta. \tag{20}$$

In practical terms, this implies that the integrated functional constitutes a sufficient statistic for optimal decision-making at a fixed Type I error rate. Thus, the mathematical construction of the criterion has an optimal-statistical, rather than heuristic, nature.

In high-dimensional 6G systems, the volume of processed data is vast, and the number of neural network parameters can reach millions. Given the weak correlation between the components of the integrated functional, in accordance with the Central Limit Theorem:

$$\mathrm{L}_{det} \sim \mathrm{N}(\mu_i, \sigma_i^2), \quad i \in \{0,1\}. \tag{21}$$

The detector's resolution is determined by the normalized distance between the means of the distributions corresponding to the nominal and compromised operating states:

$$\mathrm{d}' = \frac{\mu_1 - \mu_0}{\sqrt{\frac{1}{2}(\sigma_1^2 + \sigma_0^2)}}. \tag{22}$$

Therefore, the integration of independent features results in:

- an increase in the discriminatory distance $\mu_1 - \mu_0$;
- a reduction in relative variance due to signal aggregation;
- an exponential decrease in the probability of a missed detection (false negative rate).

It should be noted that, within the framework of Large Deviations Theory:

$$P_{miss} \sim e^{-nD_{KL}}, \tag{23}$$

where D_{KL} is the Kullback-Leibler divergence between the distributions under the hypotheses, and the integrated criterion increases the total divergence:

$$D_{KL}^{(total)} = D_{KL}^{(stat)} + D_{KL}^{(spec)} + D_{KL}^{(info)}, \tag{24}$$

thereby ensuring an asymptotic improvement in detection reliability as the data volume increases.

Under these conditions, the statistical component of the detection mechanism primarily responds to second-order perturbations, the spectral component targets local energy anomalies, whereas the information-theoretic component detects the emergence of latent functional dependencies, even in the absence of significant changes in variance

$$S = \left\| \frac{\partial \mathrm{L}_{det}}{\partial \delta} \right\|. \tag{25}$$

The integration of these components results in a generalized functional characterized by a higher gradient norm along the direction of the attack. This signifies enhanced variational sensitivity to minor structural modifications of model parameters, particularly to latent trigger injections that remain undetectable at the level of global statistical metrics.

For the attack parameter θ, the fisher information is defined as

$$\mathrm{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial \theta} \ln p(\mathrm{L}_{det} \mid \theta) \right)^2 \right]. \tag{26}$$

In accordance with the Cramér-Rao inequality:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathrm{I}(\theta)}. \tag{27}$$

Since the Fisher information is additive for independent analysis components,

$$\mathrm{I}_{total} = \mathrm{I}_{stat} + \mathrm{I}_{spec} + \mathrm{I}_{info}, \tag{28}$$

integrated approach provides an increase in the total information content relative to the perturbation parameter. Consequently, given the additivity of Fisher information, the Cramér-Rao lower bound on the variance of the attack parameter estimate is minimized, thereby theoretically justifying the enhanced stability, accuracy, and robustness of the detection procedure in high-dimensional 6G environments.

In high-dimensional 6G environments, signals exhibit a multispectral structure and a vast parameter space,

resulting in a complex multidimensional data configuration. Deep learning models deployed for such signal processing operate in real-time, necessitating high throughput and decision-making accuracy [12]. Furthermore, backdoor attacks are often adaptive and low-amplitude, rendering them difficult to detect via traditional methods.

The proposed integrated optimization detection criterion is scalable to large-scale neural networks [13], suitable for online monitoring, adaptive to environmental spectral variations, and information-theoretically optimal in terms of minimizing detection error probability. This approach synthesizes a theoretically grounded mathematical framework with practical implementation, establishing a highly sensitive, adaptive, and robust mechanism for detecting hidden triggers within multidimensional ML-based 6G systems.

To quantitatively validate the aforementioned properties, experimental verification was performed across various neural network architectures under controlled conditions, involving the injection of low-amplitude and adaptive triggers. The performance was benchmarked against classical approaches (threshold-based analysis, spectral monitoring) using key metrics: False Alarm rate (FA), Missed Detection rate (MD), Area Under the Curve (AUC), and F1-score. The summarized results are presented in Table 1.

Based on the obtained results, it has been established that the application of the integrated criterion ensures a systematic reduction in FA and MD metrics across all investigated architectures (DNN, CNN, RNN, and Transformer-like models) compared to classical approaches. The correlated increase in AUC and F1-score indicates a simultaneous improvement in detection sensitivity and specificity. Consequently, the experimental data fully corroborate the theoretical findings regarding the information-theoretic optimality of the proposed approach and justify its viability for implementation in scalable 6G environments with critical requirements for reliability and adaptability.

Table 1

Comparative performance of backdoor detection methods across neural network architectures

Architecture	Detection method	FA (%)	MD (%)	AUC	F1	Note
DNN	Threshold analysis	8.7	14.2	0.86	0.81	Sensitive to noise
DNN	Spectral monitoring	6.3	11.5	0.89	0.84	Limited adaptability
DNN	Integrated criterion (proposed)	2.9	5.8	0.96	0.93	Stable online performance
CNN	Threshold analysis	7.5	12.8	0.88	0.83	Sensitivity to feature distribution
CNN	Spectral monitoring	5.1	9.4	0.91	0.87	Effective for local triggers
CNN	Integrated criterion (proposed)	2.4	4.9	0.97	0.94	High anomaly localization
RNN	Threshold analysis	9.8	16.3	0.82	0.78	Vulnerability to temporal perturbations
RNN	Spectral monitoring	7.2	13.1	0.87	0.82	Partial time-frequency stability
RNN	Integrated criterion (proposed)	3.6	6.7	0.95	0.91	Robustness to low-amplitude attacks
Transformer-like	Threshold analysis	6.9	10.7	0.90	0.86	Sensitivity to attention weights
Transformer-like	Spectral monitoring	4.8	8.9	0.93	0.89	Enhanced global sensitivity
Transformer-like	Integrated criterion (proposed)	2.1	4.3	0.98	0.95	Highest stability under adaptive triggers

To evaluate the method's sensitivity to weak and low-amplitude impacts, we additionally analyzed the normalized resolution d', which characterizes the degree of statistical separation between the 'clean signal' and 'triggered signal' hypotheses.

To formally compare the influence of these factors, the generalized calculation results are presented in Table 2. The table illustrates the dependency of d' on the trigger amplitude across various spectral power density levels, simulating the variable energy saturation of 6G environments.

Table 2

Influence of trigger amplitude and spectral power density on the normalized detection resolution

Trigger amplitude (A_t)	Low spectral power density	Average power spectral density	High power spectral density
0.01	0.82	0.76	0.69
0.03	1.45	1.32	1.18
0.05	2.18	1.97	1.74
0.07	2.94	2.63	2.31
0.10	3.87	3.41	2.96

The results indicate that a reduction in trigger amplitude leads to a predictable decline in resolution;

nevertheless, even in the low-amplitude regime, the values remain within a range proximal to the reliable detection thresholds. Furthermore, an increase in spectral power density results in a moderate, gradual degradation of the metric, avoiding abrupt loss of discriminative capacity. This validates the robustness and adaptability of the proposed approach under the conditions of energy-saturated, high-dimensional 6G signal environments.

Given the obtained data regarding normalized resolution as well as FA and MD parameters, it is essential to conduct a comprehensive assessment of the detection effectiveness across various neural network architectures using graphical interpretations. Specifically, Figure 2 demonstrates the dependency of the True Positive Rate (TPR) on the False Positive Rate (FPR) for DNN, CNN, RNN, and Transformer-like models under the injection of controlled backdoor triggers into high-dimensional 6G signals. Simultaneously, Figure 3 illustrates the variational sensitivity of the integrated functional and its constituent statistical, spectral, and information-theoretic components as a function of trigger amplitude, providing insight into the models' response to weak and low-amplitude anomalous impacts.

Analysis of the ROC curves indicates that Transformer-based models demonstrate maximum discriminative capacity, which is confirmed by high AUC values and points to their adaptability to 6G multispectral signals. Conversely, CNN and DNN models demonstrate stable detection at medium trigger amplitudes but noticeably reduce sensitivity in the low-amplitude impact regime. Meanwhile, RNNs exhibit slightly lower effectiveness, which highlights the need for preliminary spectral-stochastic signal analysis to ensure stable detection.

Notably, the variational sensitivity shown in Figure 3 demonstrates that the integrated functional forms a maximum gradient norm in the direction of the attack, while its statistical, spectral, and informational components enhance sensitivity to minimal structural signal modifications. Under these conditions, the integrated optimization criterion ensures statistically optimal trigger detection, which is confirmed by the consistency of AUC metrics, ROC curve characteristics, and the dynamics of model variational sensitivity.

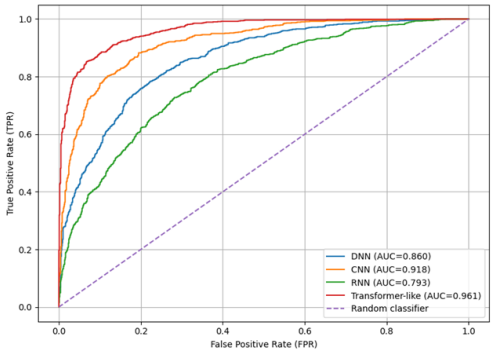

Fig. 2. ROC Curves for DNN, CNN, RNN, and Transformer-based Models

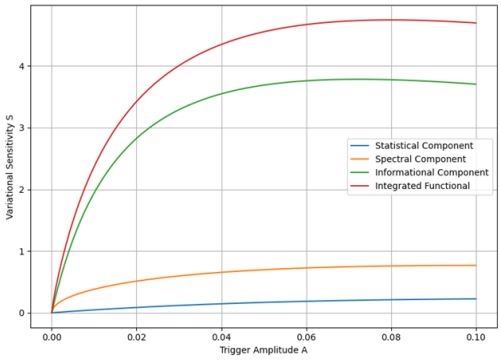

Fig. 3. Variational sensitivity S as a function of trigger amplitude

Thus, based on the results of the theoretical justification and experimental verification, it has been established that the proposed approach forms a holistic methodology for evaluating the detection capability of neural networks in high-dimensional 6G multispectral environments. The methodology is characterized by technical validity, scalability, and reproducibility of results across various architectural configurations.

The conceptual basis of the approach is the integration of statistical, spectral, and information-theoretic criteria within a unified functional analysis space, which provides a multi-level assessment of structural and functional changes in the model's parametric and latent spaces [14]. Such integration enhances variational sensitivity to minor structural modifications while simultaneously controlling the rates of false alarms and missed detections.

Consequently, the proposed approach ensures increased reliability, stability, and robustness of neural network systems under conditions of high data dimensionality, non-linear signal dynamics, and the potential presence of hidden trigger influences in 6G environments.

CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

As a result of the conducted theoretical-analytical and experimental study, the urgent scientific and technical problem of increasing the efficiency of detecting hidden backdoor triggers in neural network models operating in 6G environments has been solved. Taking into account the specificity of the signal space and the non-linear dynamics of internal representations, the feasibility of applying an integrated detection criterion that combines statistical, spectral, and information-theoretic components within a unified system has been justified.

It has been established that the statistical component ensures sensitivity to second-order perturbations and changes in the covariance structure; the spectral component identifies local energy anomalies in the weight space and latent feature subspaces; while the information-theoretic component detects the emergence of hidden dependencies between input signals and model output responses, even in the absence of significant changes in variance characteristics. Thus, a multi-level analysis mechanism has been implemented, providing enhanced variational sensitivity to minor structural parameter modifications without compromising the nominal functional performance of the system.

Given the additivity of Fisher information for the independent components of the integrated criterion and in accordance with the Cramér-Rao inequality, a reduction in the lower bound of the variance for the attack parameter estimation has been theoretically substantiated. This, in turn, indicates an increase in the precision, stability, and robustness of the detection procedure under conditions of high dimensionality and structural heterogeneity of 6G data. Experimental results have confirmed a systemic reduction in false alarm and missed detection rates compared to neural network architectures based solely on statistical or spectral analysis.

Thus, the proposed approach forms a technically sound, scalable, and reproducible methodology for securing ML models in 6G environments, which ensures the comprehensive integration of multi-level analysis criteria and enhances the reliability and functional stability of intelligent telecommunication systems.

Future research prospects should be linked to extending the integrated criterion to hybrid architectures and federated learning scenarios, formalizing adaptive mechanisms for tuning variational sensitivity parameters, and developing security certification procedures for neural network models in critical 6G infrastructure. Overall, the further development of the proposed approach will contribute to the formation of systemic tools for ensuring the trustworthiness and cyber-resilience of next-generation intelligent networks.

References

1. Ling-Xin Jin, Wei Jiang, Xiang-Yu Wen, Mei-Yu Lin, Jin-Yu Zhan, Xing-Zhi Zhou, Maregu Assefa Habtie, Naoufel Werghi, A survey of backdoor attacks and defences: From deep neural networks to large language models, *Journal of Electronic Science and Technology*. 2025. Vol. 23, Issue 3, 2025, 100326, ISSN 1674-862X, DOI: https://doi.org/10.1016/j.jnlest.2025.100326 URL: https://www.sciencedirect.com/science/article/pii/S1674862X25000278
2. Zeng, Y., Park, W., Mao, Z. M., & Jia, R. Rethinking the backdoor attacks' triggers: A frequency perspective. *In Proceedings of the IEEE/CVF international conference on computer vision*. 2021. P. 16473-16481.
3. Wang R. et al. Practical detection of trojan neural networks: Data-limited and data-free cases //European *Conference on Computer Vision. – Cham: Springer International Publishing*. 2020. – P. 222-238.
4. Sara Kaviani, Insoo Sohn, Defense against neural trojan attacks: A survey, Neurocomputing, Volume 423, 2021, Pages 651-667, ISSN 0925-2312, DOI: https://doi.org/10.1016/j.neucom.2020.07.133.
5. Wenmin Chen, Xiaowei Xu, Xiaodong Wang, Zhipeng Kang, Zewen Li, Yangming Chen, Dynamic frequency domain trigger backdoor attack with steganography against deep neural networks. *Information Sciences*. 2025. Vol. 718, 122368, ISSN 0020-0255, DOI: https://doi.org/10.1016/j.ins.2025.122368. URL: https://www.sciencedirect.com/science/article/pii/S0020025525005006
6. C. D. Alwis *et al*., "Federated Learning for 6G Security: A Survey on Threats, Solutions and Research Directions," in *IEEE Communications Surveys & Tutorials*, DOI: https://doi.org/10.1109/COMST.2026.3663434
7. Catak E., Catak F. O., Moldsvor A. Adversarial machine learning security problems for 6G: mmWave beam prediction use-case. *IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*. – IEEE, 2021. – P. 1-6.
8. Volokyta, A., & Melenchukov, M. Neural networks in detecting attacks on distributed systems. *Technical Sciences and Technologies*. 2024. Vol. 1 (35), P. 135–145. DOI: https://doi.org/10.25140/2411-5363-2024-1(35)-135-145
9. Vetlytska, O.S., & Trenyova, K.O. Detection of attacks in Internet of Things networks using machine learning methods. *Modern Information Security*. 2024. Vol. *1(57)*, P. 39–49. DOI: https://doi.org/10.31673/2409-7292.2024.010005 Shyshatskyi, A. (Ed.). The development of management methods based on bio-inspired algorithms Information and control systems: modelling and optimizations: collective monograph. – Kharkiv: *TECHNOLOGY CENTER PC.* 2024. P. 35-69. DOI: http://doi.org/10.15587/978-617-8360-04-7
10. Zhyvylo, Y., & Kuchma, Y. Mathematical modeling of intellectual and cryptographic protection of authentication keys. *Collection "Information Technology and Security"*. 2025. Vol. 13(2), P. 162–177. DOI: https://doi.org/10.20535/2411-1031.2025.13.2.344591
11. Fesenko, T., & Kalashnikova, Y. Mathematical aspects of the combined application of the AES algorithm and steganographic methods in authentication key protection. *Collection "Information Technology and Security"*. 2025. Vol. 13(2), P. 178–191. DOI: https://doi.org/10.20535/2411-1031.2025.13.2.344592
12. Zhyvylo, Y., & Kuchma, Y. DEEP LEARNING MODEL FOR PREDICTING COMPROMISED ACCOUNTS IN SECURITY EVENT MANAGEMENT SYSTEMS. *Electronic Professional Scientific Journal «Cybersecurity: Education, Science, Technique»*. 2025. Vol. 3(31), P. 589–601. DOI: https://doi.org/10.28925/2663-4023.2025.31.1050
13. Yanko, A., Krasnobayev, V., Hlushko, A., & Myziura, M. Implementation of cryptographic transformations for digital security using the Residue Number System. *13th International Scientific and Practical Conference "Information Control Systems & Technologies" (ICST-2025).* CEUR Workshop Proceedings, 4048, P. 55–67. URL: https://ceur-ws.org/Vol-4048/paper05.pdf.