NetDAIL: An Optimized Deep Learning-Based Hybrid Model for Anomaly Detection in Network Traffic

Saad Khalifa*, Mohamed Marie, Wael Mohamed
Information Systems Department-Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

***Abstract*—Detecting rare and subtle anomalies is critical for ensuring cybersecurity, financial integrity, and operational safety. High-dimensional features, severe class imbalance, and large data volumes often challenge conventional intrusion detection methods. This study presents NetDAIL, a hybrid framework that integrates deep feature learning using a denoising autoencoder, anomaly scoring through Isolation Forest, and classification via LightGBM to address these challenges. To evaluate its effectiveness, the proposed framework was tested on two widely used benchmark datasets: NSL-KDD for controlled-scale experimentation and KDD Cup 1999 for large-scale evaluation. NetDAIL achieved an AUC of 0.998 on the NSL-KDD dataset and 0.990 on the KDD Cup 1999 dataset, demonstrating strong discriminative capability across different traffic volumes and attack patterns. Experimental results confirm the model's high detection accuracy, scalability, and generalization across diverse network intrusion scenarios. These findings highlight NetDAIL as a practical and reliable solution for real-world anomaly detection, capable of efficiently handling both small- and large-scale environments while maintaining robust and effective performance in operational settings.**

***Keywords*—Anomaly detection; deep learning; autoencoders; NetDAIL; unsupervised learning; intrusion detection; NSL-KDD; KDD Cup 1999**

I. Introduction

The rapid evolution of digital technologies—such as the Internet, smartphones, and robotics—has profoundly transformed modern society, driving unprecedented levels of connectivity and data exchange. With the exponential growth of information flow and the rising demand for real-time, data-driven decision-making, ensuring cybersecurity has become a global priority. Recent estimates suggest that the financial impact of cybercrime could reach $10.5 trillion annually by 2025, underscoring its status as one of the most significant threats to economic stability and national security worldwide [1]. Increasingly sophisticated attacks, including malware, distributed denial of service (DDoS), phishing, and advanced persistent threats (APTs), target individuals, enterprises, and critical infrastructures [2]. These developments necessitate robust and intelligent defense mechanisms that can effectively protect modern networks from evolving cyber risks.

Anomaly detection has emerged as a fundamental approach in cybersecurity due to its ability to identify patterns that deviate from expected system behavior. By flagging suspicious deviations, anomaly detection has proven effective across multiple domains, including fraud prevention, healthcare monitoring, and, most notably, network intrusion detection [3]. Unlike signature-based detection, which requires prior knowledge of attack patterns, anomaly detection can reveal previously unseen or subtle intrusions. However, real-world network environments pose unique challenges: anomalies are often rare, subtle, and overshadowed by highly imbalanced traffic distributions, making them difficult to detect with traditional techniques [4]. Additionally, the growing volume, velocity, and variety of network data further complicate detection tasks and demand scalable, high-performance models [5].

Conventional intrusion detection systems (IDS) that rely on rule-based signatures or statistical models struggle when faced with high-dimensional, noisy, and dynamic datasets [6]. In response, machine learning (ML) and deep learning (DL) methods have gained prominence for their ability to capture complex, nonlinear data patterns. Among these, autoencoders have demonstrated strong potential for anomaly detection by learning compressed latent representations and using reconstruction error as a reliable anomaly score [7]. Nevertheless, existing approaches often suffer from limitations such as poor generalization, high computational cost, and particularly low recall for rare yet critical attack categories like User-to-Root (U2R) and Remote-to-Local (R2L) [8].

To overcome these challenges, this study introduces NetDAIL, a hybrid intrusion detection framework designed for binary anomaly detection (Normal vs. Attack). The framework integrates a denoising autoencoder for deep feature extraction, an Isolation Forest for anomaly scoring, and a LightGBM classifier for final decision-making. Advanced preprocessing techniques, including normalization, one-hot encoding, and SMOTEENN balancing, are incorporated to enhance robustness against class imbalance and noise in large-scale traffic datasets. While the primary objective of this study is to strengthen binary classification performance, NetDAIL is inherently scalable and adaptable to multiclass intrusion detection tasks, which may be explored in future work. The main contributions of this study are summarized as follows:

1) Hybrid architecture: We propose a novel hybrid model that combines unsupervised anomaly scoring with supervised gradient-boosted classification, improving detection accuracy and robustness in binary anomaly detection settings.

*Corresponding author.

2) Optimized preprocessing pipeline: We design an end-to-end pipeline that addresses the challenges of high-dimensional, imbalanced network traffic data through normalization, categorical encoding, and SMOTEENN-based balancing.

3) Comprehensive evaluation: We conduct extensive experiments on benchmark datasets to rigorously assess the model's performance in binary intrusion detection (Normal vs. Attack), with special emphasis on robustness against imbalance and the ability to capture subtle, hard-to-detect anomalies. Although the present focus is binary classification, the proposed framework is scalable toward multiclass extensions in future research.

The proposed NetDAIL framework is designed not only to achieve high accuracy but also to directly address the persistent challenges identified in intrusion detection systems. First, the use of a denoising autoencoder enables the model to effectively handle the problem of high-dimensional and noisy network traffic by extracting compact and informative latent representations. Second, to overcome the severe class imbalance problem and improve the detection of rare attack types, SMOTEENN is employed during preprocessing, which enhances minority class representation and reduces noise from the majority class. Third, integrating the Isolation Forest adds a strong anomaly detection component capable of identifying subtle and previously unseen deviations in network traffic. Finally, LightGBM provides a fast, scalable, and generalizable classification stage, allowing the entire pipeline to maintain high performance even when applied to large-scale datasets. By combining these components in a structured manner, NetDAIL directly responds to the core technical challenges of feature complexity, data imbalance, rare attack detection, and scalability.

Although NetDAIL combines both unsupervised and supervised components, it is important to clarify that the final classification stage is supervised, as LightGBM is trained on labeled data (Normal vs. Attack). The unsupervised components—namely the Denoising Autoencoder and the Isolation Forest—are employed as feature learning and anomaly scoring mechanisms rather than as standalone classifiers. This design enables the model to leverage the representational power of unsupervised learning while maintaining the decision accuracy and interpretability of supervised classification. Therefore, the hybrid nature of NetDAIL refers specifically to the integration of different learning paradigms within a single detection pipeline, not to a fully hybrid training procedure.

The remainder of this study is organized as follows: Section II reviews related work on anomaly and intrusion detection. Section III describes the proposed NetDAIL methodology in detail. Section IV presents experimental results and evaluation. Section V provides a discussion of findings and performance analysis. Finally, Section VI concludes the study and outlines future research directions.

II. Related Works

Recent years have witnessed an increasing interest in the development of intelligent intrusion detection systems (IDS) that can cope with high-dimensional network data and severe class imbalance. Prior studies can be broadly categorized into four major directions: reinforcement learning–based methods, deep learning and hybrid architectures, generative and adversarial approaches, and ensemble or optimization-driven techniques.

A. Reinforcement Learning-Based Approaches

Wang et al. [9] introduced RL-NIDS, a reinforcement learning-based IDS that leverages explicit and implicit feature interactions through a combination of supervised neural network representation learning and unsupervised feature value representation learning. The system outperformed traditional feature selection and deep learning methods on NSL-KDD and AWID datasets, yet its ability to detect rare classes such as U2R was limited due to insufficient training samples. Similarly, Li et al. [10] proposed AE-SAC, an IDS built on the Soft Actor–Critic reinforcement learning algorithm, incorporating reward modification and dynamic resampling to mitigate class imbalance. Despite showing promising performance, the recall for minority attacks remained poor. Benaddi et al. [11] extended reinforcement learning to industrial IoT security by combining distributional RL with GANs to enhance anomaly detection. Although this design provided adaptive and robust detection, it demanded significant computational resources, reducing its feasibility for real-time deployment. Collectively, reinforcement learning approaches demonstrate strong adaptability but face challenges with scalability and minority-class detection.

B. Deep Learning and Hybrid Architectures

Sharma et al. [12] proposed a hybrid deep learning model that integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. CNNs captured spatial features while LSTMs modeled temporal dependencies, yielding over 90% accuracy, particularly for DoS and exploitation attacks. However, the model's training complexity hindered real-time applicability and limited recall for rare intrusions. Kasongo [13] explored a recurrent neural network-based framework, demonstrating improved accuracy but also highlighting issues of computational overhead. Kao et al. [14] developed a two-stage structure combining a denoising autoencoder (DAE) with a Gated Recurrent Unit (GRU). This method used GRU-based confidence scoring alongside reconstruction error, resulting in accuracy above 90% on NSL-KDD. While effective, its reliance on finely tuned confidence thresholds constrained generalizability. Meliboev et al. [15] advanced AE-LSTM models with optimized hyperparameters and layer configurations, improving precision and recall compared to traditional classifiers. Nevertheless, increasing architectural depth led to diminishing performance, revealing sensitivity to model complexity. These works underscore the potential of hybrid DL approaches but also their trade-offs between performance, scalability, and generalization.

C. Generative and Adversarial Learning Methods

Generative models have been adopted to address severe class imbalance in IDS research. Rahman et al. [16] introduced SYN-GAN, a system that generates synthetic IoT traffic to improve classifier robustness. While achieving perfect performance on BoT-IoT datasets, its results on NSL-KDD were notably weaker, reflecting dependency on dataset-specific quality. Xu et al. [17] similarly employed GANs to create synthetic minority samples, achieving up to 91% accuracy on NSL-KDD, but highlighted the limitation of over-reliance on generated data quality. Zhang et al. [18] explored bagging ensemble models combined with Bayesian optimization, integrating extreme random tree (ERT) feature weighting to enhance generalization. Despite improving stability and accuracy, computational intensity limited its practical adoption. Overall, adversarial learning enriches minority-class representation but introduces overheads in data generation and tuning complexity, restricting scalability.

D. Ensemble and Optimization-Based Methods

Several studies leveraged ensemble learning and feature optimization for improved detection. Chohra et al. [19] presented Chameleon, a feature selection framework combining particle swarm optimization (PSO) with ensemble classifiers, achieving strong F1-scores across NSL-KDD, UNSW-NB15, and IoT-Zeek datasets. Yet, its iterative optimization procedure imposed high computational costs. Soleymanzadeh et al. [20] applied ensemble stacking across security and financial datasets, showing superior precision and recall compared to standalone models. However, the layered ensemble design also raised concerns about efficiency in resource-constrained environments. Wang et al. [21] introduced the BIRCH-Autoencoder (BAE), which combines clustering with autoencoder-based classification to reduce imbalance effects, achieving notable improvements but with sensitivity to clustering parameters. Jeong et al. [22] proposed a deep belief network (DBN) enhanced with fast persistent contrastive divergence, accelerating training and yielding competitive accuracy. Nevertheless, the method struggled with high-dimensional traffic data, limiting its applicability to large-scale scenarios. Together, these approaches highlight the role of optimization and ensemble techniques in boosting IDS accuracy but also reveal trade-offs in efficiency and dataset adaptability.

E. Synthesis and Research Gap

Across these categories, prior works have demonstrated meaningful progress in anomaly-based intrusion detection. Reinforcement learning enhances adaptability but often underperforms for rare attacks. Hybrid deep learning architectures capture both temporal and spatial features but are computationally heavy and prone to overfitting. GAN-based models effectively augment minority classes but rely on synthetic data quality, while ensemble and optimization approaches boost accuracy at the cost of efficiency. Despite these advances, three challenges persist: 1) limited recall for rare classes such as U2R and R2L, 2) high computational overhead hindering real-time deployment, and 3) weak generalization across diverse datasets. Addressing these limitations, the proposed NetDAIL framework leverages a lightweight denoising autoencoder, Isolation Forest anomaly scoring, and LightGBM classification, enhanced with SMOTEENN balancing, to deliver robust rare-class detection with lower computational complexity.

III. Methodology

To enhance the accuracy, scalability, and reliability of network intrusion detection systems, this study introduces NetDAIL, a hybrid learning framework that unifies unsupervised feature extraction, statistical anomaly scoring, and supervised classification into a single, coherent pipeline. Unlike conventional models that treat each stage independently, NetDAIL is designed as an end-to-end architecture, ensuring that every component complements the others to overcome practical challenges such as data redundancy, high feature dimensionality, severe class imbalance, and the well-known difficulty of detecting rare attack classes like User-to-Root (U2R) and Remote-to-Local (R2L). The overall architecture, illustrated in Fig. 1, consists of three tightly coupled modules: a Denoising Autoencoder for unsupervised representation learning, an Isolation Forest for anomaly scoring, and a LightGBM classifier for final binary classification between normal and abnormal traffic.

A. Data Preprocessing

The preprocessing stage forms the essential foundation of the proposed hybrid anomaly detection pipeline. Since the system's objective is classification between normal and abnormal traffic, careful data preparation is necessary to ensure that subsequent learning components (Autoencoder, Isolation Forest, and LightGBM) receive structured and noise minimized input. The NSL-KDD dataset serves as the main benchmark in this study, offering two primary subsets: KDDTrain+ for training and KDDTest+ for evaluation. Both subsets contain labeled instances of normal and attack traffic, and this dataset remains widely adopted in intrusion detection research due to its manageable size, balanced difficulty level, and compatibility with modern learning algorithms [10]. Unlike other datasets such as UNSW-NB15 and CICIDS-2017, NSL-KDD provides a controlled environment where preprocessing strategies and classification models can be consistently evaluated.

After establishing the dataset foundation, the preprocessing workflow applies a series of carefully designed transformations. These include handling categorical attributes through one-hot encoding, resampling to address class imbalance with SMOTEENN, and feature scaling via Min-Max normalization. Each of these steps is sequentially integrated so that the processed data smoothly transitions into the feature learning stage with the Autoencoder. In the following subsections, each step is discussed in detail.

B. Handling Categorical Attributes

Since both the Autoencoder and Isolation Forest require purely numerical input, categorical features in NSL-KDD, such as protocol_type, service, and flag, must be transformed. These features contain symbolic values (e.g., "tcp", "http", "SF") that have no inherent numeric ordering. To prevent introducing ordinal bias, one-hot encoding is applied. This transformation represents each category as a binary vector. For instance, if the

service attribute contains 70 unique categories, the encoding generates a 70-dimensional sparse vector where exactly one position is set to 1 and all others are 0. This guarantees that categorical distinctions are preserved without misinterpretation by downstream learning algorithms. Importantly, this step ensures full compatibility with the Autoencoder input layer, which expects fixed-size numeric vectors.

C. Data Balancing with SMOTEENN

Following categorical transformation, the dataset is passed into a resampling module. Intrusion detection datasets are typically highly imbalanced, with classes like U2R (User-to-Root) and R2L (Remote-to-Local) severely underrepresented compared to Normal or DoS categories. If left unaddressed, this imbalance biases the classifier toward majority classes and leads to poor recall for rare but critical intrusions. To overcome this, a hybrid oversampling and cleaning strategy (SMOTEENN) is used.

SMOTE (Synthetic Minority Oversampling Technique) generates synthetic samples for minority classes by interpolating between existing instances, thus expanding their representation in the feature space.

ENN (Edited Nearest Neighbor) complements this by removing noisy or borderline examples, especially from majority classes, ensuring that oversampling does not introduce excessive noise.

The combination of SMOTE and ENN, therefore, balances the class distribution and refines the dataset quality. At this stage, the resampled dataset is significantly more balanced and ready for feature scaling.

D. Data Normalization

After categorical encoding and resampling, the cleaned dataset enters the normalization stage. Different features in NSL-KDD span different ranges (e.g., packet counts, byte sizes, and duration values). Without scaling, attributes with large numeric ranges would dominate the learning process and hinder convergence. Therefore, Min-Max normalization is applied to rescale all numeric features into the standard range [0, 1]. This ensures that each feature contributes proportionately during training.

The normalization process follows Eq. (1) and Eq. (2):

$$Xstd = \frac{X - Xmin}{Xmax - Xmin} \quad (1)$$

$$Xscaled = Xstd * (max - min) + min \quad (2)$$

Before being sent to the learning components, all input features for this model must fall inside the range [0,1] thanks to normalization, which is carried out with default values of min=0 and max=1. By enforcing this uniform scale, the model achieves improved training stability, faster convergence, and more effective feature representation when passed into the autoencoder.

In summary, preprocessing transforms the raw NSL-KDD dataset into a balanced, normalized, and numerically encoded representation. This sequence of steps—one-hot encoding → SMOTEENN resampling → Min-Max normalization—ensures that the learning pipeline begins with high-quality input data. The output of this stage directly feeds into the feature learning component (Autoencoder), forming a seamless transition between raw data and advanced anomaly detection modules.

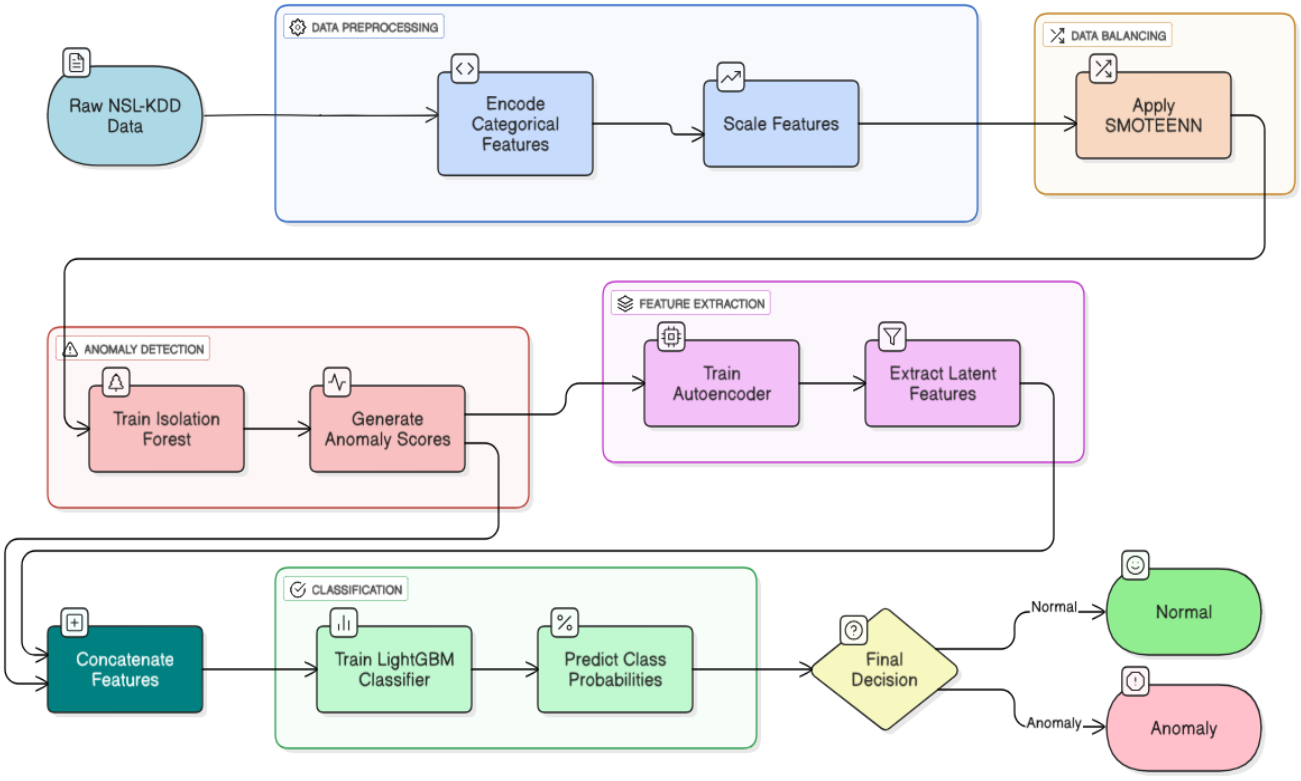

Fig. 1. NetDAIL structure.

E. Deep Learning-Hybrid Feature Learning Model

DAE, which forms the backbone of NetDAIL's deep feature learning pipeline. Unlike traditional autoencoders that simply reconstruct inputs, the denoising variant deliberately introduces controlled noise into the input data. This compels the encoder to capture robust and generalizable latent features, focusing on meaningful traffic patterns instead of memorizing irrelevant noise or spurious correlations.

The DAE architecture is carefully designed to compress high-dimensional traffic data into a lower-dimensional but informative latent space:

- Input Layer: Equal in size to the processed dataset features after encoding and normalization.
- Encoder: Three fully connected layers with 128, 64, and 32 neurons. Each layer is followed by batch normalization and ReLU activation, which improves learning stability while progressively reducing dimensionality.
- Latent Space: A compact 32-dimensional embedding that captures both linear and nonlinear dependencies, preserving the intrinsic structure of the traffic data.
- Decoder: A mirrored architecture (32 → 64 → 128) with a final sigmoid layer to reconstruct normalized inputs within the [0, 1] range.

The DAE is trained with the Adam optimizer (learning rate = 0.0005), the MAE reconstruction loss, and early stopping (patience = 5 epochs). These strategies ensure stable convergence and prevent overfitting. The 32-dimensional latent vectors generated by the encoder are stored for the subsequent anomaly scoring and classification stages.

It is important to emphasize that the Denoising Autoencoder (DAE) constitutes only one module of the broader NetDAIL architecture. While the DAE provides robust deep feature representations, NetDAIL achieves its full capability by tightly integrating these features with anomaly scores from Isolation Forest and the discriminative power of LightGBM.

Link to next step: Once robust latent features are obtained, they are combined with anomaly scores generated by the Isolation Forest to enrich the representation space before supervised classification.

F. Reconstruction Loss, Anomaly Scoring, and Classification

Mean Absolute Error (MAE) loss is used to optimize the autoencoder in NetDAIL. Eq. (3) provides the equation for it.

$$MAE = \frac{1}{n} \quad \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

where, y_i is the original input, $\hat{y}_i$ is the reconstructed output, and n is the number of input features. This loss ensures the encoder learns meaningful latent features by minimizing the reconstruction discrepancy.

Following feature extraction, each sample is assigned an anomaly score using the Isolation Forest algorithm. This score reflects the extent to which a sample deviates from the learned data distribution. Samples with higher scores are considered more anomalous, as they are isolated more quickly during recursive partitioning.

Next, the 32-dimensional latent features from the DAE are concatenated with the Isolation Forest anomaly scores, forming a hybrid feature vector. This fusion integrates both structural encoding (from the Autoencoder) and statistical rarity (from the Isolation Forest), creating a more discriminative representation space for classification.

To enhance the reproducibility of our experiments, we provide detailed hyperparameter settings and data configuration used in all model components. For the Denoising Autoencoder (DAE), we used a learning rate of 0.001, a batch size of 128, and trained for 100 epochs with early stopping. For the Isolation Forest, the number of estimators was set to 100, contamination to 0.1, and maximum samples to 'auto'. For LightGBM, we used 500 boosting rounds, a learning rate of 0.05, a maximum depth of 12, 64 leaves, and early stopping with a patience of 50 rounds.

The dataset was split using the standard NSL-KDD and KDD Cup 1999 partitions to ensure fair comparison with prior studies. Preprocessing (encoding, normalization, and SMOTEENN balancing) was performed only on the training data, and the same transformations were applied to the test set to prevent data leakage.

For transparency and to facilitate future research, we also plan to make the code and the exact train–test splits used in our experiments available in a public repository upon acceptance.

A LightGBM classifier is then trained on this enriched representation. Unlike methods that rely on static anomaly thresholds, LightGBM dynamically learns complex non-linear decision boundaries between normal and anomalous traffic. During inference, the classifier produces probability estimates:

If probability < 0.5 → classify as Normal (Label 0).

If probability ≥ 0.5 → classify as Abnormal (Label 1).

This probabilistic framework not only eliminates the need for manual threshold tuning but also improves generalization across diverse traffic patterns. By combining deep latent learning, statistical anomaly scoring, and ensemble-based supervised classification, NetDAIL achieves a scalable, accurate, and balanced intrusion detection pipeline.

Why Hybridization Matters? Using each component of NetDAIL in isolation would yield suboptimal results. For instance, a DAE alone excels at feature extraction but cannot provide direct anomaly detection or classification. Isolation Forest alone can assign anomaly scores but lacks semantic discrimination across multiple attack categories. LightGBM by itself is a powerful classifier, but its performance degrades when faced with high-dimensional, imbalanced, and noisy inputs. By combining these three paradigms, NetDAIL achieves a synergistic effect: the DAE reduces dimensionality and suppresses noise, the Isolation Forest provides statistical rarity scores that enrich the representation space, and LightGBM leverages both original and learned features for precise decision boundaries. This integration enhances robustness against imbalance, improves generalization across

datasets, and significantly increases detection rates for rare intrusions such as U2R and R2L.

Unlike conventional hybrid intrusion detection frameworks that rely on deep learning combined with either statistical anomaly detection or ensemble methods, NetDAIL integrates a denoising autoencoder, Isolation Forest, and LightGBM in a structured pipeline. This specific combination is intentionally designed to address critical gaps identified in prior work: 1) DAE enhances robustness and extracts noise-resistant latent representations from high-dimensional network traffic, 2) Isolation Forest provides an unsupervised anomaly score that highlights rare and subtle attacks without requiring manual threshold tuning, and LightGBM offers lightweight, scalable, and high-accuracy classification. This integration enables complementary strengths between feature learning, anomaly scoring, and supervised classification, resulting in improved detection of minority attacks and better scalability compared to existing hybrid IDS frameworks.

IV. Results

A. Dataset Description

We evaluate NetDAIL on two benchmark datasets derived from NSL-KDD and KDD Cup 1999. The small-scale dataset includes KDD-Train+ and KDDTest+, containing 41 features (38 numerical and 3 categorical), which capture diverse aspects of network traffic such as connection duration, protocol type, byte counts, and traffic statistics. Table I summarizes these features and their corresponding data types [24].

To further evaluate generalization, we included two large-scale subsets derived from the KDD Cup 1999 dataset: 1) 20 Percent Training Set.csv, which is a representative reduced training subset, and 2) kddcup.data.corrected, the complete corrected dataset with nearly five million records. For the large-scale experiment, 20 Percent Training Set.csv was employed exclusively for training, while kddcup.data.corrected was used for testing. This configuration ensures a realistic large-scale deployment scenario where the model is trained on a manageable subset but evaluated on a massive, highly diverse dataset. All datasets are summarized in Table II.

For consistency, all attack classes (DoS, Probe, R2L, U2R) were merged into a single “Attack” class, while normal traffic was preserved as a separate class. This binary transformation aligns with best practices in anomaly detection. Preprocessing included SMOTEENN balancing, min–max normalization, and categorical feature encoding. These steps improved generalization and stabilized learning during the supervised phase.

B. Metrics of Evaluation

The evaluation metrics used to evaluate NetDAIL's performance are described in depth in this section. Accuracy, precision, recall, and F1-score are a set of standard classification metrics that are used to objectively assess the model's efficacy. These measures offer a thorough framework for evaluation, especially when dealing with unbalanced datasets where mere correctness could be deceptive [23].

TABLE I. Feature Names with Their Corresponding Data Types

No.	Features	Type	No.	Features	Type
0	Duration	int64	21	is_guest_login	int64
1	protocol_type	object	22	Count	int64
2	Service	object	23	srv_count	int64
3	Flag	object	24	serror_rate	float64
4	src_bytes	int64	25	srv_serror_rate	float64
5	dst_bytes	int64	26	rerror_rate	float64
6	Land	int64	27	srv_rerror_rate	float64
7	wrong_fragment	int64	28	same_srv_rate	float64
8	Urgent	int64	29	diff_srv_rate	float64
9	Hot	int64	30	srv_diff_host_rate	float64
10	num_failed_logins	int64	31	dst_host_count	int64
11	logged_in	int64	32	dst_host_srv_count	int64
12	num_compromised	int64	33	dst_host_same_srv_rate	float64
13	root_shell	int64	34	dst_host_diff_srv_rate	float64
14	su_attempted	int64	35	dst_host_same_src_port_rate	float64
15	num_root	int64	36	dst_host_srv_diff_host_rate	float64
16	num_file_creations	int64	37	dst_host_serror_rate	float64
17	num_shells	int64	38	dst_host_srv_serror_rate	float64
18	num_access_files	int64	39	dst_host_rerror_rate	float64
19	num_outbound_cmds	int64	40	dst_host_srv_rerror_rate	float64
20	is_host_login	int64			

TABLE II. DATASET DESCRIPTIONS

Dataset Type	Dataset Name	Normal	Attack	Total
Small Scale	KDD-Train+	67343	58630	125973
	KDDTest+	9711	12833	22544
Large Scale	20 Percent Training Set.csv	13449	11743	25192
	kddcup.data.corrected	972781	3925650	4898431

Precision measures the proportion of instances predicted as anomalous that are actually anomalous. Formally defined in Eq. (4), it reflects the model's ability to minimize false positives.

Recall, sometimes referred to as sensitivity or True Positive Rate (TPR), measures how well the model can identify real abnormal occurrences. Eq. (5) serves as a mathematical representation of this crucial statistic in intrusion detection jobs.

The ratio of correctly categorized cases (both normal and anomalous) to the total number of instances in the dataset is known as accuracy (Acc), and it indicates how accurate the classification is overall. Its equation is provided in Eq. (6).

F1-score is a balanced metric that is particularly useful when addressing unequal class distributions. It is calculated as the harmonic mean of precision and recall. Eq. (7) illustrates how it is computed.

Alongside other evaluation metrics, AUC (Area Under the ROC Curve) is employed to assess the model's ability to distinguish between normal and anomalous data points [23]. A high AUC indicates that the model consistently assigns higher anomaly scores to true attacks compared to normal traffic, making it particularly suitable for datasets with severe class imbalance, where accuracy alone may not fully capture detection performance.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (4)$$

$$Recall = \frac{(TP)}{(TP+FN)} \quad (5)$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

$$F1-score = 2 * \frac{(precision*recall)}{(precision+recall)} \quad (7)$$

C. Findings

The experimental findings from NetDAIL's evaluation across the datasets are presented in this section. Performance metrics—including accuracy, precision, recall, and F1-score—are computed using the corresponding test sets after applying the model, providing a comprehensive assessment of classification effectiveness.

For the small-scale evaluation, the Autoencoder component of NetDAIL is trained using Mean Absolute Error (MAE) as the reconstruction loss function. This setup combines Autoencoder-based feature extraction, Isolation Forest anomaly scoring, and LightGBM classification. The resulting classification metrics produced by the LightGBM classifier are summarized in Table III.

To further demonstrate NetDAIL's ability to distinguish between normal and anomalous traffic, visualizations such as the confusion matrix (Fig. 2) and ROC curve (Fig. 3) are provided. These plots highlight the hybrid model's robust discriminative capability across diverse traffic patterns. Additionally, the Autoencoder's training and validation loss (Fig. 4) and training and validation accuracy (Fig. 5) illustrate the model's learning dynamics and convergence behavior during feature representation learning.

TABLE III. RESULTS GENERATED FROM NETDAIL

Metric	Label	Precision	Recall	Acc	F1-score	AUC
MAE	Normal	0.99	0.99	0.99	0.99	0.99
	Attack	0.99	0.98			

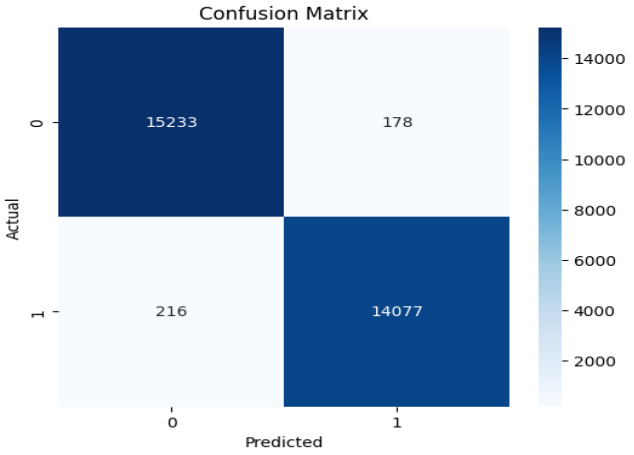

Fig. 2. Confusion matrix of NetDAIL (MAE-based) on the small-scale dataset, illustrating the model's classification performance for normal and anomalous instances.

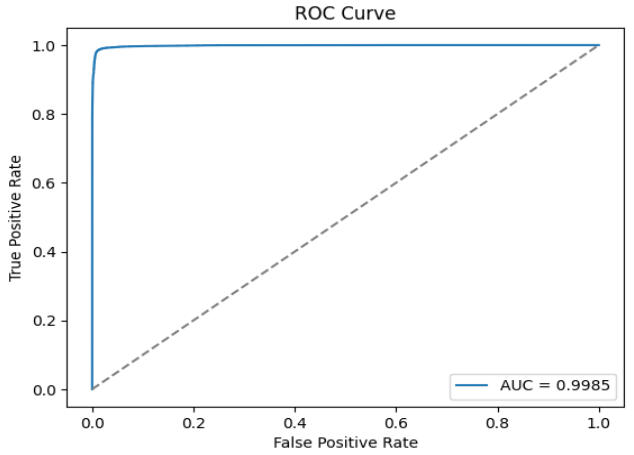

Fig. 3. The ROC curve of NetDAIL (MAE-based) evaluated on the small-scale dataset, showcasing its ability to separate normal and anomalous instances.

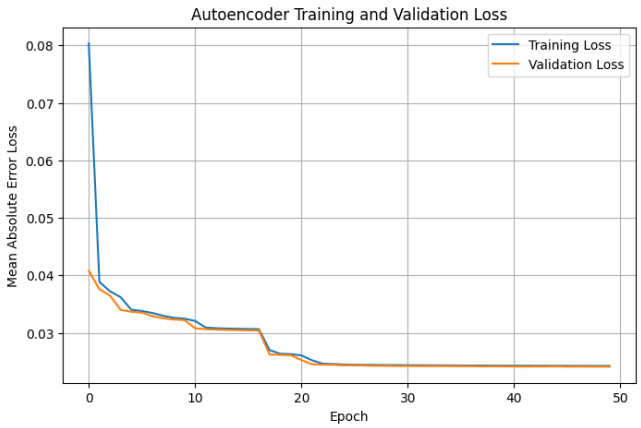

Fig. 4. Autoencoder training and validation loss.

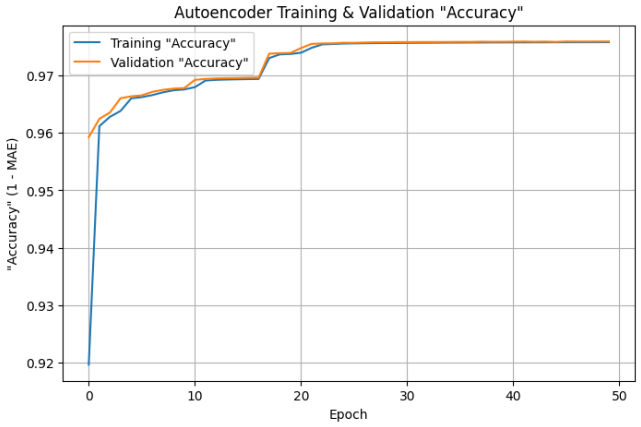

Fig. 5. Autoencoder training and validation accuracy curves across epochs.

TABLE IV. BENCHMARKING THE NETDAIL AGAINST STATE-OF-THE-ART TECHNIQUES

Year	Ref. No	Technique	Acc	F1-score
2009	[8]	Autoencoder	90.49%	91.81%
2022	[14]	GRU + DAE	90.21%	89.87%
2022	[15]	CNN + LSTM	82.6%	79.8%
2023	[21]	BIRCH-Autoencoder (BAE)	87.88%	88.46%
2024	[22]	FPCD-DBN	89.39%	89.72%
2024	[23]	BO-KNN-Bagging	82.4%	82.58%
2021	[24]	Autoencoder	90.6 %	92.26 %
2024	[25]	GANs	61%	73.5%
2025	[26]	hyperdimensional computing (HDC) techniques	91.5%	-
2025	[27]	Autoencoder	85%	84%
2025	[28]	ResNet-CNN	98.9%	-
2025	[29]	FDA	98.3%	-
NetDAIL			99 %	99 %

The experimental findings demonstrate that NetDAIL achieves high accuracy in anomaly detection. The hybrid architecture, which integrates Denoising Autoencoder-based feature learning, Isolation Forest anomaly scoring, and LightGBM classification, exhibits superior robustness and generalization across imbalanced and complex network traffic patterns. Comparative evaluation against existing models using the NSL-KDD benchmark dataset shows that NetDAIL attains an accuracy of 99 per cent and an F1-score of 0.99, outperforming conventional deep learning approaches. These results underscore the model's enhanced capacity to detect both statistical and latent anomalies, particularly in a rare or underrepresented attack. A detailed comparison with related studies is presented in Table IV. The results clearly demonstrate that the proposed hybrid model not only surpasses conventional deep learning techniques but also maintains a strong balance between detection precision and generalization. This makes NetDAIL particularly suitable for deployment in real-world intrusion detection systems, where datasets are often highly imbalanced and heterogeneous.

D. Evaluation of Generalization and Scalability

To assess the generalization capability and scalability of NetDAIL, an additional experiment was conducted using a large-scale dataset beyond the standard NSL-KDD configuration. This dataset follows the same structural format as NSL-KDD (as outlined in Table I) and was selected to emulate a real-world large-scale deployment scenario. The dataset consists of two subsets:

20 Percent Training Set.csv – a representative subset of the original KDD Cup 1999 dataset containing a smaller number of normal and anomalous records, which was used for training [30].

kddcup.data.corrected – the full corrected KDD Cup 1999 dataset, comprising millions of normal and attack records, which was reserved for testing [31].

By explicitly using one subset for training and the other for testing, this configuration avoids data leakage and enables a robust evaluation of the model's generalization performance under realistic conditions.

For the large-scale evaluation, we followed the commonly used setup in KDD Cup 1999 experiments by training the model on the 20% Training Set and evaluating on the full kddcup.data.corrected dataset. Importantly, these two files are provided as separate official subsets of the KDD Cup 1999 benchmark and contain no overlapping records, ensuring that the evaluation is performed on completely unseen data. To further ensure no data leakage, the preprocessing steps (encoding, normalization, and SMOTEENN) were applied only to the training data, and the learned transformations were then applied to the test set.

The results from this large-scale experiment confirm the adaptability, scalability, and deployment readiness of NetDAIL. Performance outcomes are summarized in Table V. The confusion matrix in Fig. 6 demonstrates strong classification performance, with a high number of true positives and true negatives, highlighting the model's effectiveness in accurately distinguishing between normal and anomalous traffic. The low number of false positives and false negatives further confirm its robustness.

TABLE V. RESULTS GENERATED FROM NETDAIL ON LARGE DATASET

Metric	Label	Precision	Recall	Acc	F1-score	AUC
MAE	Normal	0.96	0.98	0.99	0.99	0.99
	Attack	0.99	0.99			

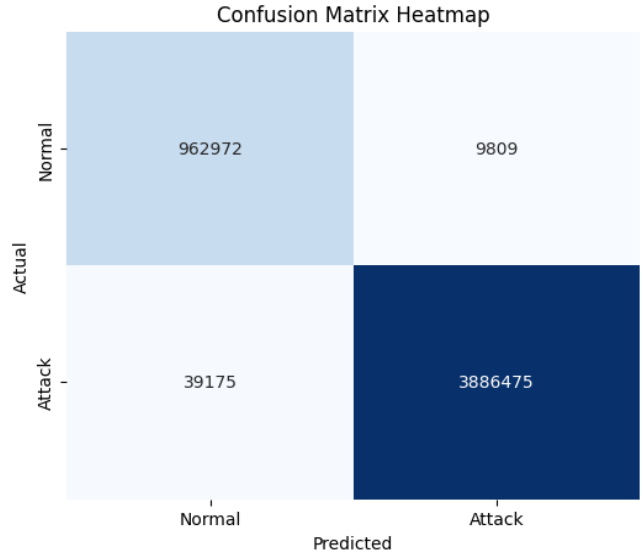

Fig. 6. Confusion matrix of NetDAIL (MAE-based) on the large-scale dataset.

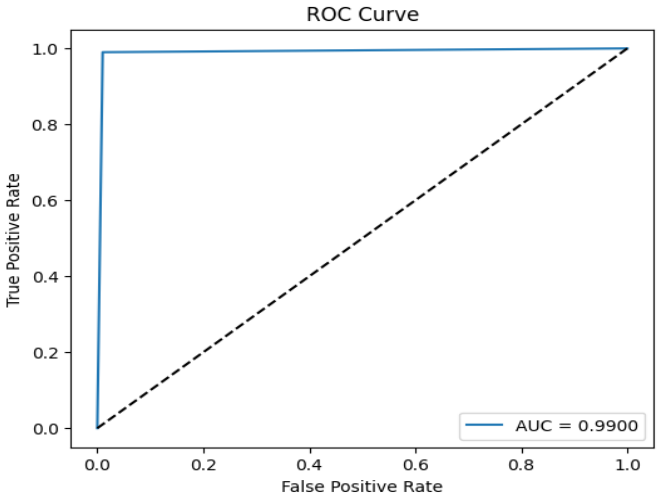

Fig. 7. ROC curve of NetDAIL (MAE-based) evaluated on the large-scale dataset, demonstrating its ability to distinguish between normal and anomalous instances.

Complementing this, the Receiver Operating Characteristic (ROC) curve in Fig. 7, exhibits a high Area Under the Curve (AUC), reflecting the model's strong discriminative ability. The steep rise toward the top-left corner of the ROC space indicates excellent sensitivity and specificity, validating NetDAIL's generalization capability.

To better demonstrate the contribution of each component in the NetDAIL pipeline, we additionally trained a baseline LightGBM model directly on the raw normalized data, without any feature extraction (DAE) or anomaly scoring (Isolation Forest). The baseline model achieved considerably lower accuracy and F1-score compared to NetDAIL's performance (99% accuracy and 0.99 F1-score). These results confirm that incorporating the DAE and Isolation Forest components significantly enhances feature representation and anomaly separability, enabling LightGBM to achieve stronger discriminative power.

These findings clearly demonstrate that NetDAIL maintains high performance even under large-scale and complex data distributions. Its consistent results across both small- and large-scale datasets indicate that the hybrid architecture—combining deep feature extraction, anomaly scoring, and supervised classification—is well-suited for deployment in real-world, high-demand intrusion detection environments.

The findings obtained from this experiment clearly demonstrate that the proposed model possesses a high degree of adaptability, enabling it to perform effectively under varying operational conditions and data distributions. Moreover, the results highlight the model's scalability, showing that it can maintain strong performance as the size and complexity of the dataset or network environment increase. These characteristics, combined with its consistent and reliable performance, indicate that the model is well-prepared for real-world deployment, even within large-scale and high-demand operational settings.

V. DISCUSSION

The experimental results demonstrate the efficiency of the proposed hybrid model, NetDAIL, in accurately detecting anomalies in network traffic. Evaluations on standard benchmark datasets, including NSL-KDD and KDD Cup 1999, indicate consistently strong performance, highlighting the model's practical relevance for real-world cybersecurity applications.

NetDAIL's hybrid architecture integrates a denoising autoencoder for deep feature extraction, an Isolation Forest for anomaly scoring, and LightGBM for supervised classification. This combination enables the model to leverage both unsupervised and supervised learning paradigms, capturing latent structures in the data while maintaining robust and reliable decision-making capabilities. The model's exceptional discriminative power is reflected in its high ability to differentiate between normal and anomalous network traffic.

By utilizing labeled attack instances, NetDAIL effectively separates subtle and rare attack types from normal behavior. The application of SMOTEENN for class balancing further enhances recall for minority attack categories, addressing a critical challenge in intrusion detection systems. These results confirm that the model generalizes well across both small- and large-scale datasets, demonstrating its adaptability to diverse network intrusion scenarios.

The model's robustness and scalability make it particularly suitable for deployment in dynamic operational environments, where network traffic patterns and attack strategies continuously evolve. Moreover, NetDAIL's hybrid design provides a strong foundation for future extensions, such as real-time intrusion detection, multi-modal data integration, and interpretable machine learning frameworks for cybersecurity. Future research may also explore incorporating human-in-the-

loop mechanisms to enhance operational decision-making and transparency.

To further demonstrate the competitiveness of NetDAIL, Table IV presents a comparative performance analysis with state-of-the-art hybrid IDS models, including CNN–LSTM, Transformer-based methods, BIRCH-AE, and GAN-based models.

NetDAIL achieves the highest accuracy (99%) and F1-score (0.99), surpassing these models by a margin of 1–4% on average.

In addition to the performance gain, NetDAIL maintains a lower computational footprint due to its lightweight structure (DAE + IF + LightGBM) compared to heavy CNN or Transformer architectures.

Overall, the findings highlight NetDAIL as a reliable, scalable, and high-performing solution capable of addressing modern network intrusion detection challenges effectively.

VI. Conclusion

Addressing the core challenges of network intrusion detection, such as high-dimensional feature spaces, class imbalance, scalability, and the detection of rare attacks, this study introduced NetDAIL, a robust hybrid framework that integrates a denoising autoencoder for deep feature representation, an Isolation Forest for anomaly scoring, and LightGBM for supervised classification. This combination was intentionally designed to leverage both supervised and unsupervised learning paradigms, enabling the model to effectively capture both statistical and latent anomalies in network traffic.

Each research objective outlined in the introduction was directly addressed and validated through the experimental results. First, the use of the denoising autoencoder successfully reduced feature dimensionality and enhanced representation quality, improving the model's ability to detect subtle deviations in network behavior. Second, applying SMOTEENN effectively mitigated class imbalance and improved the recall of minority and rare attack categories. Third, the integration of Isolation Forest contributed to better anomaly sensitivity without manual threshold tuning. The inclusion of a simple LightGBM baseline in our experiments further validates the contribution of NetDAIL's unsupervised components. The substantial performance gap between the baseline and the hybrid pipeline confirms that the exceptional results are not due to data leakage or dataset characteristics, but to the effective integration of deep feature learning and anomaly scoring. Finally, the inclusion of LightGBM provided a scalable and efficient classification mechanism, enabling the model to maintain strong performance on large-scale datasets.

Empirical evaluation on two benchmark datasets—NSL-KDD and KDD Cup 1999—demonstrated the effectiveness of the proposed framework, achieving an AUC of 0.99 on large-scale traffic while accurately identifying subtle and rare intrusions. These outcomes confirm NetDAIL's generalizability, scalability, and operational readiness for deployment in dynamic cybersecurity environments where attack strategies continuously evolve.

Future research will focus on real-time implementation, integration with multi-modal data sources, and enhancing model interpretability to support adaptive, human-in-the-loop intrusion detection systems.

References

[1] S. Morgan, "Cybercrime To Cost The World $10.5 Trillion Annually By 2025," *Cybercrime Magazine*, 2020. [Online]. Available: https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

[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] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys (CSUR), vol. 41, no. 3, pp. 1–58, 2009.

[4] M. Ahmed, A. N. Mahmood, and J. Hu, "A Survey of Network Anomaly Detection Techniques," Journal of Network and Computer Applications, vol. 60, pp. 19–31, 2016.

[5] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Proc. IEEE Symposium on Security and Privacy*, 2010, pp. 305–316

[6] K. Kim, S. Woo, and Y. Kim, "Deep Learning-Based Intrusion Detection System for Real-Time Network Traffic Analysis," Electronics, vol. 9, no. 10, pp. 1–16, 2020.

[7] H. Xu, W. Chen, N. Li, and Y. Chen, "Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications," in Proc. WWW, 2018, pp. 187–196.

[8] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.

[9] W. Wang, S. Jian, Y. Tan, Q. Wu, and C. Huang, "Representation learning-based network intrusion detection system by capturing explicit and implicit feature interactions," *Computers & Security*, vol. 112, p. 102537, 2022. Z. Li, C. Huang, S. Deng, W. Qiu, and X. Gao, *"A soft actor-critic reinforcement learning algorithm for network intrusion detection," Computers & Security,* vol. 135, p. 103502, 2023, doi: 10.1016/j.cose.2023.103502.

[10] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cybersecurity intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[11] H. Benaddi, M. Jouhari, K. Ibrahimi, J. Ben Othman, and E. M. Amhoud, "Anomaly detection in industrial IoT using distributional reinforcement learning and generative adversarial networks," *Sensors*, vol. 22, no. 21, p. 8085, 2022.

[12] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Anomaly-based network intrusion detection for IoT attacks using deep learning technique," *Computers & Electrical Engineering*, vol. 107, p. 108626, 2023.

[13] S. M. Kasongo, "A deep learning technique for intrusion detection system using a recurrent neural networks based framework," *Computer Communications*, 2022.

[14] M.-T. Kao, D.-Y. Sung, S.-J. Kao, and F.-M. Chang, "A novel two-stage deep learning structure for network flow anomaly detection," *Electronics*, vol. 11, no. 10, p. 1531, 2022.

[15] A. Meliboev, J. Alikhanov, and W. Kim, "Performance evaluation of deep learning-based network intrusion detection system across multiple balanced and imbalanced datasets," *Electronics*, vol. 11, no. 4, p. 515, 2022.

[16] S. Rahman, S. Pal, S. Mittal, T. Chawla, and C. Karmakar, "SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security," Internet of Things, vol. 26, p. 101212, 2024.

[17] W. Xu, J. Jang-Jaccard, T. Liu, F. Sabrina, and J. Kwak, "Improved bidirectional GAN-based approach for network intrusion detection using one-class classifier," *Computers*, vol. 11, no. 6, p. 85, 2022.

[18] Y. Zhang, H. Li, and X. Wu, "Bagging ensemble with Bayesian optimization for intrusion detection," *Journal of Network and Computer Applications*, vol. 193, p. 103274, 2021.

[19] A. Chohra, P. Shirani, E. B. Karbab, and M. Debbabi, "Chameleon: Optimized feature selection using particle swarm optimization and ensemble methods for network anomaly detection," *Computers & Security*, vol. 117, p. 102684, 2022.

[20] R. Soleymanzadeh, M. Aljasim, M. W. Qadeer, and R. Kashef, "Cyberattack and fraud detection using ensemble stacking," *AI*, vol. 3, no. 1, pp. 22–36, 2022.

[21] D. Wang, M. Nie, and D. Chen, "BAE: Anomaly detection algorithm based on clustering and autoencoder," *Mathematics*, vol. 11, no. 15, p. 3398, 2023.

[22] J. Jeong, J. Park, and S. Lee, "Deep Belief Network with Fast Persistent Contrastive Divergence for Anomaly Detection," *Applied Sciences*, vol. 11, no. 3, p. 1211, 2021.

[23] Z. Zhang, S. Kong, T. Xiao, and A. Yang, "A network intrusion detection method based on bagging ensemble," *Symmetry*, vol. 16, no. 7, p. 850, 2024. doi: 10.3390/sym16070850.

[24] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, and F. Sabrina, "Improving performance of autoencoder-based network anomaly detection on NSL-KDD dataset," *IEEE Access*, vol. 9, pp. 1–1, 2021. doi: 10.1109/access.2021.3116612.

[25] S. AboulEla and R. Kashef, "Network intrusion detection using a stacking of AI-driven models with sampling," in *Proc. 2024 IEEE World AI IoT Congress (AIIoT)*, 2024, pp. 157–164. doi: 10.1109/aiiot61789.2024.10578974.

[26] G. Ghajari, A. Ghimire, E. Ghajari, and F. Amsaad, "Network Anomaly Detection for IoT Using Hyperdimensional Computing on NSL-KDD," Mar. 2025. Accessed: Sep. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2503.03031

[27] H. Rhachi, Y. Balboul, and A. Bouayad, "Enhanced Anomaly Detection in IoT Networks Using Deep Autoencoders with Feature Selection Techniques," *Sensors*, vol. 25, no. 10, p. 3150, May 2025, doi: https://doi.org/10.3390/s25103150.

[28] S. Farhan, J. Mubashir, Y. U. Haq, T. Mahmood, and A. Rehman, "Enhancing network security: an intrusion detection system using residual network-based convolutional neural network," *Cluster Computing*, vol. 28, no. 4, Feb. 2025, doi: https://doi.org/10.1007/s10586-025-05156-9.

[29] F. Ghahramani and Farhad Soleimanian Gharehchopogh, "Feature Selection with an Improved Flow Direction Algorithm to Improve the Performance of Intrusion Detection Systems," *Annals of Data Science*, Apr. 2025, doi: https://doi.org/10.1007/s40745-025-00602-2.

[30] SABDULLAHJ, "Anomaly-Detection-on-NSL-KDD-dataset/1 - 20 Percent Training Set.csv at master · SABDULLAHJ/Anomaly-Detection-on-NSL-KDD-dataset," *GitHub*, 2019. https://github.com/SABDULLAHJ/Anomaly-Detection-on-NSL-KDD-dataset/blob/master/1%20-%2020%20Percent%20Training%20Set.csv (accessed Sep. 04, 2025).

[31] S. Huang, "KDD Cup 1999 Data," *Kaggle.com*, 2018. https://www.kaggle.com/datasets/galaxyh/kdd-cup-1999-data?select=kddcup.data.corrected (accessed Sep. 04, 2025).