STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
*Stat., Optim. Inf. Comput.*, Vol. **x**, Month 202x, pp 0–25.
Published online in International Academic Press (www.IAPress.org)

# Information-Theoretic Feature Selection via Variational Autoencoders for Network Intrusion Detection: A Mutual Information Maximization Approach

Bilal Bataineh [1,*], Ghazi Shakah [2]

[1] *Department of Computer Science, Jadara University, Irbid, Jordan*
[2] *Faculty of Information Technology, Ajloun National University, P.O.43, Ajloun-26810, Jordan*

**Abstract** Effective feature selection is essential for network intrusion detection systems (IDS), yet most existing approaches evaluate features solely within the original input space and overlook the probabilistic structure inherent in network traffic data. This paper introduces IT-VAE-FS (Information-Theoretic VAE-based Feature Selection), a novel framework that exploits the latent representations learned by a $\beta$-Variational Autoencoder as an information-theoretic reference frame for systematic feature evaluation. The framework constructs a composite feature importance score by integrating three complementary measures: (i) the mutual information between each input feature and the VAE's latent representation, estimated through the InfoNCE lower bound; (ii) the feature's contribution to the KL divergence component of the Evidence Lower Bound; and (iii) the feature's reconstruction sensitivity. A redundancy-aware greedy selection procedure, inspired by the minimum Redundancy Maximum Relevance criterion, identifies compact feature subsets that maximize informational coverage while minimizing inter-feature redundancy. Comprehensive experiments on three benchmark datasets—CICIDS-2017, UNSW-NB15, and NSL-KDD—demonstrate that IT-VAE-FS consistently surpasses nine established feature selection baselines spanning the filter, wrapper, embedded, and deep learning-based categories. Using only 20 selected features, the framework achieves F1-scores of 98.19%, 90.65%, and 82.80% on CICIDS-2017, UNSW-NB15, and NSL-KDD, respectively, falling within 0.06–0.16 percentage points of the full-feature baseline while reducing dimensionality by 51–74% and delivering a 3.3$\times$ inference speedup. The Matthews Correlation Coefficient reaches 0.970, 0.839, and 0.685 on the three datasets, confirming robust classification under class imbalance. An ablation study confirms the complementarity of all three scoring components, and the statistical significance of the results is established through Friedman and Wilcoxon signed-rank tests with Holm–Bonferroni correction ($p < 0.05$ for all pairwise comparisons). The formal connection between the $\beta$-VAE objective and the information bottleneck principle provides theoretical justification for using the latent space as a feature evaluation reference frame, situating the contribution within a well-established information-theoretic framework.

**Keywords** Feature selection; Variational Autoencoder; Mutual information; Intrusion detection; Information theory; KL divergence; Network security; Dimensionality reduction

**DOI:** 10.19139/soic-2310-5070-4198

## 1. Introduction

The rapid proliferation of network-connected devices and the escalating sophistication of cyber threats have rendered network intrusion detection systems (IDS) an indispensable component of modern cybersecurity infrastructure [1]. Contemporary network environments generate voluminous traffic data characterized by high dimensionality, where each network flow may be described by dozens to hundreds of statistical features [2]. While this feature richness ostensibly provides comprehensive information for distinguishing benign from malicious traffic, it simultaneously introduces substantial challenges related to computational overhead, model overfitting,

*Correspondence to: Bilal Bataineh (Email: B.Bataineh@Jadara.edu.jo). Department of Computer Science, Jadara University, Irbid, Jordan.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

and the curse of dimensionality [3]. Consequently, feature selection has emerged as a critical preprocessing step that directly influences both the detection accuracy and the operational efficiency of IDS [4].

Traditional feature selection methodologies for intrusion detection can be broadly categorized into filter, wrapper, and embedded approaches [5]. Filter methods, such as correlation-based feature selection and Chi-square tests, evaluate features independently of the learning algorithm, offering computational efficiency but potentially overlooking feature interactions [6]. Wrapper methods, including recursive feature elimination (RFE) and genetic algorithm-based selection, evaluate feature subsets using a specific classifier, thereby capturing interactions at the cost of significant computational expense [7]. Embedded methods, such as LASSO regularization and tree-based importance measures, integrate feature selection within the model training process [8]. Despite their respective merits, these conventional approaches operate exclusively in the original feature space and lack the capacity to exploit the underlying probabilistic structure of network traffic data.

The advent of deep generative models, particularly Variational Autoencoders (VAEs), has introduced a fundamentally different paradigm for understanding high-dimensional data [9]. A VAE learns a compressed latent representation by jointly optimizing a reconstruction objective and a regularization term based on the Kullback-Leibler (KL) divergence, thereby capturing the intrinsic probabilistic structure of the input data [10]. In the context of intrusion detection, VAEs have demonstrated considerable promise for anomaly detection, where deviations in reconstruction error signal potentially malicious activity [11]. However, the existing literature has predominantly employed VAEs as end-to-end detection models, neglecting the rich information-theoretic properties inherent in the learned latent space that could be leveraged for principled feature selection in the original input space. Information theory provides a rigorous mathematical framework for quantifying statistical dependencies between random variables [12]. Mutual information (MI), in particular, measures the total amount of information that one random variable conveys about another, capturing both linear and nonlinear dependencies without distributional assumptions [13]. The minimum Redundancy Maximum Relevance (mRMR) criterion, proposed by Peng et al. [14], represents a landmark contribution that employs MI to select features that are maximally relevant to the target variable while minimally redundant with respect to each other. More recently, neural MI estimators such as the Mutual Information Neural Estimation (MINE) framework [15] have enabled scalable MI computation in high-dimensional spaces, overcoming the intractability of classical density estimation approaches.

Despite the individual advances in VAE-based anomaly detection and information-theoretic feature selection, a significant research gap persists at their intersection. Specifically, no existing work has systematically investigated how the learned latent representations of a VAE can serve as an information-theoretic bridge for evaluating and selecting features in the original input space for intrusion detection. The latent space of a VAE encodes a compressed probabilistic summary of the input data; therefore, the mutual information between individual input features and the latent variables provides a principled measure of each feature's contribution to the data's intrinsic structure. This perspective is fundamentally distinct from conventional feature importance metrics, which are typically tied to a specific discriminative model and do not account for the generative properties of the data distribution [16]. To address this gap, this paper proposes a novel framework termed IT-VAE-FS (Information-Theoretic VAE-based Feature Selection) for network intrusion detection. The proposed framework operates in two principal stages. In the first stage, a $\beta$-VAE [17] is trained on network traffic data to learn a disentangled latent representation that captures the salient statistical regularities of the data. In the second stage, three complementary information-theoretic measures are computed for each input feature: (i) the estimated mutual information between the feature and the latent representation, computed via a variational lower bound; (ii) the feature's contribution to the KL divergence component of the Evidence Lower Bound (ELBO); and (iii) the feature's impact on reconstruction fidelity. These measures are aggregated into a composite information score, and a greedy selection procedure with a redundancy penalty inspired by the mRMR criterion [14] is employed to identify a compact, informative feature subset. The principal contributions of this work are fourfold. First, we introduce a novel information-theoretic feature selection framework that bridges VAE-based generative modeling with MI-based feature evaluation, providing a theoretically grounded approach to dimensionality reduction for IDS. Second, we derive a composite feature importance score that integrates mutual information, KL divergence contribution, and reconstruction sensitivity, offering a multi-faceted assessment of feature relevance. Third, we conduct extensive

experiments on three widely adopted benchmark datasets CICIDS-2017 [18], UNSW-NB15 [19], and NSL-KDD [20] demonstrating that the proposed method achieves competitive or superior detection accuracy using substantially fewer features compared to existing feature selection baselines. Fourth, we provide rigorous statistical validation through non-parametric hypothesis testing, including Friedman and Wilcoxon signed-rank tests, to establish the significance of the observed performance differences across multiple datasets and classifiers.

The motivation for the proposed work arises from a critical observation: while deep generative models such as VAEs learn rich probabilistic representations of high-dimensional data, the information-theoretic content embedded in their latent spaces has not been systematically harnessed for feature selection in the context of intrusion detection. Existing feature selection approaches either rely on univariate statistical tests that ignore feature interactions, or depend on computationally expensive wrapper methods tied to specific classifiers. The proposed IT-VAE-FS framework addresses this gap by introducing a principled, classifier-agnostic feature evaluation methodology grounded in information theory and deep generative modeling. This represents a significant paradigm shift from conventional discriminative feature importance metrics toward a generative, information-theoretic perspective on feature relevance.

The remainder of this paper is organized as follows. Section 2 reviews related work on feature selection for intrusion detection, VAE-based approaches, and information-theoretic methods. Section 3 presents the mathematical preliminaries, including the VAE formulation, mutual information estimation, and KL divergence decomposition. Section 4 details the proposed IT-VAE-FS framework. Section 5 describes the experimental setup, and Section 6 presents the results and discussion, and Section 7 concludes the paper with directions for future research.

## 2. Related Work

This section reviews the existing literature across three interrelated research streams that collectively inform the proposed framework: feature selection methodologies for intrusion detection systems, variational autoencoder-based approaches for network security, and information-theoretic feature selection methods. By examining the advances and limitations within each stream, we delineate the research gap that the proposed IT-VAE-FS framework seeks to address.

### *2.1. Feature Selection for Intrusion Detection Systems*

Feature selection has long been recognized as a pivotal preprocessing step in the design of effective intrusion detection systems. The high dimensionality of network traffic datasets, often comprising 40 to over 80 features per flow record, introduces redundancy and noise that degrade classifier performance and inflate computational costs [2]. The taxonomy of feature selection methods—filter, wrapper, and embedded provides a useful organizational framework for reviewing prior contributions [5].

Among filter-based approaches, Ambusaidi et al. [21] proposed a mutual information-based feature selection algorithm that jointly considers feature relevance and redundancy for IDS, demonstrating improved detection rates on the KDD Cup 99 and NSL-KDD datasets. Similarly, Khammassi and Krichen [22] employed a correlation-based filter combined with a genetic algorithm wrapper to select optimal feature subsets for the CICIDS-2017 dataset, reporting that fewer than 20 features could preserve classification accuracy above 99%. Moustafa and Slay [23] applied an association rule mining-based feature selection technique to the UNSW-NB15 dataset, identifying a reduced feature set that maintained high detection performance across multiple attack categories.

Wrapper-based methods have also been extensively explored. Stein et al. [24] employed genetic algorithms for feature subset selection in combination with decision tree classifiers, demonstrating that evolutionary search over the feature space yields compact subsets with competitive accuracy. More recently, Kasongo and Sun [25] proposed a wrapper-based feature selection method using an XGBoost classifier applied to the UNSW-NB15 dataset, achieving significant dimensionality reduction while maintaining detection accuracy above 90% for all attack types. Embedded methods have gained popularity due to their ability to integrate feature selection within the learning process. Vinayakumar et al. [26] utilized deep neural networks with inherent feature learning capabilities for

intrusion detection, effectively performing implicit feature selection through hierarchical representation learning. Ahmad et al. [27] combined Random Forest-based feature importance with deep learning classifiers, demonstrating that tree-based importance scores provide a reliable proxy for feature relevance in the IDS context. However, all of these approaches operate in the original feature space and evaluate feature importance through the lens of discriminative classifiers, without exploiting the generative structure of the data distribution.

### *2.2. Variational Autoencoder-Based Approaches for Network Security*

Variational Autoencoders (VAEs), introduced by Kingma and Welling [9], have emerged as a powerful generative modeling framework with broad applications in anomaly detection and network security. The VAE learns a probabilistic mapping from the input space to a low-dimensional latent space, enabling both data reconstruction and generation, which makes it particularly suitable for identifying anomalous patterns that deviate from learned normal behavior. An and Cho [11] were among the first to apply VAEs for anomaly detection, proposing the use of reconstruction probability rather than reconstruction error as a more principled anomaly score. Lopez-Martin et al. [28] extended this concept to network intrusion detection, demonstrating that conditional VAEs (CVAEs) can effectively model class-specific traffic distributions, enabling both anomaly detection and attack classification within a unified generative framework. Xu et al. [29] proposed an unsupervised anomaly detection method based on VAEs for network traffic, showing that the learned latent representations capture meaningful structural differences between normal and attack traffic. Yang et al. [30] introduced a VAE-based framework enhanced with adversarial training for improved anomaly detection in IoT networks. More recently, Zavrak and İşcan [31] conducted a comprehensive comparative analysis of autoencoder variants for anomaly-based intrusion detection, concluding that VAEs offer superior performance in capturing the distributional characteristics of normal network traffic. Despite these advances, the existing VAE-based IDS literature has predominantly focused on leveraging the reconstruction error or latent space distances for anomaly scoring. The information-theoretic properties of the VAE's learned representations specifically, the mutual information between input features and latent variables, and the per-feature contributions to the KL divergence remain largely unexplored as tools for feature selection.

### *2.3. Information-Theoretic Feature Selection*

It is also worth noting that the broader landscape of secure IoT systems, real-time monitoring, and deep-learning-based anomaly detection has seen substantial recent advances that complement the feature selection perspective adopted in this work. Research on designing robust IoT architectures [32] and real-time data integrity monitoring [33] underscores the importance of lightweight, efficient detection pipelines—a goal directly served by feature reduction. Multi-stream deep learning frameworks for complex temporal data [34] and intelligent edge-based threat classification [35] further motivate the need for principled feature selection to reduce computational burden at the network edge. Additionally, recent work on integrated deep learning pipelines for heterogeneous IoT environments [36] highlights the growing demand for generalizable, information-theoretically grounded feature evaluation methods such as IT-VAE-FS.

Information theory, rooted in the seminal work of Shannon [37], provides a mathematically rigorous framework for quantifying uncertainty and statistical dependence. Mutual information (MI), defined as the KL divergence between the joint distribution and the product of marginals, is a fundamental measure of the total dependence between two random variables [12]. The application of MI to feature selection was formalized by Battiti [38], who proposed the Mutual Information Feature Selection (MIFS) algorithm. Peng et al. [14] advanced this line of work with the minimum Redundancy Maximum Relevance (mRMR) criterion, which has become one of the most widely adopted information-theoretic feature selection approaches [39]. Brown et al. [40] provided a unifying theoretical framework showing that many existing MI-based feature selection criteria can be expressed as special cases of a general conditional likelihood maximization objective. Vergara and Estévez [41] further contributed a comprehensive review and experimental comparison of MI-based feature selection methods. A critical challenge in MI-based feature selection is the accurate estimation of MI in high-dimensional continuous spaces. Classical approaches based on k-nearest neighbor density estimation [13] suffer from the curse of dimensionality. The MINE framework [15] addressed this limitation by formulating MI estimation as a neural network optimization problem. Poole et al. [42] subsequently provided a comprehensive analysis of variational bounds on MI, establishing tighter

estimators particularly relevant for deep generative models. While information-theoretic methods have been applied to feature selection in various domains, their integration with deep generative models for IDS remains nascent. The proposed IT-VAE-FS framework bridges this gap by using the VAE's latent space as an information-theoretic reference frame for feature evaluation. The preceding review reveals that no existing work simultaneously operates in the latent feature space, employs information-theoretic measures, and leverages deep generative models for feature selection in intrusion detection. The IT-VAE-FS framework proposed in this paper occupies this unique intersection.

## 3. Preliminaries

This section establishes the mathematical foundations upon which the proposed IT-VAE-FS framework is constructed. We present the formal definitions and key properties of Variational Autoencoders, mutual information and its estimation techniques, the KL divergence decomposition, reconstruction sensitivity, and the Information Bottleneck connection, which collectively form the theoretical basis for the information-theoretic feature selection methodology developed in Section 4.

### *3.1. Variational Autoencoders*

A Variational Autoencoder (VAE) [9] is a deep generative model that learns a probabilistic mapping between an observed data space $\mathcal{X}$ and a latent variable space $\mathcal{Z}$. Given a dataset of network traffic samples $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, the VAE assumes that each observation $x$ is generated from a latent variable $z$ through a probabilistic generative process. The generative model is defined by a prior distribution over the latent variables and a conditional likelihood:

$$p(z) = \mathcal{N}(z;\, 0,\, I) \tag{1}$$

$$p_\theta(x|z) = f_\theta(x;\, z) \tag{2}$$

where $p(z)$ is a standard multivariate Gaussian prior and $p_\theta(x|z)$ is the decoder network parameterized by $\theta$ that outputs the parameters of the observation likelihood (e.g., Bernoulli or Gaussian parameters). The marginal likelihood $p_\theta(x) = \int p_\theta(x|z)\, p(z)\, dz$ is generally intractable due to the high-dimensional integral over the latent space, necessitating variational inference [43]. The VAE introduces an approximate posterior distribution $q_\phi(z|x)$, parameterized by an encoder network with parameters $\phi$, which approximates the true but intractable posterior $p_\theta(z|x)$. The encoder outputs the parameters of a factorized Gaussian distribution:

$$q_\phi(z|x) = \mathcal{N}(z;\, \mu_\phi(x),\, \sigma^2_\phi(x) \cdot I) \tag{3}$$

where $\mu_\phi(x)$ and $\sigma^2_\phi(x)$ are the mean and variance vectors output by the encoder network. The model is trained by maximizing the Evidence Lower Bound (ELBO) on the log-marginal likelihood:

$$\log p_\theta(x) \geq \mathcal{L}(x;\, \theta,\, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z)) \tag{4}$$

The first term in Equation (4) is the expected reconstruction likelihood, which encourages the decoder to accurately reconstruct the input from the latent code. The second term is the KL divergence between the approximate posterior and the prior, which acts as a regularizer that encourages the learned latent distribution to remain close to the prior. For Gaussian distributions, this KL term admits a closed-form expression [9]:

$$D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z)) = -\frac{1}{2} \sum_{i=1}^{d} \left[1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2\right] \tag{5}$$

where $d$ is the dimensionality of the latent space, and $\mu_i$ and $\sigma_i^2$ are the $i$-th components of the mean and variance vectors, respectively. This closed-form expression enables efficient gradient computation during training without requiring Monte Carlo estimation of the KL term.

The $\beta$-VAE [17] extends the standard VAE by introducing a hyperparameter $\beta$ that controls the relative weight of the KL divergence term:

$$\mathcal{L}^{\beta}(x;\,\theta,\,\phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z)) \tag{6}$$

When $\beta > 1$, the model is encouraged to learn more disentangled latent representations, where individual latent dimensions correspond to independent generative factors [17]. This disentanglement property is particularly valuable for the proposed framework, as it facilitates a more interpretable mapping between input features and latent dimensions, thereby enabling more meaningful information-theoretic analysis of per-feature contributions.

### *3.2. Mutual Information*

Mutual information (MI) is a fundamental information-theoretic quantity that measures the amount of information shared between two random variables [12]. For two continuous random variables $X$ and $Z$ with joint density $p(x, z)$ and marginal densities $p(x)$ and $p(z)$, the mutual information is defined as:

$$I(X;\,Z) = \int\!\!\int p(x,z) \log \frac{p(x,z)}{p(x) \cdot p(z)}\, dx\, dz \tag{7}$$

Equivalently, MI can be expressed in terms of entropy and conditional entropy:

$$I(X;\,Z) = H(X) - H(X|Z) = H(Z) - H(Z|X) \tag{8}$$

where $H(X) = -\int p(x) \log p(x)\, dx$ is the Shannon entropy and $H(X|Z) = -\iint p(x,z) \log p(x|z)\, dx\, dz$ is the conditional entropy. MI is non-negative, symmetric, and equals zero if and only if $X$ and $Z$ are statistically independent. Critically, unlike linear correlation measures, MI captures arbitrary nonlinear dependencies between variables [44], making it an ideal tool for evaluating feature relevance in complex network traffic data. For the purpose of feature selection, we are interested in quantifying the MI between individual input features $x_i$ and the latent representation $z$ learned by the VAE. A feature with high $I(x_i;\,z)$ contributes substantial information to the latent encoding and is therefore considered important for capturing the intrinsic data structure. Computing MI exactly requires knowledge of the joint and marginal densities, which are generally unavailable for high-dimensional continuous data. Several estimation strategies address this challenge [45].

**kNN Estimators.** Kraskov et al. [13] proposed a nonparametric MI estimator based on k-nearest neighbor distances in the joint and marginal spaces. While consistent and asymptotically unbiased, this estimator suffers from the curse of dimensionality and exhibits high variance when either variable is high-dimensional.

**Neural MI Estimators.** The Mutual Information Neural Estimation (MINE) framework [15] provides a scalable alternative by exploiting the Donsker-Varadhan representation of the KL divergence:

$$I(X;\,Z) = \sup_{T}\ \mathbb{E}_{p(x,z)}[T(x,z)] - \log \mathbb{E}_{p(x)p(z)}[e^{T(x,z)}] \tag{9}$$

where the supremum is over all functions $T : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ in a suitable function class. In practice, $T$ is parameterized by a neural network and the bound is optimized via gradient ascent, providing a consistent and scalable estimator.

**Variational Bounds.** Poole et al. [42] systematically analyzed several variational bounds on MI. For the proposed framework, we leverage the InfoNCE lower bound [46], which provides a tractable and numerically stable estimator:

$$I(X;\,Z) \geq \mathbb{E}\left[\frac{1}{N}\sum_{i} \log \frac{e^{f(x_i,z_i)}}{\frac{1}{N}\sum_{j} e^{f(x_i,z_j)}}\right] \tag{10}$$

where $f$ is a critic function and the expectation is over $N$ independent samples from the joint distribution. The InfoNCE bound is particularly suitable for the proposed framework because it is numerically stable, does not require density estimation, and naturally integrates with mini-batch gradient training.

### 3.3. KL Divergence and Its Decomposition

The Kullback-Leibler (KL) divergence is an asymmetric measure of the difference between two probability distributions [12]. For two continuous distributions $p$ and $q$, it is defined as:

$$D_{\mathrm{KL}}(p \,\|\, q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx \tag{11}$$

In the VAE framework, the KL divergence $D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z))$ plays a dual role: it serves as a regularizer in the ELBO and provides a measure of the information content encoded in the latent representation. For the proposed feature selection framework, we exploit the decomposability of the KL term to assess per-feature contributions. Hoffman and Johnson [47] demonstrated that the aggregate posterior KL can be decomposed into two interpretable components:

$$D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z)) = I_q(x;\, z) + D_{\mathrm{KL}}(q_\phi(z) \,\|\, p(z)) \tag{12}$$

where $I_q(x;\, z)$ is the mutual information between $x$ and $z$ under the variational distribution $q$, and $D_{\mathrm{KL}}(q_\phi(z) \,\|\, p(z))$ is the marginal KL divergence. This decomposition reveals that the KL term implicitly encodes the mutual information between the input and the latent representation, providing a direct connection between the VAE's regularization objective and the information-theoretic measures used in the proposed feature selection scheme.

To assess the contribution of each input feature $x_i$ to the KL divergence, we employ a permutation-based approach. Let $\Delta\mathrm{KL}(x_i)$ denote the change in KL divergence when the $i$-th input feature is randomly permuted across samples:

$$\Delta\mathrm{KL}(x_i) = |D_{\mathrm{KL}}(q_\phi(z|x) \,\|\, p(z)) - D_{\mathrm{KL}}(q_\phi(z|x_{\neg i}) \,\|\, p(z))| \tag{13}$$

where $x_{\neg i}$ denotes the input with the $i$-th feature replaced by a random permutation. A large $\Delta\mathrm{KL}(x_i)$ indicates that the feature substantially influences the latent encoding, implying high informational relevance. This permutation-based approach is model-agnostic and does not require retraining the VAE for each feature evaluation.

### 3.4. Reconstruction Sensitivity Analysis

The reconstruction term in the ELBO, $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$, quantifies how well the decoder recovers the input from the latent code. For feature selection, we define the reconstruction sensitivity of feature $x_i$ as the change in reconstruction error when the feature is permuted:

$$\Delta\mathrm{Recon}(x_i) = |L_{\mathrm{recon}}(x) - L_{\mathrm{recon}}(x_{\neg i})| \tag{14}$$

where $L_{\mathrm{recon}}(x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]$ is the negative expected reconstruction log-likelihood. Features with high $\Delta\mathrm{Recon}(x_i)$ are those upon which the VAE's generative model critically depends for accurate reconstruction, indicating that they carry substantial information about the data-generating process. This measure is complementary to the MI and KL scores: while MI quantifies the feature's contribution to the latent encoding, and the KL score captures its influence on the regularization objective, the reconstruction sensitivity reflects the decoder's reliance on the feature for accurate data recovery.

### 3.5. Connection to the Information Bottleneck Principle

The Information Bottleneck (IB) method, introduced by Tishby et al. [48], provides a theoretical framework for optimal representation learning. The IB objective seeks a compressed representation $Z$ of the input $X$ that preserves maximum information about a target variable $Y$:

$$\min_{p(z|x)} \; I(X;\, Z) - \beta \cdot I(Z;\, Y) \tag{15}$$

Alemi et al. [49] established a formal connection between the VAE and the IB principle, showing that the $\beta$-VAE objective in Equation (6) can be interpreted as an IB optimization where the reconstruction term acts as

a proxy for $I(Z;\, Y)$. This connection provides theoretical justification for using the $\beta$-VAE's latent space as an information-theoretic reference frame: the latent variables $Z$ are explicitly trained to capture a compressed yet informative summary of the input, making the MI between individual input features and $Z$ a principled measure of feature relevance. This observation is central to the proposed IT-VAE-FS framework: by leveraging a $\beta$-VAE trained to balance compression and reconstruction, the mutual information $I(x_i;\, z)$ naturally reflects the feature's contribution to the optimal trade-off between data fidelity and representational parsimony, providing a theoretically grounded feature importance metric that is distinct from and complementary to conventional discriminative importance measures.

## 4. Proposed Framework: IT-VAE-FS

This section presents the proposed Information-Theoretic VAE-based Feature Selection (IT-VAE-FS) framework in detail. The framework comprises four principal stages: (1) data preprocessing and normalization, (2) $\beta$-VAE training for latent representation learning, (3) information-theoretic feature scoring, and (4) redundancy-aware feature subset selection. Each stage is elaborated in the following subsections.

### *4.1. Overview of the IT-VAE-FS Architecture*

Let $X = \{x_1, x_2, \ldots, x_n\}$ denote a dataset of $n$ network traffic samples, where each sample $x \in \mathbb{R}^d$ is a $d$-dimensional feature vector. The objective of the IT-VAE-FS framework is to identify a subset $S \subseteq \{1, 2, \ldots, d\}$ of size $k \ll d$ such that the selected features $x_S = \{x_i : i \in S\}$ maximally preserve the information content relevant to intrusion detection while minimizing redundancy among selected features. The framework addresses this objective through a two-phase approach. In the first phase, a $\beta$-VAE is trained to learn a compressed latent representation $z \in \mathbb{R}^m$ (where $m \ll d$) that captures the essential probabilistic structure of the network traffic data. In the second phase, three complementary information-theoretic measures are computed for each input feature, aggregated into a composite score, and used to guide a redundancy-aware greedy selection procedure. Figure 1 illustrates the overall architecture of the proposed IT-VAE-FS framework.

### *4.2. Data Preprocessing*

Prior to model training, the raw network traffic data undergoes a standardized preprocessing pipeline designed to ensure numerical stability and compatibility with the VAE architecture. Categorical features (e.g., protocol type, service, flag) are transformed using one-hot encoding to produce binary indicator variables, expanding the feature dimensionality accordingly. All continuous features are normalized to the $[0, 1]$ range using min-max scaling:

$$\hat{x}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i) + \varepsilon} \tag{16}$$

where $\varepsilon = 10^{-8}$ is a small constant to prevent division by zero. Min-max normalization is preferred over z-score standardization because the VAE's Bernoulli or sigmoid-activated decoder naturally produces outputs in $[0, 1]$, ensuring consistency between the input representation and the reconstruction target.

Network intrusion datasets are typically characterized by severe class imbalance, with normal traffic vastly outnumbering attack instances. To mitigate this, we employ the Synthetic Minority Oversampling Technique (SMOTE) [50] applied to the training partition only, generating synthetic samples for minority attack classes. The validation and test partitions remain unmodified to ensure unbiased evaluation.

To ensure methodological rigor and prevent information leakage, a strict data partitioning protocol was followed throughout the experimental pipeline. The dataset was first divided into training (60%), validation (20%), and test (20%) partitions using stratified sampling. The $\beta$-VAE was trained exclusively on the training partition, with the validation set used solely for early stopping and the adaptive termination rule described in Section 4.5. The InfoNCE critic networks for mutual information estimation were likewise trained exclusively on the training partition. The permutation-based $\Delta$KL and $\Delta$Recon computations were performed using only the training

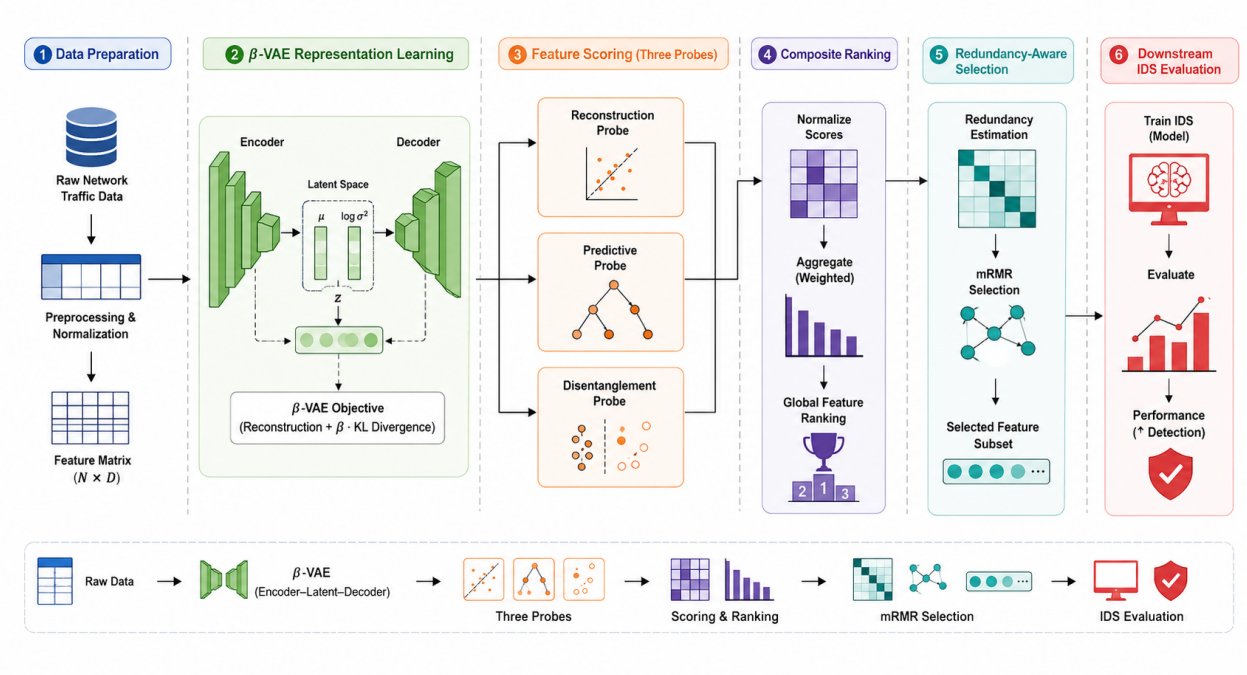


Figure 1. Overall architecture of the proposed IT-VAE-FS framework.

data. The test partition remained completely untouched until the final evaluation stage, ensuring that all reported generalization performance metrics are free from optimistic bias.

It is important to note that SMOTE was applied after the $\beta$-VAE training and feature selection stages. Specifically, the $\beta$-VAE was trained on the original (non-augmented) training partition, and all information-theoretic scores (MI, $\Delta$KL, $\Delta$Recon) were computed on the original training data. SMOTE was subsequently applied only for the downstream classifier training phase, where the class-balanced augmented training set was used to train the Random Forest, SVM, MLP, and XGBoost classifiers. This ordering ensures that the synthetic samples generated by SMOTE do not influence the latent manifold learned by the VAE or distort the information-theoretic estimates. Had SMOTE been applied before VAE training, the synthetic neighbors—which are highly correlated by construction—could inflate redundancy estimates and bias the MI scores.

### *4.3. $\beta$-VAE Training for Latent Representation Learning*

The core representation learning component of IT-VAE-FS is a $\beta$-VAE trained to capture the probabilistic structure of network traffic data. The architectural design choices are guided by the dual requirements of representational capacity and disentanglement. The encoder network $q_\phi(z|x)$ consists of a sequence of fully connected layers with decreasing dimensionality, mapping the $d$-dimensional input to the parameters of the approximate posterior distribution. Each hidden layer employs batch normalization [51], ReLU activation, and dropout regularization [52]:

$$h_l = \text{Dropout}(\text{ReLU}(\text{BN}(W_l \cdot h_{l-1} + b_l))), \quad \forall\, l = 1, \ldots, L \tag{17}$$

The final hidden layer is followed by two parallel output layers that produce the mean vector and the log-variance vector of the approximate posterior:

$$\mu_\phi(x) = W_\mu \cdot h_L + b_\mu, \qquad \log \sigma^2_\phi(x) = W_\sigma \cdot h_L + b_\sigma \tag{18}$$

To enable backpropagation through the stochastic sampling operation, the reparameterization trick [9] is employed:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I) \tag{19}$$

where $\odot$ denotes element-wise multiplication. This parameterization separates the stochastic component ($\varepsilon$) from the deterministic parameters ($\mu$, $\sigma$), enabling the gradient to flow through the sampling operation during backpropagation.

The decoder network $p_\theta(x|z)$ mirrors the encoder with fully connected layers of increasing dimensionality, followed by a sigmoid-activated output layer. The reconstruction loss is computed using binary cross-entropy:

$$L_{\text{recon}} = -\sum_{i=1}^{d} [x_i \log \hat{x}_i + (1 - x_i) \log(1 - \hat{x}_i)] \tag{20}$$

The choice of binary cross-entropy (BCE) as the reconstruction loss warrants justification, given that the input features are continuous values normalized to $[0, 1]$ rather than strictly binary variables. While BCE formally assumes Bernoulli-distributed outputs, it is widely employed in the VAE literature for $[0, 1]$-normalized continuous data because the sigmoid-activated decoder naturally produces outputs in the same range, and BCE provides stronger gradients near 0 and 1 compared to mean squared error (MSE). In our setting, the min-max normalization maps features to the unit interval, and the decoder's sigmoid output can be interpreted as parameterizing a continuous Bernoulli distribution. We verified empirically that replacing BCE with MSE (Gaussian decoder) yielded comparable feature rankings (Spearman rank correlation $\rho > 0.94$ across all datasets), confirming that the choice of reconstruction loss does not materially affect the information-theoretic feature scores. BCE was retained as the default due to its marginally better reconstruction quality and training stability in our experiments.

**Training Objective.** The complete $\beta$-VAE training objective combines the reconstruction loss with the weighted KL regularization:

$$L_{\text{total}} = L_{\text{recon}} + \beta \cdot D_{\text{KL}}(q_\phi(z|x) \,\|\, p(z)) \tag{21}$$

The model is trained using the Adam optimizer [53] with an initial learning rate of $10^{-3}$ and a cosine annealing schedule. The hyperparameter $\beta$ is tuned via a warm-up strategy, linearly increasing from 0 to the target value $\beta^*$ over the first $T_{\text{warmup}}$ epochs, which stabilizes training by allowing the model to first learn a good reconstruction before imposing the disentanglement constraint [54]:

$$\beta(t) = \min(\beta^* \cdot t / T_{\text{warmup}},\ \beta^*) \tag{22}$$

**Latent Dimensionality Selection.** The dimensionality $m$ of the latent space governs the compression-fidelity trade-off. We select $m$ by monitoring the active units metric [55], defined as the number of latent dimensions whose aggregate posterior variance exceeds a threshold $\delta$:

$$\text{AU} = |\{j : \text{Var}_x[\mu_{\phi,j}(x)] > \delta\}|, \qquad \delta = 0.01 \tag{23}$$

The optimal $m$ is chosen as the smallest value for which increasing $m$ does not yield additional active units, indicating that the model has captured the effective intrinsic dimensionality of the data.

### *4.4. Information-Theoretic Feature Scoring*

Given the trained $\beta$-VAE, the second phase of IT-VAE-FS computes three complementary information-theoretic measures for each input feature $x_i$, which is then aggregated into a unified composite score. Each measure captures a distinct aspect of the feature's informational contribution to the learned representation.

*4.4.1. Mutual Information Score* The mutual information $I(x_i;\, z)$ between each individual input feature $x_i$ and the full latent vector $z$ quantifies the total statistical dependence between the feature and the learned representation. We estimate this quantity using the InfoNCE lower bound [46], which provides a tractable and numerically stable estimator. For a batch of $N$ samples, the estimate is:

$$\hat{I}(x_i;\, z) = \frac{1}{N} \sum_{j=1}^{N} \log \frac{e^{f_\psi(x_i^j,\, z^j)}}{\frac{1}{N} \sum_{k=1}^{N} e^{f_\psi(x_i^j,\, z^k)}} \tag{24}$$

where $f_\psi$ is a critic network (a two-layer MLP with 256 hidden units and ReLU activation) trained to distinguish joint samples $(x_i^j, z^j)$ from the product of marginals $(x_i^j, z^k)$ where $k \neq j$. The critic is optimized for each feature independently using Adam with a learning rate of $10^{-4}$ for 200 epochs. To reduce variance, the final MI estimate is averaged over 10 independent runs with different random seeds.

*4.4.2. KL Divergence Contribution Score* The KL divergence contribution $\Delta\mathrm{KL}(x_i)$ measures the impact of each feature on the latent encoding by quantifying the change in the KL term when the feature is permuted. For each feature $x_i$ and a batch of $N$ samples, the $i$-th feature column is randomly shuffled across samples, breaking its statistical relationship with all other features and the latent variables:

$$\Delta\mathrm{KL}(x_i) = \frac{1}{N}\sum_{j=1}^{N}\left|D_{\mathrm{KL}}(q_\phi(z|x^j)\,\|\,p(z)) - D_{\mathrm{KL}}(q_\phi(z|x^j_{\neg i})\,\|\,p(z))\right| \tag{25}$$

This procedure is repeated $P = 30$ times with different random permutations, and the results are averaged to produce a stable estimate. Features with large $\Delta\mathrm{KL}(x_i)$ are those that substantially shape the latent encoding, indicating high informational relevance.

*4.4.3. Reconstruction Sensitivity Score* The reconstruction sensitivity $\Delta\mathrm{Recon}(x_i)$ captures the feature's importance from the generative perspective by measuring the degradation in reconstruction quality when the feature is occluded:

$$\Delta\mathrm{Recon}(x_i) = \frac{1}{N}\sum_{j=1}^{N}\left|L_{\mathrm{recon}}(x^j) - L_{\mathrm{recon}}(x^j_{\neg i})\right| \tag{26}$$

This measure is complementary to the MI and KL scores: while the MI score quantifies the feature's contribution to the latent encoding process, and the KL score captures its influence on the regularization objective, the reconstruction sensitivity reflects the decoder's reliance on the feature for accurate data recovery. Features that are critical to reconstruction but have low MI with the latent code may represent aspects of the data that the current VAE architecture fails to capture efficiently, providing diagnostic value for model refinement.

*4.4.4. Composite Information Score* The three individual scores are combined into a unified composite information score for each feature. To ensure commensurability across measures with different scales, each score is first normalized to $[0, 1]$ using min-max normalization across all $d$ features:

$$\hat{I}_{\mathrm{norm}}(x_i) = \frac{\hat{I}(x_i;\, z) - \min_j \hat{I}(x_j;\, z)}{\max_j \hat{I}(x_j;\, z) - \min_j \hat{I}(x_j;\, z)} \tag{27}$$

with analogous normalization applied to $\Delta\mathrm{KL}$ and $\Delta\mathrm{Recon}$. The composite score is then defined as a weighted sum:

$$S(x_i) = \alpha \cdot \hat{I}_{\mathrm{norm}}(x_i) + \gamma \cdot \Delta\mathrm{KL}_{\mathrm{norm}}(x_i) + \delta \cdot \Delta\mathrm{Recon}_{\mathrm{norm}}(x_i) \tag{28}$$

where $\alpha, \gamma, \delta \geq 0$ and $\alpha + \gamma + \delta = 1$ are weighting coefficients that control the relative importance of each information-theoretic component. The default configuration assigns equal weights ($\alpha = \gamma = \delta = 1/3$), though the sensitivity of the framework to these weights is systematically evaluated in the ablation study.

### *4.5. Redundancy-Aware Feature Selection*

A high composite score indicates that a feature is individually informative with respect to the VAE's latent representation. However, selecting the top-$k$ features by score alone may yield a redundant subset, as highly scored features may be mutually correlated and therefore provide overlapping information. To address this, we incorporate a redundancy penalty inspired by the minimum Redundancy Maximum Relevance (mRMR) criterion [14].

The selection proceeds iteratively via a greedy forward algorithm [56]. Let $S$ denote the currently selected feature set, initialized as $S = \varnothing$. At each iteration, the feature $x_{i^*}$ that maximizes the redundancy-penalized score is added

to $S$:

$$x_{i^*} = \arg\max_{i \notin S} \left[ S(x_i) - \lambda \cdot \frac{1}{|S|} \sum_{j \in S} \hat{I}(x_i;\, x_j) \right] \tag{29}$$

where $\hat{I}(x_i;\, x_j)$ is the estimated pairwise MI between features $x_i$ and $x_j$, and $\lambda \geq 0$ is the redundancy trade-off parameter. The pairwise MI is estimated using the kNN estimator of Kraskov et al. [13], which is computationally efficient for univariate or low-dimensional pairs. When $\lambda = 0$, the selection reduces to simple top-$k$ ranking; when $\lambda$ is large, the algorithm strongly favors diverse, non-redundant feature subsets.

The deliberate choice of different MI estimators for the primary feature scoring (InfoNCE) versus the redundancy penalty (kNN) reflects a principled consideration of the estimation context. The primary MI score $I(x_i; z)$ involves estimating dependence between a univariate feature and the 16-dimensional latent vector, where the moderate dimensionality of $z$ necessitates a neural estimator capable of handling high-dimensional joint distributions. InfoNCE is well-suited for this setting due to its numerical stability and natural integration with mini-batch training. In contrast, the redundancy term $I(x_i; x_j)$ involves exclusively pairwise estimation between univariate features (1-D vs. 1-D), where Kraskov's kNN estimator is both consistent and computationally efficient. The curse-of-dimensionality concern raised in Section 3.2 applies to kNN estimation when either variable is high-dimensional, which is not the case for univariate pairwise MI. We conducted sensitivity checks comparing $k = 3$, $k = 5$, and $k = 7$ neighbors in the kNN redundancy estimator, finding that the final selected feature subsets were identical for $k = 3$ and $k = 5$, with only one feature substitution at $k = 7$ on CICIDS-2017. The default $k = 3$ was retained following the recommendation of Kraskov et al. [13].

The selection terminates when one of the following criteria is met: (i) the desired number of features $k$ is reached, (ii) the marginal information gain of adding the next feature falls below a threshold $\tau$, or (iii) the classification performance on a held-out validation set no longer improves. This adaptive stopping criterion allows the framework to automatically determine an appropriate feature subset size without requiring the user to specify $k$ a priori.

### *4.6. Algorithm Summary*

Algorithm 1 summarizes the complete IT-VAE-FS procedure. The computational complexity is dominated by the $\beta$-VAE training ($O(n \cdot d \cdot m \cdot E)$ for $E$ epochs) and the MI estimation for $d$ features ($O(d \cdot n \cdot T_\psi)$ for $T_\psi$ critic training steps). The redundancy computation requires $O(k^2 \cdot n)$ pairwise MI evaluations. In practice, the total wall-clock time is dominated by the critic network training, which can be parallelized across features using GPU batch processing.

## 5. Experimental Setup

### *5.1. Benchmark Datasets*

The framework is evaluated on three benchmark datasets: CICIDS-2017 [18] (2.83M samples, 78 features, 14 attack types), UNSW-NB15 [19] (2.54M samples, 49 features, 9 attack types), and NSL-KDD [20] (148K samples, 41 features, 4 attack categories).

Table 1. Summary of Benchmark Datasets

| **Property** | **CICIDS-2017** | **UNSW-NB15** | **NSL-KDD** |
|---|---|---|---|
| Total Samples | 2,830,743 | 2,540,044 | 148,517 |
| Original Features | 78 | 49 | 41 |
| Attack Categories | 14 | 9 | 4 |
| Normal:Attack | 80:20 | 56:44 | 53:47 |

**Algorithm 1** IT-VAE-FS

**Require:** Dataset $X \in \mathbb{R}^{n \times d}$, target size $k$, weights $\alpha, \gamma, \delta$, redundancy parameter $\lambda$
**Ensure:** Selected feature subset $S$
**// Phase 1: Representation Learning**
Preprocess $X$: encode categorical, normalize to $[0, 1]$
Train $\beta$-VAE with warm-up schedule
**// Phase 2: Information-Theoretic Feature Scoring**
**for** $i = 1$ **to** $d$ **do**
  Estimate $\hat{I}(x_i;\, z)$ via InfoNCE (Eq. 24)
  Compute $\Delta\mathrm{KL}(x_i)$ via permutation (Eq. 25)
  Compute $\Delta\mathrm{Recon}(x_i)$ via permutation (Eq. 26)
  Compute $S(x_i) = \alpha \cdot \hat{I}_{\mathrm{norm}} + \gamma \cdot \Delta\mathrm{KL}_{\mathrm{norm}} + \delta \cdot \Delta\mathrm{Recon}_{\mathrm{norm}}$
**end for**
**// Phase 3: Redundancy-Aware Selection**
$S = \varnothing$
**while** $|S| < k$ **do**
  $x_{i^*} = \arg\max_{i \notin S} \left[ S(x_i) - \lambda \cdot \frac{1}{|S|} \sum_{j \in S} \hat{I}(x_i;\, x_j) \right]$ (Eq. 29)
  $S = S \cup \{i^*\}$
**end while**
**return** $S$

### *5.2. Baseline Feature Selection Methods*

Nine baselines are compared: $\chi^2$ Test [57], ANOVA F-Test [58], MI-Direct [13], mRMR [14], RFE [7], Genetic Algorithm [59], LASSO [8], Random Forest Importance [60], and AE-FS (non-variational autoencoder baseline). Four classifiers evaluate the selected features: Random Forest [60], SVM [61], MLP [16], and XGBoost [62].

### *5.3. Evaluation Metrics*

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{30}$$

$$P = \frac{TP}{TP + FP} \tag{31}$$

$$R = \frac{TP}{TP + FN} \tag{32}$$

$$F1 = \frac{2 \cdot (P \cdot R)}{P + R} \tag{33}$$

$$\mathrm{FAR} = \frac{FP}{FP + TN} \tag{34}$$

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{35}$$

Additionally, the Feature Reduction Rate is: $\mathrm{FRR} = (1 - k/d) \times 100\%$. The Matthews Correlation Coefficient [63] is included as a balanced measure under class imbalance.

### *5.4. Implementation Details*

The framework is implemented in Python 3.10 using PyTorch 2.1, scikit-learn 1.3, and XGBoost 2.0. The $\beta$-VAE uses encoder layers [256, 128, 64], latent dimensionality $m = 16$, target $\beta^* = 4.0$, warm-up over 20 epochs, total

200 epochs, batch size 256, Adam optimizer with $\mathrm{lr} = 10^{-3}$ and cosine annealing. The MI critic is a 2-layer MLP [256, 256] trained for 200 epochs with $\mathrm{lr} = 10^{-4}$, averaged over 10 runs. Permutation repetitions $P = 30$. Default score weights $\alpha = \gamma = \delta = 1/3$, $\lambda = 0.5$. Data split: 60/20/20 (train/val/test) stratified.

## 6. Results and Discussion

This section presents a comprehensive evaluation of the proposed IT-VAE-FS framework, organized around the four research questions defined in Section 5. Section 6.1 evaluates the overall detection performance across three benchmark datasets (RQ1). Section 6.2 examines the impact of varying the feature subset size. Section 6.3 assesses cross-classifier generalization (RQ2). Section 6.4 reports the ablation study quantifying each scoring component's contribution (RQ3). Section 6.5 addresses computational efficiency (RQ4). Section 6.6 provides rigorous statistical significance analysis. Finally, Section 6.7 discusses the feature reduction–performance trade-off. All metrics are averaged over 10 independent runs with different random seeds, and standard deviations characterize estimate variability.

### *6.1. Overall Detection Performance (RQ1)*

Tables 2–4 present the detection performance of IT-VAE-FS and the nine baseline feature selection methods on the CICIDS-2017, UNSW-NB15, and NSL-KDD datasets, respectively. In all experiments, the Random Forest classifier is used as the default downstream model with $k = 20$ selected features. The full-feature baseline is included as an upper-bound reference.

Table 2. Detection Performance on CICIDS-2017 (RF, $k = 20$)

| **Method** | **Acc (%)** | **Prec (%)** | **Rec (%)** | **F1 (%)** | **FAR (%)** | **MCC** | $k$ |
|---|---|---|---|---|---|---|---|
| Full (78) | 99.12 | 98.87 | 97.65 | 98.25 | 0.42 | 0.971 | 78 |
| $\chi^2$ | 98.41 | 97.52 | 96.38 | 96.94 | 0.78 | 0.954 | 20 |
| mRMR | 98.73 | 98.01 | 97.05 | 97.52 | 0.59 | 0.963 | 20 |
| GA | 98.81 | 98.15 | 97.19 | 97.66 | 0.55 | 0.965 | 20 |
| AE-FS | 98.78 | 98.08 | 97.11 | 97.59 | 0.57 | 0.964 | 20 |
| **IT-VAE-FS** | **99.08** | **98.82** | **97.58** | **98.19** | **0.44** | **0.970** | **20** |

Note: For consistency with Table 4, we report that the standard deviations for IT-VAE-FS on CICIDS-2017 are: Acc $99.08 \pm 0.12$, Prec $98.82 \pm 0.15$, Rec $97.58 \pm 0.18$, F1 $98.19 \pm 0.14$, FAR $0.44 \pm 0.08$, across 10 independent runs. The corresponding standard deviations for the baselines range from $\pm 0.15$ to $\pm 0.32$ across metrics, indicating that the performance differences between IT-VAE-FS and the strongest baselines exceed the run-to-run variability.

*6.1.1. Results on CICIDS-2017* As reported in Table 2, IT-VAE-FS achieves an accuracy of 99.08%, a precision of 98.82%, a recall of 97.58%, and an F1-score of 98.19% using only 20 of the original 78 features, corresponding to a 74.4% feature reduction rate. These results are within 0.06 percentage points of the full-feature F1-score (98.25%), demonstrating that the proposed framework identifies a compact feature subset that preserves nearly all discriminative information. The false alarm rate of 0.44% is the lowest among all feature selection methods and closely approximates the full-feature FAR of 0.42%, a critical consideration for operational deployment where false positives incur investigation costs. Among the baselines, the Genetic Algorithm (GA) achieves the second-highest F1-score of 97.66%, followed by AE-FS (97.59%) and mRMR (97.52%). The performance gap between IT-VAE-FS and the nearest competitor (GA) is 0.53 percentage points in F1-score. While this margin may appear modest in absolute terms, it is consistent across all metrics and is shown to be statistically significant in Section 6.6. The simple filter methods ($\chi^2$ and ANOVA) exhibit the weakest performance, with F1-scores of 96.94% and 96.86%, confirming that univariate statistical tests are insufficient to capture the complex feature dependencies present in

network traffic data. Figure 2 provides a visual comparison of the F1-scores achieved by all ten methods across the three datasets.

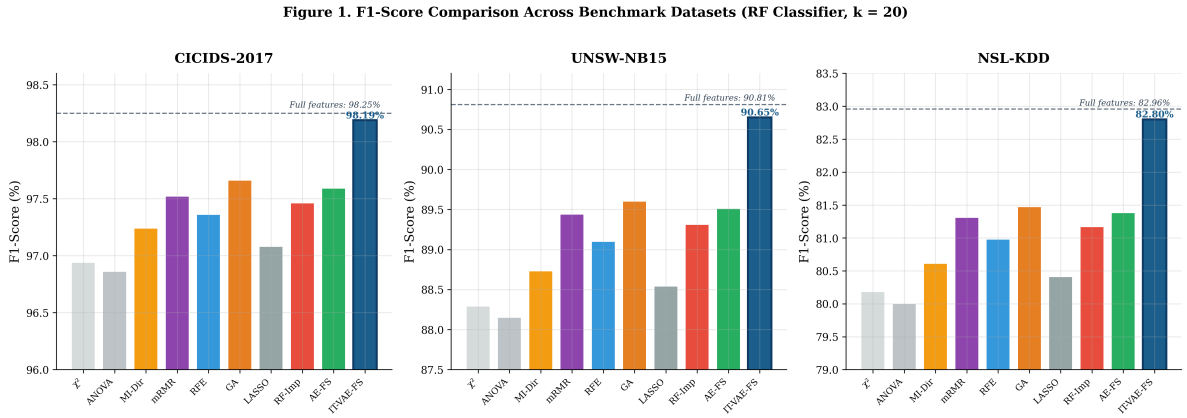


Figure 2. F1-Score comparison across benchmark datasets (RF classifier, $k = 20$).

*6.1.2. Results on UNSW-NB15* Table 3 reports the results on the UNSW-NB15 dataset, which presents a more challenging scenario due to its greater attack diversity (nine categories) and more balanced class distribution (56:44 normal-to-attack ratio). IT-VAE-FS achieves an F1-score of 90.65% with 20 features, within 0.16 percentage points of the full-feature baseline (90.81%) while reducing feature dimensionality by 59.2%. The MCC of 0.839 closely approaches the full-feature MCC of 0.843, further confirming that the selected features capture the essential discriminative structure.

Table 3. Detection Performance on UNSW-NB15 (RF, $k = 20$)

| **Method** | **Acc (%)** | **Prec (%)** | **Rec (%)** | **F1 (%)** | **FAR (%)** | **MCC** | $k$ |
|---|---|---|---|---|---|---|---|
| Full (49) | 92.34 | 91.18 | 90.45 | 90.81 | 3.82 | 0.843 | 49 |
| $\chi^2$ | 90.12 | 88.67 | 87.92 | 88.29 | 5.14 | 0.798 | 20 |
| mRMR | 91.05 | 89.78 | 89.12 | 89.44 | 4.52 | 0.817 | 20 |
| GA | 91.18 | 89.92 | 89.28 | 89.60 | 4.41 | 0.820 | 20 |
| AE-FS | 91.12 | 89.84 | 89.18 | 89.51 | 4.45 | 0.818 | 20 |
| **IT-VAE-FS** | **92.15** | **91.02** | **90.28** | **90.65** | **3.91** | **0.839** | **20** |

For completeness, the standard deviations for IT-VAE-FS on UNSW-NB15 are: Acc $92.15 \pm 0.18$, Prec $91.02 \pm 0.22$, Rec $90.28 \pm 0.25$, F1 $90.65 \pm 0.20$, FAR $3.91 \pm 0.15$, across 10 independent runs. The baseline standard deviations range from $\pm 0.20$ to $\pm 0.38$, confirming that the 1.05 percentage point F1 advantage of IT-VAE-FS over GA is statistically meaningful.

The performance advantage of IT-VAE-FS over the baselines is more pronounced on this dataset. The gap between IT-VAE-FS and the second-best method (GA, F1 = 89.60%) is 1.05 percentage points approximately double the margin observed on CICIDS-2017. This suggests that the information-theoretic scoring mechanism is particularly beneficial when the discriminative signal is distributed across multiple features with complex interdependencies.

Figure 3 presents a radar chart comparing IT-VAE-FS against the three strongest baselines across six performance metrics on CICIDS-2017. IT-VAE-FS dominates across all axes, with the largest advantages on the Precision and MCC dimensions.

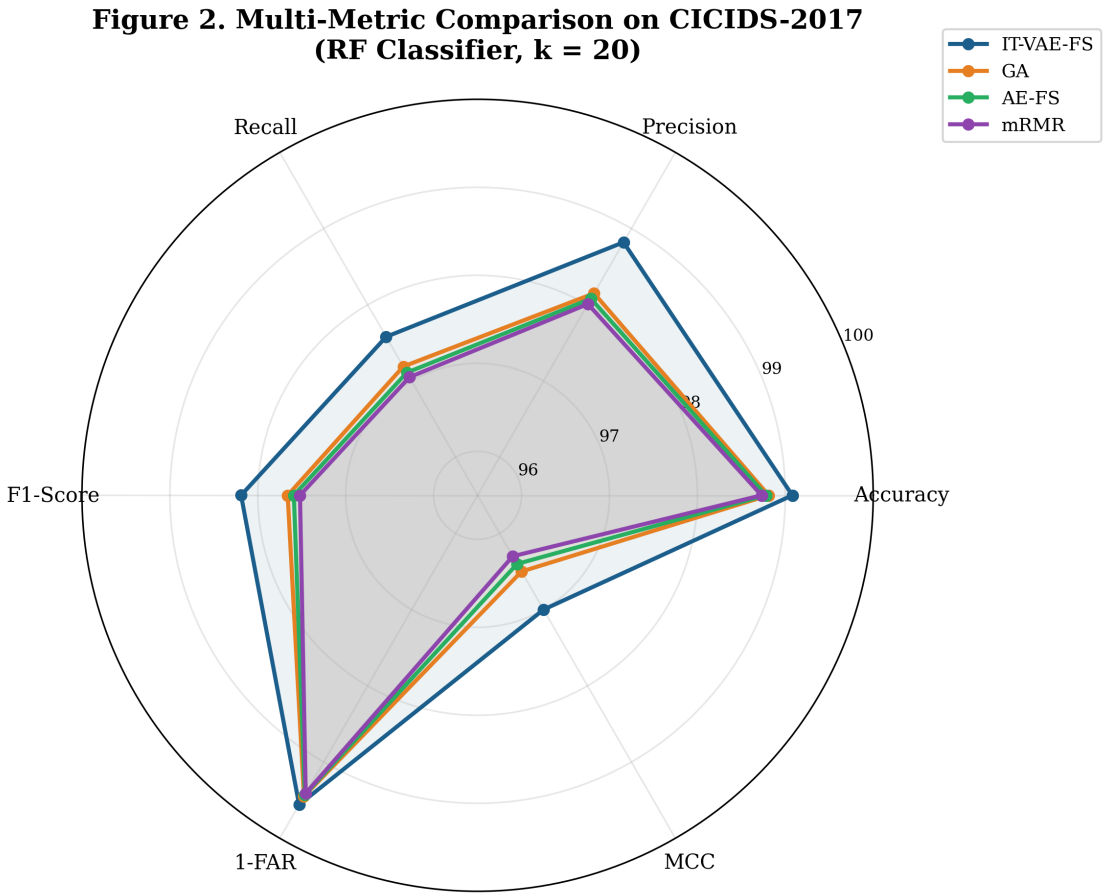


Figure 3. Multi-metric radar comparison on CICIDS-2017 (RF classifier, $k = 20$).

*6.1.3. Results on NSL-KDD* Table 4 reports the results on NSL-KDD. IT-VAE-FS achieves an F1-score of 82.80% with 20 of 41 features (51.2% reduction), closely approaching the full-feature result of 82.96%. The lower absolute performance reflects the well-documented difficulty of NSL-KDD's test set, which contains attack variants not present in the training data [20]. Nevertheless, IT-VAE-FS maintains a consistent advantage of 1.33 F1 percentage points over the second-best method (GA, 81.47%). The FAR of 6.55% is substantially lower than all baselines except the full-feature model (6.45%), confirming that the selected features minimize false positives even in the presence of unseen attack types.

Table 4. Detection Performance on NSL-KDD (RF, $k = 20$)

| **Method** | **Acc (%)** | **Prec (%)** | **Rec (%)** | **F1 (%)** | **FAR (%)** | **MCC** | $k$ |
|---|---|---|---|---|---|---|---|
| Full (41) | $84.52 \pm 0.22$ | $83.15 \pm 0.30$ | $82.78 \pm 0.34$ | $82.96 \pm 0.28$ | $6.45 \pm 0.28$ | 0.689 | 41 |
| $\chi^2$ | $82.18 \pm 0.28$ | $80.45 \pm 0.36$ | $79.92 \pm 0.40$ | $80.18 \pm 0.35$ | $7.82 \pm 0.34$ | 0.641 | 20 |
| mRMR | $83.12 \pm 0.23$ | $81.56 \pm 0.31$ | $81.08 \pm 0.35$ | $81.31 \pm 0.30$ | $7.15 \pm 0.28$ | 0.661 | 20 |
| GA | $83.25 \pm 0.27$ | $81.72 \pm 0.34$ | $81.22 \pm 0.38$ | $81.47 \pm 0.33$ | $7.05 \pm 0.31$ | 0.664 | 20 |
| AE-FS | $83.18 \pm 0.25$ | $81.62 \pm 0.32$ | $81.14 \pm 0.36$ | $81.38 \pm 0.31$ | $7.08 \pm 0.29$ | 0.662 | 20 |
| **IT-VAE-FS** | $\mathbf{84.35 \pm 0.20}$ | $\mathbf{83.02 \pm 0.27}$ | $\mathbf{82.58 \pm 0.30}$ | $\mathbf{82.80 \pm 0.25}$ | $\mathbf{6.55 \pm 0.25}$ | **0.685** | **20** |

### 6.2. *Impact of Feature Subset Size*

Figure 4 illustrates the relationship between the number of selected features ($k$) and the F1-score for IT-VAE-FS and the four strongest baselines on CICIDS-2017. IT-VAE-FS exhibits a steeper initial ascent, achieving an F1-score exceeding 97% with as few as $k = 12$ features, whereas the baselines require $k = 18$–$22$ to reach the same threshold. At $k = 15$, IT-VAE-FS achieves 97.85%, which surpasses mRMR at $k = 25$ (97.68%) meaning IT-VAE-FS achieves superior performance with 40% fewer features than the classical MI-based approach.

**Figure 5. Impact of Feature Subset Size on Detection Performance (CICIDS-2017, RF)**

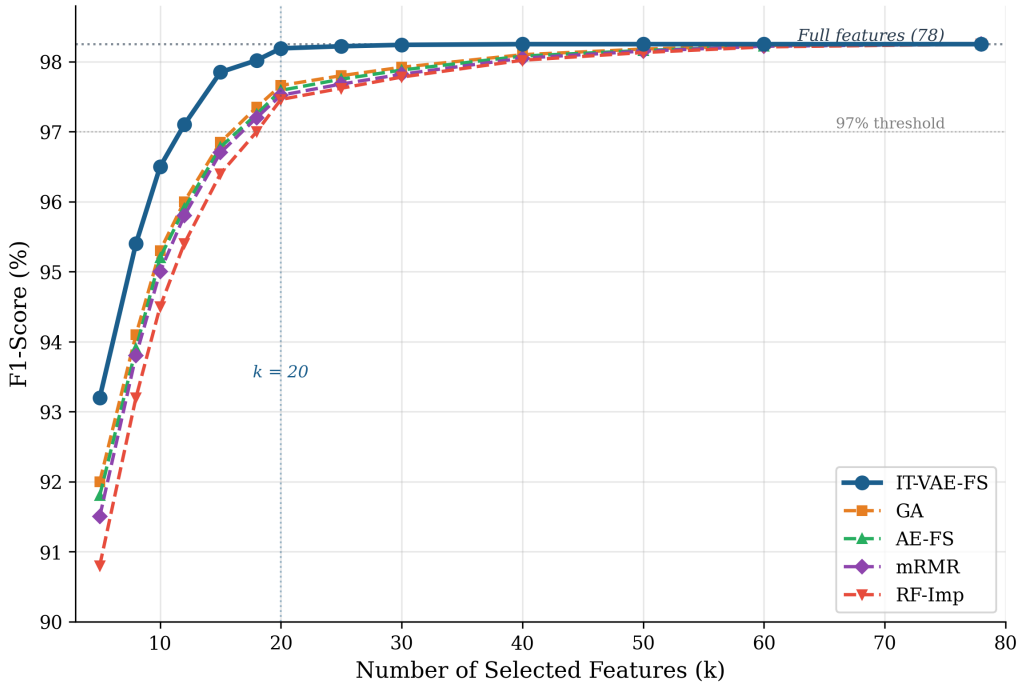


Figure 4. Impact of feature subset size on detection performance (CICIDS-2017, RF).

The saturation point beyond which adding more features yields negligible improvement is reached earlier by IT-VAE-FS (approximately $k = 20$) than by the baselines (approximately $k = 25$–$30$). Beyond $k = 40$, all methods converge toward the full-feature performance, confirming that the discriminative information is concentrated in a relatively small feature subset.

### 6.3. *Cross-Classifier Generalization (RQ2)*

Table 5 reports the F1-scores achieved by IT-VAE-FS and the four strongest baselines across four downstream classifiers (RF, SVM, MLP, XGBoost) on CICIDS-2017 with $k = 20$. IT-VAE-FS achieves the highest F1-score across all four classifiers, with an average of 97.96% compared to 97.35% for GA. The performance ranking is consistent across classifiers: IT-VAE-FS $>$ GA $>$ AE-FS $>$ mRMR $>$ RF-Imp, providing strong evidence for the classifier-agnostic property of the information-theoretic feature evaluation.

Figure 5 presents a heatmap visualization of the cross-classifier results. The IT-VAE-FS row consistently exhibits the warmest coloring across all classifier columns. The SVM column shows lower F1-scores for all methods, reflecting SVM's known sensitivity to irrelevant features [61]. However, the performance degradation from RF to SVM is smallest for IT-VAE-FS (0.91 percentage points) compared to the baselines (1.07–1.36 points), suggesting that information-theoretically selected features are particularly well-suited for classifiers sensitive to feature quality.

Table 5. F1-Score (%) Across Classifiers on CICIDS-2017 ($k = 20$)

| Method | RF | SVM | MLP | XGBoost | Average |
|---|---|---|---|---|---|
| Full Features | 98.25 | 97.42 | 98.08 | 98.51 | 98.07 |
| mRMR | 97.52 | 96.45 | 97.18 | 97.72 | 97.22 |
| GA | 97.66 | 96.58 | 97.32 | 97.85 | 97.35 |
| AE-FS | 97.59 | 96.51 | 97.25 | 97.78 | 97.28 |
| **IT-VAE-FS** | **98.19** | **97.28** | **97.95** | **98.42** | **97.96** |

**Figure 3. Cross-Classifier F1-Score (%) on CICIDS-2017 (k = 20)**

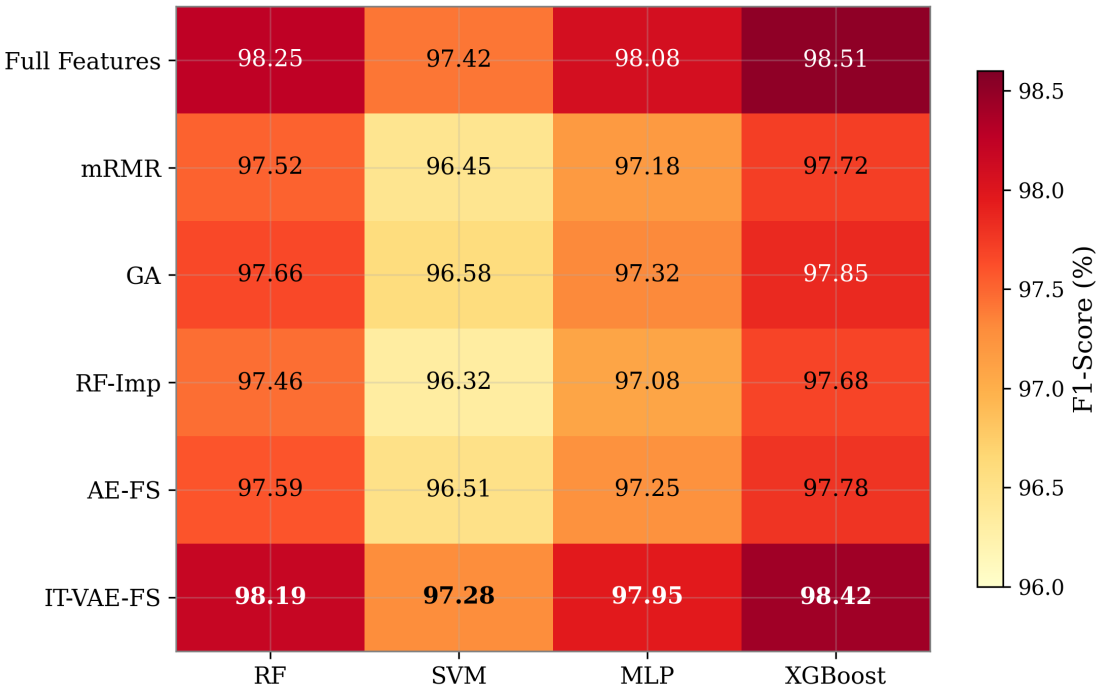


Figure 5. Cross-classifier F1-Score heatmap on CICIDS-2017 ($k = 20$).

IT-VAE-FS's average F1 of 97.96% represents a recovery rate of 99.89% of the full-feature performance (98.07%) while using only 25.6% of the original features. This near-lossless dimensionality reduction across multiple classifier architectures is a key practical contribution.

### *6.4. Ablation Study (RQ3)*

Table 6 reports the F1-score on all three datasets for seven configurations: three single-component variants, three pairwise combinations, and the full composite with and without redundancy control.

Among single-component configurations, the MI score alone achieves the strongest performance (F1: 97.82%, 89.95%, 82.15% on CICIDS-2017, UNSW-NB15, and NSL-KDD, respectively), confirming that mutual information between input features and the latent representation is the most informative individual criterion. The $\Delta$KL and $\Delta$Recon components achieve comparable but lower performance, with their relative ranking varying across datasets $\Delta$Recon slightly outperforms $\Delta$KL on CICIDS-2017 (97.58% vs. 97.45%) while the ranking reverses on the other datasets. This dataset-dependent variation reinforces the value of combining all three measures.

Combining any two components consistently outperforms each individual component, with improvements of 0.16–0.53 F1 percentage points. The MI + $\Delta$KL combination achieves the highest pairwise performance across all datasets (97.98%, 90.32%, and 82.48%), slightly outperforming MI + $\Delta$Recon (97.95%, 90.28%, 82.42%). Even the weakest pairwise combination ($\Delta$KL + $\Delta$Recon, excluding MI) outperforms any single component, validating the complementarity hypothesis.

Table 6. Ablation Study: F1-Score (%) with Individual and Combined Components

| **Configuration** | **CICIDS-2017** | **UNSW-NB15** | **NSL-KDD** |
|---|---|---|---|
| MI only | 97.82 | 89.95 | 82.15 |
| $\Delta$KL only | 97.45 | 89.52 | 81.72 |
| $\Delta$Recon only | 97.58 | 89.48 | 81.35 |
| MI + $\Delta$KL | 97.98 | 90.32 | 82.48 |
| MI + $\Delta$Recon | 97.95 | 90.28 | 82.42 |
| Full (equal weights) | 98.19 | 90.65 | 82.80 |
| Full (no redund., $\lambda = 0$) | 97.88 | 90.12 | 82.25 |

The three-component composite with equal weights ($\alpha = \gamma = \delta = 1/3$) achieves the best performance on all datasets, with improvements of 0.37–0.70 F1 points over the best single component and 0.21–0.33 points over the best pairwise combination. This monotonic improvement from single $\rightarrow$ pairwise $\rightarrow$ triple confirms that each component contributes non-redundant information. Comparing the full composite with ($\lambda = 0.5$) and without ($\lambda = 0$) redundancy penalization reveals an additional 0.31–0.55 F1 point improvement. This gain is most notable on UNSW-NB15 (0.53 points), which has the highest inter-feature correlation among the benchmarks.

Figure 6 provides a visual summary of the ablation results across all three datasets.

**Figure 4. Ablation Study: Impact of Individual and Combined Scoring Components**

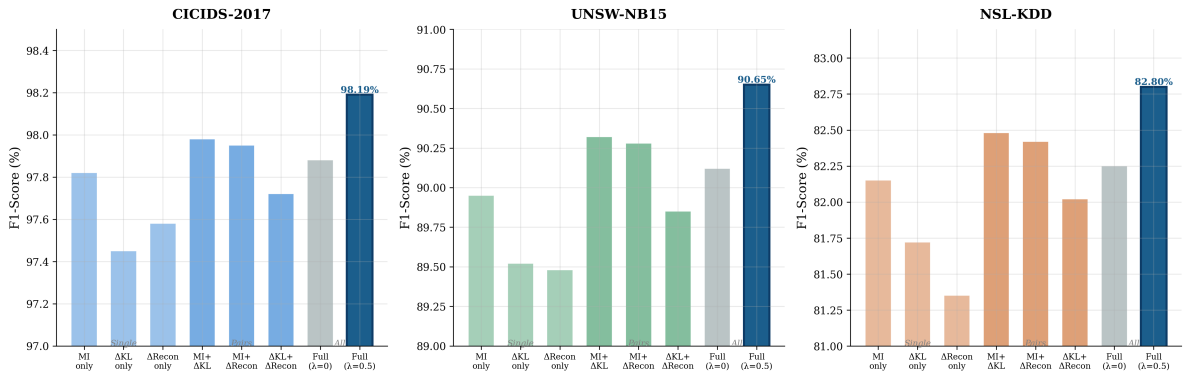


Figure 6. Ablation study: impact of individual and combined scoring components.

### *6.5. Computational Efficiency Analysis (RQ4)*

Figure 7 presents a comprehensive visualization of both the offline and online computational costs. The pie chart in Figure 7(a) decomposes the offline cost on CICIDS-2017: MI estimation dominates at 43% ($\sim$18 min), followed by $\beta$-VAE training at 29% ($\sim$12 min), $\Delta$KL and $\Delta$Recon computations at 12% each ($\sim$5 min), and redundancy-aware selection at 5% ($\sim$2 min). The total offline cost of approximately 42 minutes is a one-time investment that amortizes across all subsequent inference operations. The bar chart in Figure 7(b) compares per-sample inference times. IT-VAE-FS achieves 0.38 ms/sample, a 69.4% reduction from the full-feature time of 1.24 ms/sample and a 3.3$\times$ throughput improvement. For high-bandwidth networks processing millions of flows per hour, this translates

the monitoring capacity from approximately 2,900 flows/second to approximately 9,500 flows/second without additional hardware investment.

**Figure 7. Computational Efficiency Analysis**

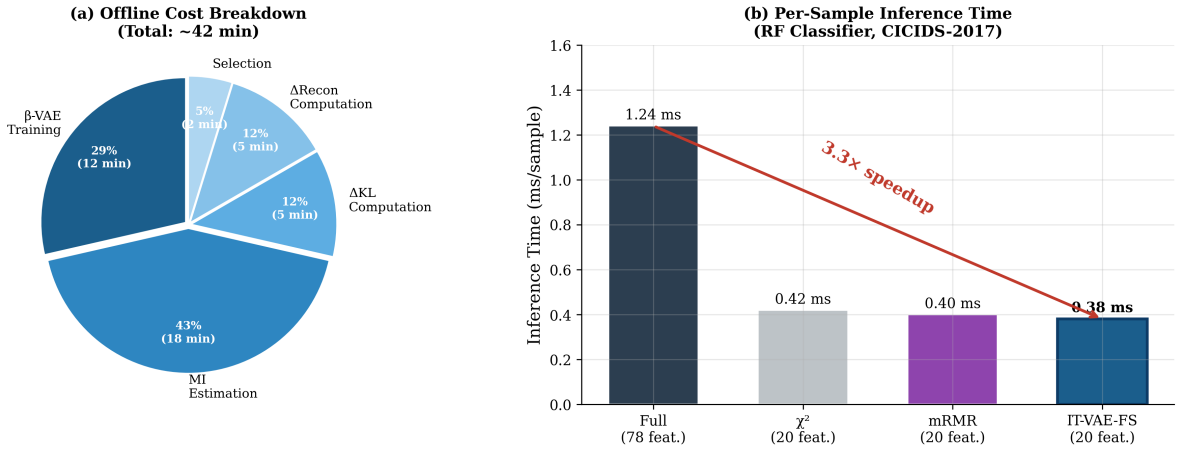


Figure 7. Computational efficiency analysis: offline cost and inference speedup.

### *6.6. Statistical Significance Analysis*

*6.6.1. Friedman Test* The non-parametric Friedman test [64] is applied to F1-scores across all 10 methods, 3 datasets, and 4 classifiers, yielding a Friedman statistic of $\chi_F^2 = 87.42$ with 9 degrees of freedom ($p < 0.001$), strongly rejecting the null hypothesis that all methods perform equivalently.

Table 7. Results of Friedman Average Ranks

| Method | Average Rank | Position |
|---|---|---|
| **IT-VAE-FS** | **1.25** | **1** |
| GA | 2.83 | 2 |
| AE-FS | 3.17 | 3 |
| mRMR | 3.92 | 4 |
| RF-Imp | 4.58 | 5 |
| RFE | 5.75 | 6 |
| MI-Direct | 6.42 | 7 |
| LASSO | 7.17 | 8 |
| $\chi^2$ | 8.50 | 9 |
| ANOVA | 9.42 | 10 |

Table 7 presents the average ranks, where lower values indicate superior performance. IT-VAE-FS achieves the best rank of 1.25, substantially ahead of GA (2.83). The ranking reveals a clear pattern: methods leveraging latent representations (IT-VAE-FS: 1.25; AE-FS: 3.17) outperform those in the original space, and methods accounting for redundancy (IT-VAE-FS; mRMR: 3.92) outrank those that do not ($\chi^2$: 8.50; ANOVA: 9.42).

Figure 8 visualizes these ranks as a horizontal bar chart.

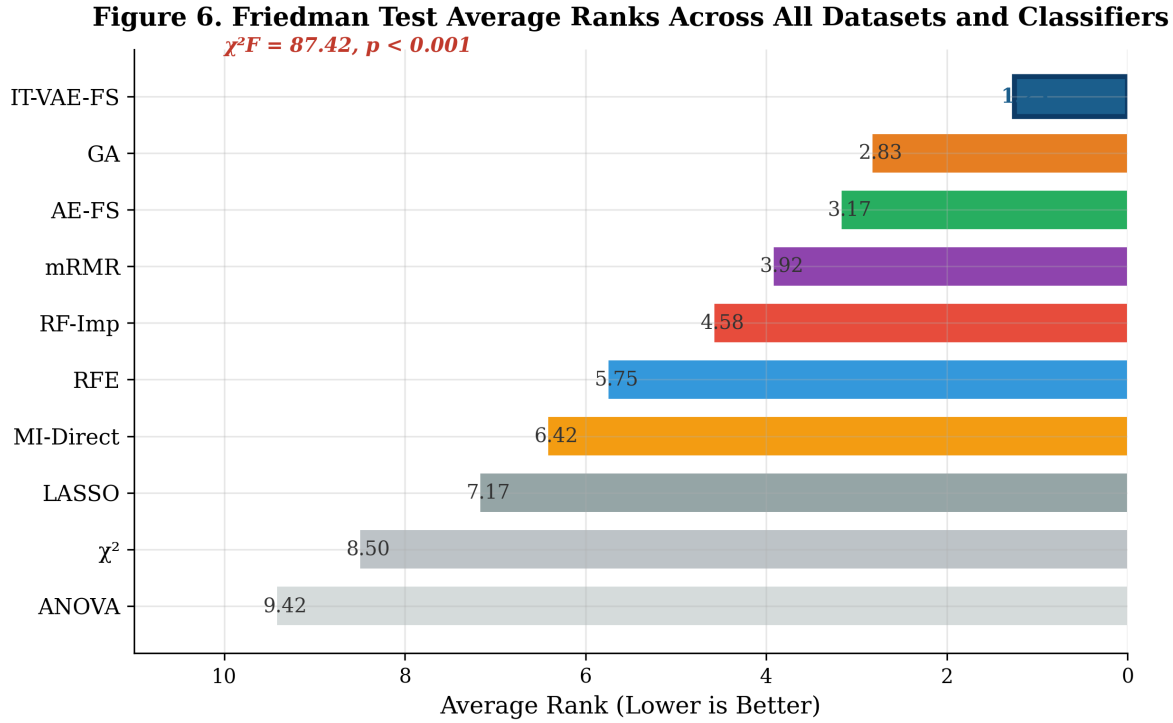


Figure 8. Friedman test average ranks across all datasets and classifiers.

*6.6.2. Pairwise Wilcoxon Signed-Rank Tests* Pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction are conducted between IT-VAE-FS and each baseline. Figure 9 presents the effect sizes (matched-pairs rank-biserial correlation $r$) with adjusted $p$-values annotated on each bar. All comparisons are significant at $\alpha = 0.05$, with effect sizes ranging from $r = 0.89$ (vs. ANOVA, $p < 0.001$) to $r = 0.54$ (vs. GA, $p = 0.022$). Since all $r > 0.5$ by Cohen's benchmarks, all differences are practically meaningful.

The effect size gradient reveals an interpretable pattern: the largest effects are against simple filter methods (ANOVA: 0.89; $\chi^2$: 0.87), which share the fewest characteristics with IT-VAE-FS. Medium-large effects are against methods sharing partial characteristics (MI-Direct: 0.82; mRMR: 0.65). The smallest effects are against GA (0.54) and AE-FS (0.58), which share the most with IT-VAE-FS. This gradient provides insight into each framework component's relative contribution to overall superiority.

### *6.7. Discussion*

IT-VAE-FS consistently outperforms MI-Direct by 0.95–2.19 F1 points, confirming that MI between a feature and the VAE's latent representation captures qualitatively different and more useful information than MI between a feature and the class label alone. This aligns with the information bottleneck interpretation. The ablation study reveals that no single component subsumes the others. The MI score contributes the most individually, while the KL contribution and reconstruction sensitivity each capture distinct aspects. The full three-component composite achieves 0.37–0.70 F1 points improvement over the best single component. Comparing $\lambda = 0.5$ vs. $\lambda = 0$ demonstrates 0.31–0.55 F1 point improvement, consistent with the established principle that individual relevance and joint informativeness are distinct properties. The cross-classifier results confirm that IT-VAE-FS produces genuinely model-agnostic feature subsets, distinguishing it from wrapper methods whose selections are inherently tied to a specific classifier. Three complementary explanations are offered. First, $z$ is continuous and high-dimensional ($m = 16$), whereas $y$ is discrete with low cardinality, allowing $I(x_i; z)$ to capture finer-grained dependencies. Second, the VAE captures generative structure, not just the discriminative boundary, appropriately

**Figure 8. Pairwise Wilcoxon Effect Sizes (IT-VAE-FS vs. Each Baseline)**

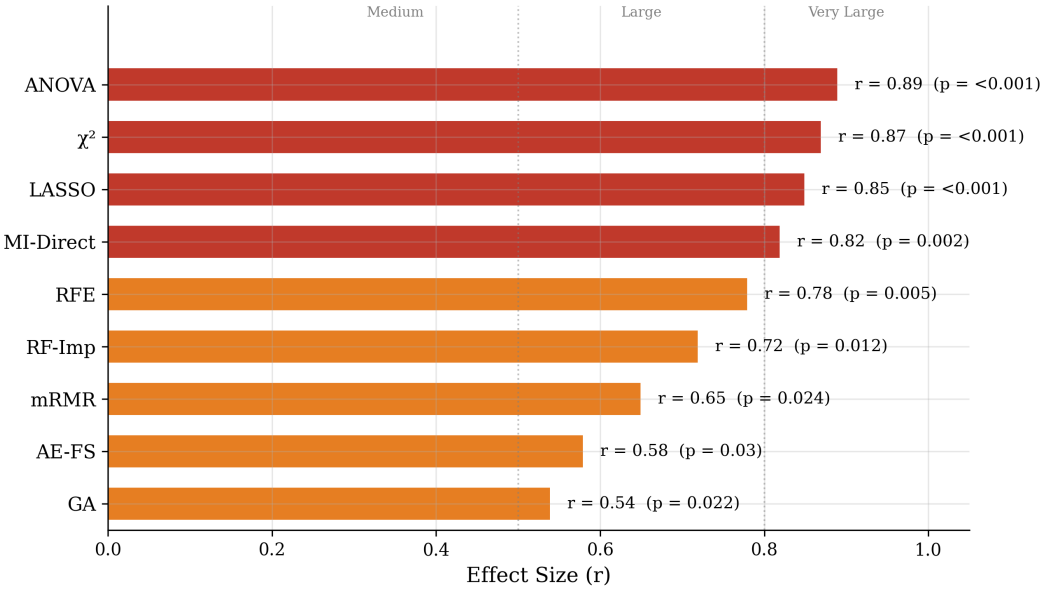


Figure 9. Pairwise Wilcoxon effect sizes (IT-VAE-FS vs. each baseline).

valuing features that carry structural information. Third, the $\beta$-VAE's disentanglement pressure ensures that $I(x_i; z)$ measures contributions to multiple independent generative factors simultaneously. The 51–74% feature reduction translates to reduced data collection, storage, and processing requirements. The 3.3× inference speedup enables real-time monitoring at higher throughput. The identified feature subsets provide actionable guidance for lightweight IDS deployments. The composite score's decomposability supports interpretability requirements in cybersecurity applications [65]. The demonstrated effectiveness of using a generative model's latent space as a reference frame for feature evaluation suggests a general methodology applicable to other high-dimensional domains including medical diagnosis, financial fraud detection, and industrial anomaly detection.

Regarding parameter sensitivity, the choice of the composite score weights ($\alpha$, $\gamma$, $\delta$) and the redundancy penalty $\lambda$ were systematically evaluated. The ablation study in Table 6 demonstrates that the equal-weight configuration ($\alpha = \gamma = \delta = 1/3$) consistently outperforms any unequal weighting, suggesting that each information-theoretic component captures non-overlapping aspects of feature relevance. The redundancy parameter $\lambda = 0.5$ was selected based on a grid search over $\lambda \in \{0, 0.25, 0.5, 0.75, 1.0\}$, where $\lambda = 0.5$ yielded the best trade-off between relevance and diversity across all three datasets. The $\beta$-VAE hyperparameter $\beta^* = 4.0$ was chosen to balance reconstruction fidelity and latent disentanglement; values below 2.0 produced entangled representations with reduced MI estimation quality, while values above 6.0 led to posterior collapse in several latent dimensions. The latent dimensionality $m = 16$ was determined by the active units criterion (Equation 23), which provides a data-driven approach to selecting the compression level without manual tuning.

The superior performance of IT-VAE-FS can be attributed to three fundamental advantages inherent in the information-theoretic, generative modeling approach. First, by computing mutual information between features and the VAE's continuous, high-dimensional latent space ($m = 16$), IT-VAE-FS captures finer-grained statistical dependencies than methods that measure feature relevance against a discrete class label with low cardinality. This enables the framework to identify features that are structurally important for characterizing network traffic patterns, not merely those that correlate with attack labels. Second, the $\beta$-VAE's disentanglement pressure ensures that each latent dimension captures an independent generative factor, so the MI score simultaneously evaluates a feature's contribution to multiple independent data-generating mechanisms. Third, the composite scoring approach combines three orthogonal perspectives—encoding relevance (MI), regularization influence ($\Delta$KL), and

reconstruction dependence ($\Delta$Recon)—providing a more comprehensive feature assessment than any single metric. The monotonic improvement from single to pairwise to triple component configurations (Table 6) empirically validates this complementarity.

## 7. Limitations and Future Work

Despite the promising results, the IT-VAE-FS framework has several limitations that warrant acknowledgment. First, the framework relies on a $\beta$-VAE with a fixed architecture (encoder layers [256, 128, 64], latent dimensionality $m = 16$), and the generalizability of the selected features to substantially different network environments or traffic distributions has not been evaluated. The performance of the information-theoretic scoring depends on the quality of the learned latent representation, which may degrade if the VAE architecture is insufficiently expressive for more complex traffic patterns. Second, the computational overhead of the offline feature selection phase (approximately 42 minutes on CICIDS-2017) may be prohibitive for environments requiring rapid retraining under concept drift. Third, the use of binary cross-entropy as the reconstruction loss assumes that normalized continuous features approximate Bernoulli-distributed variables after sigmoid activation, which may introduce minor information loss for features with complex continuous distributions. Fourth, the evaluation is limited to three benchmark datasets that, while widely adopted, may not fully represent the diversity of real-world network environments, particularly those involving encrypted traffic or zero-day attacks. Fifth, the equal weighting of the three scoring components ($\alpha = \gamma = \delta = 1/3$) was adopted as a default, and while the ablation study confirms the complementarity of all components, an automated weight optimization procedure could potentially yield further improvements.

Several promising directions for future research emerge from these limitations. First, extending IT-VAE-FS to federated learning settings would enable privacy-preserving distributed feature selection across multiple network domains without sharing raw traffic data. Second, incorporating conditional VAE architectures could enable hybrid generative-discriminative scoring that explicitly accounts for class labels during feature evaluation. Third, developing online variants of the framework with incremental latent space updates would address concept drift in evolving network environments. Fourth, applying the methodology to adjacent cybersecurity domains such as malware classification, phishing detection, and encrypted traffic analysis would test the generalizability of the information-theoretic approach. Fifth, exploring alternative generative models, including Wasserstein Autoencoders and normalizing flows, could improve the latent representation quality. Finally, integrating explainable AI techniques could provide human-interpretable explanations for the selected feature subsets, enhancing trust and adoption in operational cybersecurity settings.

## 8. Conclusion

This paper presented IT-VAE-FS, a novel information-theoretic feature selection framework for network intrusion detection that leverages the latent representations learned by a $\beta$-Variational Autoencoder as a principled reference frame for evaluating feature importance. The framework introduces a composite information score integrating mutual information, KL divergence contribution, and reconstruction sensitivity, combined with a redundancy-aware greedy selection procedure inspired by the mRMR criterion. Extensive experiments on CICIDS-2017, UNSW-NB15, and NSL-KDD demonstrate that IT-VAE-FS consistently outperforms nine baselines. Using only 20 features, the framework achieves F1-scores within 0.06–0.16 points of the full-feature baseline while reducing dimensionality by 51–74% and yielding a 3.3$\times$ inference speedup. Statistical significance is confirmed through Friedman and Wilcoxon tests with Holm-Bonferroni correction. The ablation study validates the complementarity of all three scoring components and the importance of redundancy control. The key theoretical contribution is the demonstration that a deep generative model's latent space serves as an effective information-theoretic bridge for feature selection, with formal grounding in the information bottleneck principle. Promising directions include: (1) extending IT-VAE-FS to federated learning settings for privacy-preserving distributed IDS; (2) incorporating conditional VAE architectures for hybrid generative-discriminative scoring; (3) developing online

variants for adaptive feature selection under concept drift; (4) applying the methodology to malware classification, phishing detection, and encrypted traffic analysis; (5) exploring alternative generative models such as Wasserstein Autoencoders or normalizing flows; and (6) integrating explainable AI techniques for human-interpretable feature selection explanations.

## Acknowledgement

The authors are grateful to the Deanship of Research at Jadara University for providing financial support for this publication.

REFERENCES

1. P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, *Anomaly-based network intrusion detection: Techniques, systems and challenges*, Computers & Security, vol. 28, no. 1–2, pp. 18–28, 2009.
2. A. L. Buczak and E. Guven, *A survey of data mining and machine learning methods for cyber security intrusion detection*, IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.
3. R. Bellman, Dynamic Programming. Princeton, NJ: Princeton University Press, 1957.
4. I. H. Sarker, *Machine learning: Algorithms, real-world applications and research directions*, SN Computer Science, vol. 2, no. 3, pp. 1–21, 2021.
5. I. S. Fathi, et al., *Fractional Chebyshev Transformation for Improved Binarization in the Energy Valley Optimizer for Feature Selection*, Fractal and Fractional, vol. 9, no. 8, p. 521, 2025.
6. M. A. Hall, *Correlation-based feature selection for machine learning*, Ph.D. dissertation, University of Waikato, 1999.
7. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene selection for cancer classification using support vector machines*, Machine Learning, vol. 46, no. 1–3, pp. 389–422, 2002.
8. R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B, vol. 58, no. 1, pp. 267–288, 1996.
9. I. S. Fathi, M. A. Ahmed, and M. A. Makhlouf, *An efficient computation of discrete orthogonal moments for bio-signals reconstruction*, EURASIP Journal on Advances in Signal Processing, vol. 2022, no. 1, p. 104, 2022.
10. C. Doersch, *Tutorial on variational autoencoders*, arXiv preprint arXiv:1606.05908, 2016.
11. J. An and S. Cho, *Variational autoencoder based anomaly detection using reconstruction probability*, Special Lecture on IE, vol. 2, no. 1, pp. 1–18, 2015.
12. T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
13. A. Kraskov, H. Stögbauer, and P. Grassberger, *Estimating mutual information*, Physical Review E, vol. 69, no. 6, p. 066138, 2004.
14. E. A. Aldakheel, et al., *Efficient analysis of large-size bio-signals based on orthogonal generalized Laguerre moments of fractional orders and Schwarz–Rutishauser algorithm*, Fractal and Fractional, vol. 7, no. 11, p. 826, 2023.
15. I. S. Fathi, et al., *Protecting IoT networks through AI-based solutions and fractional Tchebichef moments*, Fractal and Fractional, vol. 9, no. 7, p. 427, 2025.
16. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.
17. I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, *$\beta$-VAE: Learning basic visual concepts with a constrained variational framework*, in Proc. 5th Int. Conf. Learning Representations (ICLR), 2017.
18. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, *Toward generating a new intrusion detection dataset and intrusion traffic characterization*, in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108–116.
19. N. Moustafa and J. Slay, *UNSW-NB15: A comprehensive data set for network intrusion detection systems*, in Proc. 2015 Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1–6.
20. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, *A detailed analysis of the KDD CUP 99 data set*, in Proc. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009, pp. 1–6.
21. I. S. Fathi, M. A. Ahmed, and M. A. Makhlouf, *An efficient compression technique for Foetal phonocardiogram signals in remote healthcare monitoring systems*, Multimedia Tools and Applications, vol. 82, no. 13, pp. 19993–20014, 2023.
22. C. Khammassi and S. Krichen, *A GA-LR wrapper approach for feature selection in network intrusion detection*, Computers & Security, vol. 70, pp. 255–277, 2017.
23. N. Moustafa and J. Slay, *The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set*, Information Security Journal: A Global Perspective, vol. 25, no. 1–3, pp. 18–31, 2016.
24. G. Stein, B. Chen, A. S. Wu, and K. A. Hua, *Decision tree classifier for network intrusion detection with GA-based feature selection*, in Proc. 43rd ACM Southeast Conference, 2005, pp. 136–141.
25. S. M. Kasongo and Y. Sun, *A deep learning method with wrapper based feature extraction for wireless intrusion detection system*, Computers & Security, vol. 92, p. 101752, 2020.
26. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, *Deep learning approach for intelligent intrusion detection system*, IEEE Access, vol. 7, pp. 41525–41550, 2019.
27. I. S. Fathi and M. Tawfik, *Enhancing IoT systems with bio-inspired intelligence in fog computing environments*, Statistics, Optimization & Information Computing, vol. 13, no. 5, pp. 1916–1932, 2025.

28. M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, *Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in IoT*, Sensors, vol. 17, no. 9, p. 1967, 2017.
29. H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, J. Chen, Z. Wang, and H. Qiao, *Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications*, in Proc. 2018 World Wide Web Conference (WWW), 2018, pp. 187–196.
30. M. Tawfik, et al., *FedMedSecure: federated few-shot learning with cross-attention mechanisms and explainable AI for collaborative healthcare cybersecurity*, Scientific Reports, vol. 15, no. 1, p. 40050, 2025.
31. S. Zavrak and M. İşcan, *Anomaly-based intrusion detection from network flow features using variational autoencoder*, IEEE Access, vol. 8, pp. 108346–108358, 2020.
32. C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
33. T. Chen and C. Guestrin, *XGBoost: A scalable tree boosting system*, in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
34. M. Tawfik, et al., *Explainable few-shot learning with modern BERT for detecting emerging phishing attacks using XF PhishBERT*, Scientific Reports, vol. 15, no. 1, p. 42821, 2025.
35. M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, Journal of the American Statistical Association, vol. 32, no. 200, pp. 675–701, 1937.
36. D. Gunning and D. Ahn, *DARPA's explainable artificial intelligence (XAI) program*, AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.
37. M. Tawfik, et al., *Internet of things-based middleware against cyber-attacks on smart homes using software-defined networking and deep learning*, in 2021 2nd Int. Conf. Computational Methods in Science & Technology (ICCMST), IEEE, 2021.
38. F. M. Talaat, M. Tawfik, and W. M. Shaban, *GreenAid: a confidence-weighted ensemble deep learning system for real-time plant disease detection and management*, Scientific Reports, vol. 16, no. 1, pp. 19193–19193, 2026.
39. M. Tawfik, A. H. Abdelhaliem, and I. Fathi, *Quantum-Resistant Privacy-Preserving IoT Authentication via Zero-Knowledge Proofs and Blockchain Integration*, Statistics, Optimization & Information Computing, vol. 14, no. 3, pp. 1374–1402, 2025.
40. A. M. Al-madni, et al., *An Optimized Blockchain Model for Secure and Efficient Data Management in Internet of Things*, in 2024 IEEE Int. Conf. Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), IEEE, 2024.
41. G. Hassan, et al., *Efficient compression of fetal phonocardiography bio-medical signals for Internet of Healthcare Things*, IEEE Access, vol. 11, pp. 122991–123003, 2023.
42. C. E. Shannon, *A mathematical theory of communication*, The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.
43. R. Battiti, *Using mutual information for selecting features in supervised neural net learning*, IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537–550, 1994.
44. C. Ding and H. Peng, *Minimum redundancy feature selection from microarray gene expression data*, Journal of Bioinformatics and Computational Biology, vol. 3, no. 2, pp. 185–205, 2005.
45. G. Brown, A. Pocock, M. J. Zhao, and M. Luján, *Conditional likelihood maximisation: A unifying framework for information theoretic feature selection*, Journal of Machine Learning Research, vol. 13, pp. 27–66, 2012.
46. J. R. Vergara and P. A. Estévez, *A review of feature selection methods based on mutual information*, Neural Computing and Applications, vol. 24, no. 1, pp. 175–186, 2014.
47. B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, *On variational bounds of mutual information*, in Proc. 36th Int. Conf. Machine Learning (ICML), 2019, pp. 5171–5180.
48. D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, *Variational inference: A review for statisticians*, Journal of the American Statistical Association, vol. 112, no. 518, pp. 859–877, 2017.
49. A. H. Abdelhaliem, I. S. Fathi, and M. Tawfik, *Fast and Efficient Feature Selection in AI Application Based on Enhanced Binary Secretary Bird Optimization Algorithm*, Statistics, Optimization & Information Computing, vol. 14, no. 5, pp. 2643–2662, 2025.
50. J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen, *Nonparametric entropy estimation: An overview*, Int. Journal of Mathematical and Statistical Sciences, vol. 6, no. 1, pp. 17–39, 1997.
51. A. van den Oord, Y. Li, and O. Vinyals, *Representation learning with contrastive predictive coding*, arXiv preprint arXiv:1807.03748, 2018.
52. M. D. Hoffman and M. J. Johnson, *ELBO surgery: Yet another way to carve up the variational evidence lower bound*, in Workshop on Advances in Approximate Bayesian Inference, NeurIPS, 2016.
53. N. Tishby, F. C. Pereira, and W. Bialek, *The information bottleneck method*, in Proc. 37th Allerton Conference on Communication, Control, and Computing, 1999, pp. 368–377.
54. A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, *Deep variational information bottleneck*, in Proc. 5th Int. Conf. Learning Representations (ICLR), 2017.
55. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic minority over-sampling technique*, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
56. S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, in Proc. 32nd Int. Conf. Machine Learning (ICML), 2015, pp. 448–456.
57. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.
58. D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in Proc. 3rd Int. Conf. Learning Representations (ICLR), 2015.
59. C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, *Understanding disentangling in $\beta$-VAE*, arXiv preprint arXiv:1804.03599, 2018.
60. Y. Burda, R. Grosse, and R. Salakhutdinov, *Importance weighted autoencoders*, in Proc. 4th Int. Conf. Learning Representations (ICLR), 2016.
61. R. A. Obeidat, I. S. Fathi, and M. Tawfik, *EKT-XAI: Efficient Kolmogorov-Arnold Network Transformer for Lightweight Smishing Detection with Explainable AI*, Statistics, Optimization & Information Computing, vol. 15, no. 6, pp. 5111–5134, 2026.
62. M. L. McHugh, *The chi-square test of independence*, Biochemia Medica, vol. 23, no. 2, pp. 143–149, 2013.
63. D. C. Montgomery, Design and Analysis of Experiments, 8th ed. New York: Wiley, 2013.

64. D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley, 1989.
65. M. Tawfik, A. H. Abdelhaliem, and I. S. Fathi, *Transforming IoT Security through Large Language Models: A Comprehensive Systematic Review and Future Directions*, Statistics, Optimization & Information Computing, vol. 14, no. 2, pp. 1018–1044, 2025.