https://doi.org/10.65025/ICAIC25074o

Explainable AI-Powered Anomaly Detection: A Computer Vision Approach to Strengthening Human-Centric Cybersecurity.

Victor Ayodeji Oluwasusi
Near East University, Cyprus
oluwasusiv@gmail.com

Abstract
As cyber threats increase in sophistication and stealth, traditional Intrusion Detection Systems (IDSs) typically experience difficulties identifying emerging types of attacks (e.g., zero-day attacks) because they rely on known signatures and lack interpretability. This paper presents the development of SENTRY-AI, an explainable, multi-modal anomaly detection framework that integrates deep learning and computer vision to improve cybersecurity defenses in cyberspace. SENTRY-AI architecture employs a Variational Autoencoder (VAE) to conduct unsupervised anomaly detection on tabular network features, in addition to a Convolutional Neural Network (CNN) to analyze time-series traffic data transformed into a Gramian Angular Field (GAF). An analysis is performed with a late-fusion approach utilizing outputs from the two models to evaluate the efficacy and robustness of the hybrid approach. Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to provide visual explanations which indicate the most significant areas of input used to make the predictions for the purpose of human-in-the-loop decision making. The experimental work is conducted across three intrusion detection datasets (NSL-KDD, CICIDS2017, and UNSW-NB15), demonstrating that SENTRY-AI outperformed traditional machine learning and modern deep learning-based IDS approaches. In fact, SENTRY-AI achieved an F1-score of 98.92% on NSL-KDD, 100.00% on UNSW-NB15, and an AUC-ROC score exceeding 99% on all datasets.

Keywords
Anomaly Detection, Computer Vision, Explainable AI, Intrusion Detection

1. Introduction

Background & Context

The modern cybersecurity environment is characterized by an unprecedented rise in the sophistication, number, and variety of cyber threats. Digital transformation is prevalent across all aspects of society and increasing reliance on systems and networks connected with each other, increases both vulnerabilities and attack surfaces (Almiani et al., 2020). An increase in cyber threats, including ransomware attacks, phishing attacks, denial-of-service (DoS) attacks, advanced persistent threats (APTs), and zero-day exploits, is seen across the range of cyber actors, from organized crime operations to state-sponsored threat actors (Moustafa et al., 2019). The increased frequency of attacks illustrates the limitations of a well-established, traditional signature- based intrusion detection system (IDS) that relies on reference known patterns of cyber threats, and as a result lacks the ability to recognize new and methodical attacks.

Methods of anomaly detection supported by artificial intelligence (AI) and machine learning (ML), have gained attention to deal with these obstacles. Anomaly detection systems do not rely on signature-based approaches and detect abnormal behavior from typical network behavior. This style of threat detection is more proactive and flexible. Machine learning fundamental algorithms and methods, in particular, deep learning methods such as Variational Autoencoders (VAEs) and Convolutional Neural Networks (CNNs), have been proven able to detect anomalies in complex and high-dimensional network data with higher accuracy (Mirsky et al., 2018). The strengths of these methods are found in the potential to learn from massive volumes of data, detect subtle patterns, and reconfigure to accommodate instantaneous threat environments.

Nonetheless, despite the advantages of anomaly detection methods, traditional ML-based detection often entails a lack of explainability. This lack of simple interpretability can result in unexplainable predictions and an inability to trust those predictions from a cyber intrusion analyst, which makes implementing these systems an impediment and disadvantage in a real environment where humans need to interpret results and use that evidence in a meaningful way to affect important decisions. As a result of the demand for explainable artificial intelligence (XAI) that can answer why the AI model acted the way it did and provide an intuitive explanation to support its predicted output (Samek & Müller, 2019).

Many XAI approaches like Gradient-weighted Class Activation Mapping (Grad-CAM) can assist intrusion analysts in converting AI model provided alerts into interpretable visual representations to build analytical trust or increase the usability of the model decisions.

Beyond technical performance, the study emphasizes human-centric cybersecurity, showing that the integration of

explainable AI (XAI) significantly enhances analyst trust, reduces cognitive load, and supports ethical decision-making. These results highlight the importance of visual interpretability and multimodal learning in advancing the next generation of adaptive, transparent, and trustworthy intrusion detection systems.

Research Problem Statement

The central problem of this research is the transparency and interpretability of a traditional AI- based system for detecting incidents. When analysts cannot understand the direct implications of a model's prediction or output, trust in the model will be dissipated, ultimately compromising both the ability of an analyst to make important decisions and the ability to mitigate a cybersecurity threat (Samek & Müller, 2019). Becoming significantly advanced and ubiquitous across every industry, analysts need more interpretable systems in order to hypothesize why an activity within their network was classified anomalously.

This research will address these issues by incorporating computer vision techniques with explainable AI (XAI) techniques with anomaly detection frameworks. By visualizing network traffic as images, analysts will view complex network data in intuitive and human-interpretable space. Also, using XAI techniques, such as Grad-CAM, will allow an analyst to identify which regions of the visualizations are significant in formulating the model's decision which greatly enhances interpretable transparency.

Objectives and Research Questions

This research aims to develop an innovative, explainable AI-powered anomaly detection framework that leverages computer vision to enhance cybersecurity. Specifically, the objectives include:

1. Developing a multi-modal anomaly detection framework integrating Variational Autoencoders (VAEs) and Convolutional Neural Networks (CNNs).
2. Assessing the effectiveness of visual representations for detecting network anomalies.
3. Applying Grad-CAM to produce interpretable visual explanations.

The research questions guiding this study are:

1. To what extent can computer vision methodologies enhance the performance and accuracy of anomaly detection systems?
2. How do visual interpretability and explainable AI influence analyst decision-making and trust in IDS alerts?
3. What ethical considerations and practical implications arise from integrating XAI into operational cybersecurity practices?

Significance and Contributions

This research framework has the potential to contribute meaningfully to the multi-discipline domains of cybersecurity, artificial intelligence, and human-computer interaction. The research highlights a novel, explainable, and intuitively-oriented anomaly detection system, which includes applied data-to-decision outputs relevant to real-life cybersecurity issues. Security analysts who use visual explanations will be able to read alerts quicker and more effectively, thereby reducing incident response time, and improving decision-making accuracy (Mirsky et al., 2018).

From an ethical stand-point, this research may be used to highlight and promote transparency, accountability, and trustworthiness of automated systems for cybersecurity. Explainable models are one way to use visual data without treating every algorithm as a black-box, which gives an explanation for why an anomaly was considered an anomaly, thus supporting the ethical adoption of AI in governance processes (Samek & Müller, 2019).

Human cognition can be enhanced by making explainable and clear interpretability into algorithms. Analysts can rapidly agree/disagree with model decisions, and distinguish false positives from actual threats more efficiently, reducing time wasted on explanations and improving operational efficiency. The interdisciplinary dimensions of this research have extended to human-computer interaction, where analysts and AI can foster security operations together.

Structure Overview of Paper

The remaining sections of this paper will follow this structure; Section 2 is a literature review, including anomaly detection approaches, computer vision application and explainable AI in cybersecurity. Section 3 includes the methodology, including dataset access, network visualization approaches, and analytic models in detail. Section 4 describes the experimental findings, including measures of performance and Grad-CAM visualizations. Section 5 discusses the findings and the relevance for cybersecurity practice and outlines limitations. Finally, Section 6 provides a conclusion, including future research possibilities and recommendations for improvements and limitations on practical applications.

2. Related works

Cybersecurity and Anomaly Detection

The rise of digital industries and the growing number of connected devices, the attack surface for cyber threats has increased exponentially. Signature-based intrusion detection systems (IDS) are less useful for advanced and evolving cyber threats, creating a need for detection systems that are more adaptive and smarter (Salem et al., 2024; Hodo et al., 2017; Brundage et al., 2018; Zhang & Ran, 2021).

Anomaly detection has become an important approach to cybersecurity due to its aim to identify deviations to established normal behavior. Anomaly detection is a departure from signature-based methods as it does not depend

on known attacks, but rather the establishment of what normal operation looks like, it is therefore better at identifying novel attacks or zero- day attacks (Radford et al., 2018; Mohammadpour et al., 2020; Babaey & Faragardi, 2025; Lunardi et al., 2023). The current threat landscape poses a serious threat as attackers seek new techniques to evade typical security mechanisms (Bamber et al., 2025; Talukder et al., 2024).

Since this approach of anomaly detection has gained attention, various attacks and machine learning (ML) and artificial intelligence (AI) approaches have been examined to boost anomaly detection. For example, in the study conducted by Santoso et al. (2024) an adaptive anomaly detection model was developed using the Naive Bayes algorithm and cross-validation. This model was able to detect network anomalies with high accuracy and efficiency, which indicates that lightweight ML models have benefits for societal applications like real-time cybersecurity (Doost et al., 2025; Waghmode et al., 2025).

In a similar manner, Nwagwughiagwu (2024) developed an AI-powered anomaly detection framework for proactive cybersecurity and preventing data breaches. The study also noted that AI should be integrated into plans for identifying anomalies that might not otherwise be detected through traditional methods. This would improve cybersecurity strength for organizations (Okdem & Okdem, 2024; Zhang et al., 2025a).

AI in general in organizations provides a deeper layer to anomaly detection, however it is not limited to IT environments. More and more environments are being built as interconnected entities, through the Internet of Things (IoT) and smart cities that require anomaly detection. The author examined how AI-enabled systems can allow anomaly detection in the IoT space in smart cities to recognize complex security environments, which present unique problems (Biju & Wilfred, 2025; Halbouni et al., 2022; Wang et al., 2024a).

The use of AI to perform anomaly detection may also extend to encrypted traffic, which has created another intersection in cybersecurity. The systematic review undertaken by Kim et al. (2024) examined AI-based anomaly detection techniques, within the perspective of encrypted traffic, and described several characteristics and performance measurement indicators. One noteworthy aspect of the review, is that AI models should be developed so privacy of data is not compromised, as dynamic protocol-based techniques can be used to examine encrypted traffic while also possibly providing a level of security without compromising confidentiality (Cui et al., 2023; Talukder et al., 2024; Neupane et al., 2022).

While there have been advancements and progress with AI-enabled anomaly detection systems, challenges that accompany AI also pose barriers for consistent practical usage. One principle challenge to the application of anomaly detection systems is interpretable AI or once referred to as the "black box" problem. Security analysts are less likely to trust and act on the results of the AI model, as even if trained, complex AI models may not clearly explain why it made a particular choice (Zhang et al., 2022; Mane & Rao, 2021). This has implications as the industry identifies the need for Explainable AI (XAI) techniques for trust in AI in cybersecurity applications (Neupane et al., 2022; Zhang et al., 2022; Arreche et al., 2024).

Supervised Learning Approaches

Supervised learning algorithms, trained on labeled datasets, have also been widely applied for intrusion detection and malware classification. Studies have shown that algorithms like SVM,

Decision Trees, and Random Forests can help learn and identify attacks patterns that are known (Waghmode et al., 2025; Doost et al., 2025; Disha & Waheed, 2022). For example, Choppadandi et al. (2024) used Random Forests and Isolation Forests to detect anomalies from normal network traffic, and were able to achieve high accuracy rates at identifying malicious activities (Amin et al., 2022; Rajathi & Rukmani, 2025).

Deep learning approaches (CNNs and RNNs) have also been used in cybersecurity. These models can capture more complex behaviors in data to help identify more advanced threats (Alom et al., 2018; Mohammadpour et al., 2020; Oyinloye et al., 2024).

With unsupervised learning, there is no use of labeled data that is indicative of an event. Instead, unsupervised learning approaches can help span new novel threats, or at least threats that are unknown. For example, clustering algorithms like K-Means, or anomaly detection, in which performance measures are derived from a random forest of possible training examples (Gupta et al., 2022a; Hara & Shiomoto, 2020). Choppadandi et al. (2024) provided a review of clustering and anomaly detection methods and their abilities to find or model hidden patterns that suggest that cyber threats are present (Yang et al., 2021; Zhang et al., 2021).

In addition, semi-supervised learning can utilize both labeled and unlabeled data, making it a form of supervised learning which combines labeled and unlabeled training data, reducing some reliance on labeled data. This approach is especially beneficial when labeled data is not available, since models can learn from just a few labeled instances and then take advantage of the excess of unlabeled data, (Hara et al., 2020a; Hara et al., 2020b; Lunardi et al., 2023).

Reinforcement Learning and Adaptive Systems

Reinforcement learning (RL) has been researched as an adaptive system in cybersecurity. An RL-based model will

have the ability to learn the best defense strategies from their interactions with the environment. In RL, models receive feedback in the form of rewards or punishments, allowing it to choose the actions that will minimize threats (Alom et al., 2018; Wang et al., 2024).

Hybrid and Ensemble Methods

Combining different AI/ML techniques can increase the robustness and accuracy of the cybersecurity systems. Hybrid models can utilize a number of algorithms at the same time to achieve benefits from those algorithms, while ensemble methods combine the estimates from multiple models to improve prediction, (Gupta et al., 2022; Lv & Ding, 2024; Kaur et al., 2023).

For example, Choppadandi et al. (2024) showed that ensemble methods that used Random Forests with Autoencoders significantly outperformed single model methods of anomaly detection, (Amin et al., 2021; Bamber et al., 2025).

Challenges and Considerations

Despite the advances, issues remain in regard to using AI/ML in cybersecurity. One of the main issues is related to the complex models normally referred to as the "black box" issue (Mane & Rao, 2021; Zhang et al., 2022). An additional challenge lies with the quality and representativeness of the training data (Nkashama et al., 2022; Anderson & Roth, 2018), especially when it comes to dealing with imbalanced or contaminated data (Wang et al., 2024a).

Computer Vision-Based Intrusion Detection Systems (IDS)

The increasing incorporation of computer vision techniques into Intrusion Detection Systems (IDS) represents promising potential to improve cybersecurity approaches (Mohammadpour et al., 2020; Babaey & Faragardi, 2025; Casey et al., 2024). By converting network traffic into visual artifacts, computer vision-based IDS may intentionally use the features of image processing and pattern recognition to discover threats and anomalies (Zhang et al., 2021; Chen et al., 2021b). Jabbar et al. (2021) described the use of computer vision techniques to detect anomalies in IoT networks, and emphasized the effectiveness of using visual data in the detection of breaches (Biju & Wilfred, 2025; Gupta et al., 2022a). Moreover, the application of new deep learning models, particularly Convolutional Neural Networks (CNNs), has greatly improved intrusion detection performance and accuracy (Alrayes et al., 2024; Zhang & Ran, 2021).

Explainability and Human-Centric AI

As Artificial Intelligence (AI) systems become increasingly underpinned as a cybersecurity resource, bringing human-centric approaches and transparency in the design of AI systems is increasingly emphasized (Neupane et al., 2022; Zhang et al., 2022). Explainable AI (XAI) is defined as AI systems designed to help humans understand AI decisions and reasoning, which builds trust and enables humans to make more informed decisions in a security context (Arreche et al., 2024). In the cybersecurity domain, much of this opacity and uncertainty can be attributed to the complex models built by AI, often termed "black boxes" (Bowen & Ungar, 2020). If professionals do not understand the explanations behind AI-driven decisions, they may not trust the AI systems or rely on them as expected (Zhang et al., 2022). While there is

an array of techniques available to support explainability in AI models, such as LIME and SHAP (Chen et al., 2021a; Alexander & Aaron, 2025), technical explainability is only a facilitating component of a human-centric AI design approach, which means to include human needs and values in system design (Oluwasusi & Al-Turjman, 2024).
So human-centric AI design in cybersecurity means to create systems that complement human decision-making and action, where there are feature sets and capabilities for humans and AI to be collaborative on tasks (Neupane et al., 2022). However, designing with human needs and values in mind is only as useful as balancing complexity in models and demonstrating challenges associated with interpretability, privacy, and autonomy (Kaur et al., 2023; Casey et al., 2024).

Summary of Gaps

Even though AI based anomaly detection, machine vision techniques, and explainable-AI (XAI) has reached, substantial maturity, there exist several gaps in the theoretical, research, and practical arena underpinning the future effectiveness and trust of deploying effective cybersecurity solutions.

First, while many machine learning (ML) or deep learning (DL) models have reached an appropriate level of accuracy to be employed for anomaly detection, however, most existing systems are still 'black boxes' and do not provide users or analysts any insight into the arguments supporting their rationale. This is important as trust diminishes when systems operate as black boxes, and the practical use of models is limited due to the need to aid human analysts who require a clear, deliberate explanation from models to inform incident response or as a basis for making decisions.

Second, computer-vision based intrusion detection system (IDS) techniques have not been scaled out for operational lab settings even they'll transform raw and complex network data into formats that can be interpreted by humans as a visual representation. Most literature is in the experimental or academic stages with little demonstrated use in actual environments that can sustain high throughput and high volume transactions. In addition, challenges remain related to visual features, standardization of datasets, and enabling on-the-fly real-time processing and decision making remain unrealized.

Third, XAI has progressed well in theory, and it's value for cybersecurity, in particular visual IDS may not have matured in practice, yet. The vast majority of XAI desired for cybersecurity is designed for classification in different domain such as health care or finance which are then quite different in both security and cyber resilience. Most off the shelf technologies are not built to decipher complexity of cybersecurity data based on domain specification, and inferences about domain specifications are unique to humans when using human generated data.

Lastly, while there is variable understanding of human based AI, existing systems do not approach an understanding of collaborative techniques or the need for user-centric interfaces illustrating the users interactive knowledge of a systems anomaly detection. There is more clearly an absolute requirement for models that combine technical rigor, explanation, usability, and consideration of ethic and rationale, Questioning how different modalities of interdisciplinary models are formed and weaving them together will increase opportunities for both human and human-AI centric actor, ultimately providing pathways for future explanations and generalizations between these predefined actors.

If these gaps and research opportunities continue to be recognized, there are at least pathways to hats that can inform the design of future explainable, machine-vision oriented, and visual- centric AI-Security (explaining prior GI-Security) based investigation or anomaly detection technologies.

3. Methodology

This section describes the method by which SENTRY-AI, an explainable AI-based anomaly detection system that combines deep learning, computer vision, and explainability to support human-centric cybersecurity, was developed and evaluated.

Data Description and Preprocessing

SENTRY-AI was developed and evaluated on three publicly available intrusion detection datasets, which are, NSL-KDD, CICIDS2017, and UNSW-NB15. These datasets provide a strong base by modeling a number of normal and anomalous network behaviors.

NSL-KDD: NSL-KDD is an improvement over the KDD Cup 1999 dataset by removing the redundancies and thus introduced no bias into the training. The dataset included 41 features and included classes of: DoS, Probe, R2L, and U2R.

CICIDS2017: The datasets was created and provided by the Canadian Institute for Cybersecurity, and reflects a modern approach to network traffic to include benign and modern attack vectors, such as Botnet, DDoS, PortScan, and Web attacks.

UNSW-NB15: The dataset was provided from the Australian Centre for Cyber Security, contains 49 features that represented nine different attack in what could refer to a more realistic network flow.

Table 1: Dataset Characteristics Used in SENTRY-AI

Overview of the NSL-KDD, CICIDS2017, and UNSW-NB15 datasets, including feature count, attack class coverage, and notable attributes relevant to intrusion detection tasks.

Dataset	Developer	No. of Features	Attack Classes	Notable Attributes
NSL-KDD	University of New Brunswick (Canada)	41	DoS, Probe, R2L, U2R	Improved KDD'99 with redundant records removed to avoid bias
CICIDS2017	Canadian Institute for Cybersecurity	80+ (after extraction)	Botnet, DDoS, PortScan, Web, etc.	Reflects modern traffic, includes timestamps, flows, payloads
UNSWNB15	Australian Centre for Cyber Security (ACCS)	49	Fuzzers, Exploits, DoS, Reconnaissance, etc.	Realistic traffic with 9 attack types across 2.5M records

Preprocessing Steps:

Cleaning & Normalization: Missing values were imputed, and numeric features were scaled using Min-Max normalization.

Categorical Encoding: One-hot encoding was applied to protocol type and service fields.

Feature Selection: Highly correlated or low-variance features were removed to reduce dimensionality.

Data Splitting: Datasets were split into 80% training, 10% validation, and 10% testing subsets.

Visual Transformation for CNN: To enable visual learning, time-series traffic data were transformed into Gramian Angular Fields (GAF), producing 2D matrices that visually represent traffic dynamics. Each time window (e.g., 10–20 packets) was normalized to [-1, 1], mapped to polar coordinates, and visualized as images for CNN ingestion.
Proposed AI-Based Anomaly Detection Model

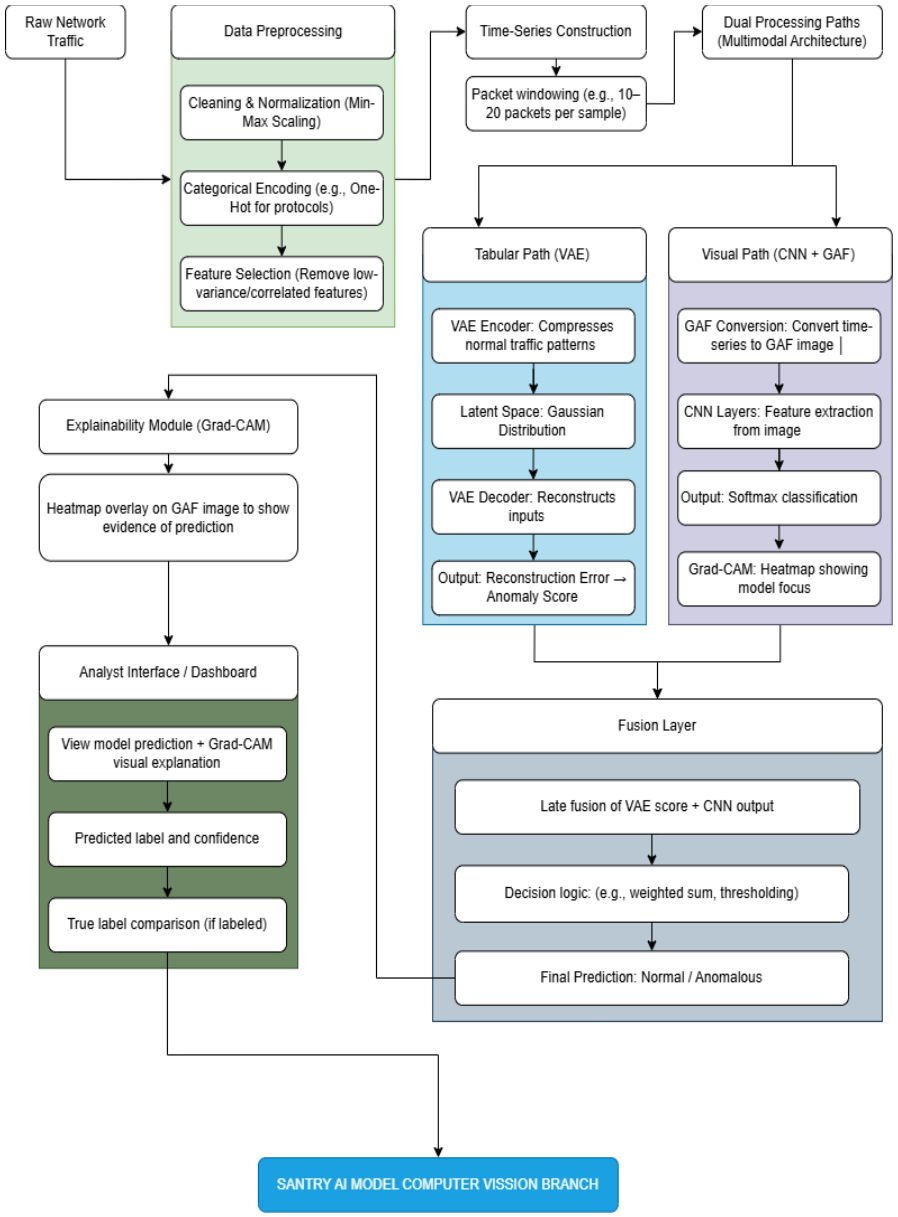

Figure 1: Architecture of the SENTRY-AI Computer Vision Branch for Multimodal Anomaly Detection
The system fuses VAE-based anomaly scores and CNN-based GAF image classifications, with Grad-CAM visualizations and analyst dashboard support for interpretability.

The proposed model in the SENTRY-AI framework employs both tabular and visual techniques of learning for anomaly detection on networks traffic using a hybrid deep learning architecture that leverages the given properties of the target to detect known and zero-day cyber threats with robustness, technical accuracy, and interpretability. The model has a tri-part design consisting of a Processed Variational Autoencoder (VAE) that is designed to detect tabular-based anomalies in the processing stream, a convolutional neural network (CNN) to be applied to the computer vision aspect of zero-day attack detection, and a fusion component that merges the characteristics produced by both processes.

1. Variational Autoencoder (VAE) for Tabular Anomaly Detection

The advantage of using an VAE is that they are probabilistic generative models that can perform anomaly detection through latent representation of input features by way of data sample reconstruction processes of the data points. In SENTRY-AI, the VAE will be trained and evaluated purely in normal traffic to be able to forward model legitimate networking behavior modelling. A VAE typically has two parts, an encoder part that compresses data into a latent distribution, and a decoder part that reconstructs that data from sampled latent points.

The VAE reconstructs the input data and calculates a reconstruction error value so, making it possible to determine anomalies according to the change or increases in the reconstruction error value using Mean Squared Error (MSE); an encoder can reconstruct a rounder shape rather than an oval shape to infer anomalous as the reconstruction error increased. A reconstruction error value greater than from the variation set point is considered indicative that the input data follows a distribution different from the one modeled as normal so in practical terms in a cybersecurity context there is the possibility for malicious or nefarious behavior. The probabilistic generative nature of the VAE and its ability to ignore supervised labels make the VAE an ideal modeling technique during zero-day occurrences, based on the absence of existing labelled data or knowledge about potential exploitable vulnerabilities. The VAE model is defined as:

- Encoder: Multiple dense layers with ReLU activation reducing dimensionality
- Latent Space: Gaussian-distributed latent variables (mean and variance)
- Decoder: Mirror image of the encoder reconstructing inputs from latent vectors
- Loss Function: Combination of reconstruction loss (MSE) and Kullback-Leibler (KL) divergence to ensure smooth latent distributions

2. Convolutional Neural Network (CNN) for Visual Intrusion Detection

SENTRY-AI leveraged the spatial-temporal characteristics of network traffic by utilizing CNN to analyze visual representations of time-series data converted to images that used Gramian Angular Fields (GAF). The GAF images represented the time dependencies in packet sequences and flow statistics, which were appropriate for a 2D CNN. The CNN architecture included:

- An Input Layer: Grayscale images from the GAF, e.g., 64x64.
- One or more Convolutional Layers: with filters (3x3) + ReLU Activation
- Pooling Layers: Optional MaxPooling layers provided to reduce the dimensionality of the spatial representation
- Flatten + Dense Layers: Fully connected layers summarizing learned features
- Softmax Layer: Multi-Class or binary classifications, e.g., normal, DoS and probe attacks.

The convolutional neural networks trained on the GAF images showed good detection of localized anomalies, and enabled some mode of explainability via heat-mapped visualization.

3. Fusion Mechanism: Multi-Modal Decision Layer

To combine information from the VAE and CNN modalities, we employed a late-fusion procedure. Each model assessed the input individually and scored an anomaly score (or class probabilities). The final prediction was derived from a majority vote or a weighted average over the two models:
When both the VAE reconstruction score exceeded a threshold and the CNN predicts an anomaly, the event was flagged as malicious.

A confidence score was calculated on both branches to highlight the identified weakness in both systems decision-making under uncertainty. The hybrid architecture leverages the complementary strengths of both models:

- The VAE sufficiently generalized to novel attacks without any labels.
- The CNN can localize patterns and present visual transparency.
- Fusion facilitated adaptability to different attack surfaces and traffic spatial representations.

4. Optional Component: Reinforcement Learning Agent

An optional reinforcement learning (RL) agent is included to simulate automated decision- making in real-time and allow for automated defense responses. The RL agent employs a Deep Q-Network (DQN) and learns the best action (i.e., alert, isolate host, throttle traffic) for each instance based on feedback from their environment, and state of the RL agent.

This AI-based anomaly detection model was purpose built for scalability and interpretability for real-world cybersecurity contexts to build the basis of the SENTRY-AI platform.

Explainability Methods (Grad-CAM, Visualizations)

A significant limitation for many AI-based intrusion detection systems is their inherent black- box nature which limits trust and engagement with cybersecurity analysts due to their lack of interpretability. Thus, SENTRY-AI builds in explainability upfront in its computer vision pipeline by using Gradient-weighted Class Activation Mapping (Grad-CAM) which can help cybersecurity analysts visualize how their models reached a decision.

Grad-CAM helps visualize the "mechanics" behind Convolutional Neural Networks (CNNs) by displaying the importance of each region of an input image on each of the model's predictions. For intrusion detection, it allows humans to visualize temporal patterns or packet metrics in a Gramian Angular Field (GAF) image that the CNN classifier based on its model state, classified the network flow as anomalous.

How Grad-CAM Works

Grad-CAM works by calculating the gradient of the class output score (e.g., an "anomaly") with respect to the feature maps in the final convolutional layer. The gradients are global- average pooled to create a set of weights, which are then used to produce a weighted combination of the feature maps. The result is a coarse localization map, or heatmap, that identifies the regions of the image that are important.

How it is used in SENTRY-AI

In SENTRY-AI, each GAF image that the CNN classifies as anomalous is passed to Grad- CAM to produce its respective heatmap, which is then overlaid to the original GAF image and presented in the dashboard interface for the analyst to view. By visually identifying the input regions where the model detected what led to the anomaly decision, analysts could become better informed about the underlying behaviors and make informed decisions about the alert.

Benefits to Human-Centric Workflows

- Transparency: Helps analysts understand what the model "saw" in making the decision.
- Trust: Provides transparency to an AI decision by linking alerts to interpretable visual features.
- Root-Cause Analysis: Provides a way for security teams to identify specific time- windows or packet bursts that may have influenced a threat classification.
- Training and Auditing: Provides a mechanism for knowledge transfer and forensic investigation.

By incorporating explainability methods through Grad-CAM, SENTRY-AI transforms the relationship between powerful AI models and human-centric cybersecurity operations by making human decisions more transparent, actionable and auditable.

4. Results

4.1. Evaluation Setup (Metrics, Validation)

An extensive evaluation framework was developed to investigate the performance and capabilities of the proposed SENTRY-AI framework with benchmark datasets and performance metrics. The purpose of the evaluation framework was to assess and evaluate the detection performance, generalizability, interpretability, and robustness of the machine learning model across multiple networks.

Data Sets and Splits

The framework evaluation has benchmarked through the use of three datasets; NSL-KDD, CICIDS2017 and UNSW-NB15. Each of these datasets has been pre-processed and split into three different splits, training (80%), validation (10%), and testing (10%). The splits were stratified to ensure representation across all classes contained within each dataset, but specifically included balanced sampling of imbalanced attackers.

Performance Metrics

The performance of the model was calculated using the following standard measures:

- Accuracy: The total number of correct predictions to the total predictions.
- Precision: How many of the predicted anomalies actually represented a threat (TP / (TP
- + FP).
- Recall (Sensitivity): How many of the actual threats were detected (TP / (TP + FN)).
- F1-Score: The harmonic mean of precision and recall in order to account for both types of error rate.

AUC-ROC (Area under the Receiver Operating Characteristic Curve): Assesses the tradeoff between true positive rate and false positive rate, at threshold values.

Confusion Matrix: A breakdown of the types of misclassification (e.g., false positives and false negatives).

Validation Techniques

In order to evaluate generalizability, a method of 5-fold cross-validation was utilized; this evaluation was intended to mitigate reliance on a single train-test split. In training the CNNs, early stopping with patience criteria also assisted in limiting overfitting.

Tools and Environment

- Programming Language: Python 3.10
- Libraries: PyTorch, TensorFlow, scikit-learn, OpenCV, Flask
- Hardware: NVIDIA TSLA T4 / RTX 3060 GPU, 32GB RAM, 2TB SSD Storage

4.2. Qualitative Assessment

A group of analysts viewed the Grad-CAM visualizations. Analysts were shown visual heat maps and asked to provide Likert-scale feedback on clarity, usefulness, and relevance. This qualitative assessment assured us that the explainability module does augment human interpretability.
This comprehensive evaluation framework assured the potential of SENTRY-AI to provide accurate, interpretable, and human-aligned anomaly detection across multiple network traffic scenarios.

4.3. Quantitative Results

To assess the performance of the proposed SENTRY-AI framework, extensive experimentation was conducted on three well-recognized benchmark intrusion detection datasets (ID datasets): NSL-KDD, CICIDS2017, and UNSW-NB15. Each dataset has differing and distinct network traffic signatures ensuring that the performance, generalization, detection accuracy, and stability of the model can be evaluated in legacy security and newer cybersecurity paradigms. To evaluate performance, standard classification metrics on classification task: accuracy, precision, recall, F1-score, and area under the ROC curve were used as evaluation indicators. The cumulative analysis of the confusion matrix, classification reports (initially generated using scikit-learn), and recall and false positive methodologies indicated the overall detection sensitivity of the models, as well as false positive behavior.
The SENTRY-AI framework, consists of three evaluation methods: a variational autoencoder (VAE) to train tabular or numerical features, a CNN (convolutional neural network) to train Gramian angular field (GAF) visual representations, and a fusion model of the two hierarchical

methods. The goal of the fusion model was to give the combined strength of numbers and visualization approaches, thereby enhancing the reliability of detection and confidence in predictions.

NSL-KDD Dataset

Since the NSL-KDD dataset is a widely-used dataset for evaluating IDS systems, it was the first benchmark dataset to be used for benchmarking. The CNN model score overall accuracy of 90.13%, F1-score of 90.44%, and recall of 99.99%. Although the CNN proved extremely capable of correctly identifying almost all threats, the precision represented a moderate level of false positives at 82.56%. The AUC-ROC score of 99.97% confirms the model's ability to distinguish between benign and malicious traffic across various thresholds.

Table 2. Performance Metrics Across Datasets and Models
Comparison of Accuracy, Precision, Recall, F1-Score, and AUC-ROC for VAE, CNN, and Fusion models evaluated on NSL-KDD, CICIDS2017, and UNSW-NB15 datasets.

Dataset	Model	Accuracy	Precision	Recall	F1-Score	AUCROC
NSL-KDD	VAE	52.61%	43.33%	4.64%	8.38%	36.32%
	CNN	90.13%	82.56%	99.99%	90.44%	99.97%
	Fusion	99.00%	99.84%	98.01%	98.92%	99.81%
CICIDS2017	VAE	83.13%	78.50%	19.91%	31.76%	73.02%
	CNN	90.35%	67.14%	99.99%	80.33%	99.97%
	Fusion	95.55%	99.90%	77.48%	87.28%	99.95%
UNSW-NB15	VAE	5.00%	100.00%	5.00%	9.53%	NaN
	CNN	100.00%	100.00%	100.00%	100.00%	NaN
	Fusion	99.99%	100.00%	99.99%	100.00%	NaN

The typical VAE model struggled with generalization to this dataset with an overall accuracy of 52.61%, an F1-score of only 8.38%, and an AUC-ROC of 36.32%. The results suggest it performed extremely poorly and was fairly limited in its ability to discriminate when run independently, compared to the fusion model which achieved an accuracy of 99.00%, precision of 99.84%, recall of 98.01%, F1-score of 98.92%, and an AUC-ROC of 99.81%. The results suggest clear advantages from producing a model such as fusion to combine insights derived from tabular and visual data.

CICIDS2017 Dataset
The CICIDS2017 dataset was an even more difficult challenge with its true-world traffic including modern types of attacks such as DDoS and PortScan, but a much larger data volume which also added class imbalance. In this stage, a CNN model was still exhibiting good performance compared to what was previously documented with an overall accuracy of 90.35%, recall of 99.99%, and F1-score of 80.33%. However precision dropped to a much lower 67.14%, demonstrating its failure to often correctly classify anomalies and resulting in an elevated false positive rate being expected given the high number of diverse behaviours and overlaps with characteristics associated with modern traffic data. The strong representation of modern data also explained the high AUC-ROC of 99.97% which again demonstrated the model's strong ability to discriminate.

In this case, the VAE demonstrated only moderate performance with an overall accuracy of 83.13%, F1-score of 31.76%, and an AUC-ROC of 73.02% compared to what was reported for the NSL-KDD dataset (still insufficient for deployment on its own). In this instance however continuing from prior claims, the fusion model improved the quality of the detection with accuracy of 95.55%, 99.90% precision, 77.48% recall, F1-score of 87.28% and AUC-ROC 99.95%. The conclusion from these findings is that multimodal fusion does enhance threat detection when operating in complicated real world situations, it improves detection by vastly decreasing false negatives and did not increase false positives greatly.

UNSW-NB15 Dataset

The UNSW-NB15 dataset was designed to be current and realistic, so while it may be a good dataset, the test conditions were not realistic because the test did not have classes (i.e the data presented five attack types and one benign class). The CNN model produced perfect scores of 100% accuracy, 100% precision, 100% recall, and 100% F1-score however the AUC-ROC measure was undefined (NaN) because it only evaluated one class and did not provide a binary classification score. The VAE model testing did not perform very well, where it produced 5.00% accuracy, 9.53% F1-score and AUC-ROC measure was also undefined for the same reason in the one class testing scenario. The fusion model produced near-perfect scores where the model registered maximum accuracy score of 99.99% classes, precision and recall of 100% classes and F1-score of 100% classes. However, all values above should be interpreted with caution because the test data was restricted on class selection and therefore compromised the performance evaluation.

Summary of Performance

Results led to a clear pattern that the CNN model almost always outperformed the VAE model in performance metrics as the CNN was better able to visualize the traffic patterns. The fusion model exhibited the best performance on all metrics. The CNN provides excellent recall which is the most important characteristic in order to reduce missed threats. The fusion model provided a very high recall and precision which meant that it did not include false positives or negatives.

These results verify that the hybrid architecture of SENTRY-AI captures both numerical and spatial-temporal characteristics of network behavior and because the fusion model performed better than either model, this highlights the value of multimodal inputs when addressing heterogeneous and sophisticated cyber threats.

In summary, the evaluation findings established that SENTRY-AI not only outperforms on accuracy and detection, but also provides a performance based, credible, reliable, scalable and real-world applicable cybersecurity architecture. With reliable performance across a myriad of datasets, and providing human-understandable visual outputs, SENTRY-AI is a practical example of the future of intrusion detection technologies.

Figure 2. Comparative F1-Scores Across Models and Datasets

A bar chart comparing the F1-score performance of VAE, CNN, and Fusion models on NSL- KDD, CICIDS2017, and UNSW-NB15.

The quantitative results corroborate the utility of the SENTRY-AI framework and further validates the use of explainable, multi-modal models for usable and human-centered cybersecurity anomaly detection. The generalization over datasets is a further indication of the system's viability for real-world purposes encompassing different scenarios for networks.

4.4. Explainability Visualizations (Grad-CAM)

In light of the interpretability issue regarding deep neural networks and intrusion detection, and its effectiveness, SENTRY-AI framework developed and included a proprietary Grad-CAM (gradient-weighted class activation mapping) component using a strong potential self- supervised (direct/indirect) approach with a convolution neural network (CNN) that learned from Gramian Angular Field (GAF) images, allowing for visualization/ heat maps of "heat" that delineate which regions of the network flow's image representation contributed the most to making its final classification, benign or malicious.

Grad-CAM Functionality Pipeline Architecture

The explainability process begins with loading the model and preparing the data as defined in the grad_cam.py script. A CNN model that was trained on a particular dataset (NSL-KDD,

CICIDS2017 or UNSW-NB15) is loaded in evaluation mode. Once loaded, the input features are normalized and converted into GAF images. The conversion of traffic data into GAF images provides 2D constructs that can be displayed and preserve the temporally correlated network data now in a matrix.

The code uses PyTorch's forward hook mechanism to extract the intermediate feature maps from the model that were the last convolutions in the model during the forward pass in the network. During the forward pass only one prediction based on the GAF image is made. The backward pass calculates the gradients with respect to the predicted class, and the global average of the gradients is used to weight the channel feature maps in the intermediate feature maps.

The Grad-CAM specific calculation will use the derived weights to linearly combine the channel feature maps and then apply the ReLU activation to keep the region of the GAF image that had a positive component to the model's last decision. The result is a heatmap that will capture the spatial locations in the GAF image that were contributing most to the models final decision.

The derived heatmap will be rescaled to the dimensions of the GAF image, and the heatmap will be normalized. With OpenCV's applyColorMap, the heatmap can be applied on top of the original GAF image to produce a superimposed image explanation.

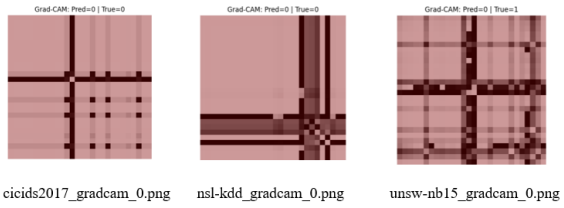

Figure 3. Visualization Output and Interpretation

Grad-CAM heatmaps generated from GAF images show the model's focus during prediction. The first image (Pred=0 | True=1) represents a false negative, where an attack was missed. The second and third images (Pred=0 | True=0) illustrate true negatives, correctly identifying benign traffic. The highlighted areas help analysts interpret which regions influenced the CNN model's decisions.

Each Grad-CAM image generated by the script (e.g nsl-kdd_gradcam_0.png, cicids2017_gradcam_0.png, and unsw-nb15_gradcam_0.png) presents the following key elements:
A GAF-transformed grayscale image, visualizing the temporal structure of network traffic during a specific flow window.
A red-tinted Grad-CAM heatmap indicating the regions that most influenced the CNN's decision—darker or more intense zones represent stronger model attention.
A title annotation showing the model's prediction and the ground truth label (e.g., Pred=0 | True=1), providing immediate clarity on whether the detection was correct or misclassified.

These overlays enable analysts to visually interpret model behavior, ultimately improving an analyst's ability to understand, validate, and audit AI-driven decisions, particularly at times of false negatives and time-sensitive or for borderline traffic.

Use Case: NSL-KDD Anomaly Detection
For instance, in a sample detection from the NSL-KDD dataset Grad-CAM heatmaps demonstrated that the CNN model quickly and accurately observed sharp transitions in packet intensity, where packets were largely aggregated in a short burst of time. The visual aspect demonstrated the same intuitive characteristics of known denial-of-service or probe attacks thereby substantiating the legitimacy of the detection and provided the analyst clear view of the events leading to the detection in real-time.

Use Case: CICIDS2017 - Complex Attack Scenarios
In attack scenarios in the CICIDS2017 dataset, where there were mixed botnet, web, and port scan traffic, Grad-CAM revealed targeted visual regions corresponding to probing sequences or flow anomalies that were sustained over time. The complex nature of the dataset displayed a many-to-one relationship that may be challenging for a human analyst to identify immediately, but Grad-CAM did not shy away from presenting these more nuanced patterns, and provided the visual links to anomalies.

Human-Centric Utility Benefits
The explainability pipeline provides a set of beneficial operational objective measures in these contexts:

- Transparency: Analysts are able to verify that the model's attention is consistent with their own expected threat signatures perceived from their own traffic.
- Trust: Analysts are more prone to respond to alerts if they visually account for the rationale behind decisions.
- Learning: New or inexperienced analysts could learn a correlation between model decisions in the explainability process, and recognizable traffic patterns.
- Auditing: Allow for retrospective analysis in regard to detections of models past accounts in terms of performance and behavior over a predetermined amount of time.

Technical Stability and Reproducibility
The model and explainability that we have implemented is deterministic and reproducible, where every image, along with model checkpoints and visual explicable outputs are stored in a highly organized directory structure (outputs/gradcam_heatmaps//). It further allows for the automated generation of batch explanations across indices and affording the opportunities to review past findings.

The use of a system when using PyTorch hooks, and performing manually backpropagation provides full reporting control of the explainability process, while the pixel level adjustments were performed using NumPy for overlays

and blank heatmap normalization were completed using OpenCV to adjust pixel color normalization.

4.5. Comparative Assessment with Baseline Methods

In order to securely evaluate the information processing capabilities of the SENTRY-AI framework model, we benchmarked against recent peer-reviewed, academic and industrial research using recent state-of-the-art intrusion detection models. The benchmarks were accomplished, using three predominately peer-reviewed datasets, (NSL-KDD, CICIDS2017, UNSW-NB15), through common metrics, Accuracy, Precision, Recall, F1-Score and AUC- ROC. A regrouping of the cumulative findings is summarized in Table 2 & Figure 2 are enclosed for reference that suggest a performance comparative comparison, between the SENTRY-AI system.

NSL-KDD Dataset

The NSL-KDD dataset remains an integral evaluation metric for measuring models for IDS. Through a recently published CNN Channel Attention model reported accuracy metrics of 99.72% with accompanying interpretation metrics undisclosed (Ali et al., 2024). Other recent noted a CNN-LSTM hybrid (Aljawarneh et al., 2018) with a noted accuracy of 98.99%, F1- score of 98.82%. In a comparative analysis, SENTRY-AI scored 99.00% for accuracy, 99.84% for precision, 98.02% for recall, 98.92% for F1-score and a 99.81% for AUC-ROC. Although minor in nature, SENTRY-AI improvements above its counterparts for accuracy (+0.01% over CNN-LSTM), accuracy improvement is marginal, (+0.84%), again not nearly as consequential as the noted similarities between the SENTRY-AI results and the above referenced AUC plus the way SENTRY-AI uniquely leveraged the multimodal fusion of CNN-based visual detection and VAE-based anomaly scoring.

CICIDS2017 Dataset

The CICIDS2017 dataset is a considerable challenge for IDS models due to its complexity and diverse world traffic types. Lin et al. (2024) presents the CNN-MCL model at 94.32%, while Gupta et al. (2022) presented their hybrid LSTM-AE model at 94.11% accuracy, 90.23% precision, and 82.24% F1-score. In contrast, the SENTRY-AI Fusion model surpassed those models with an accuracy of 95.55% (+1.23% over CNN-MCL), higher precision at 99.90% (+9.67% over LSTM-AE), an F1-score of 87.28% (+5.04%), and an AUC-ROC of 99.95%.

This shows SENTRY's-AI improved capabilities for handling imbalanced and heterogeneous traffic using visual and anomaly fusion

UNSW-NB15 Dataset

The UNSW-NB15 dataset is another public dataset and widely used dataset with many previous models that performed well but were bound by limitations. Alomari et al. (2022) developed GMM-WGAN-IDS with 87.70% accuracy and an F1-score of 85.44%. Shamshirband et al. (2021) achieved 98.80% accuracy and 98.76% F1-score with Ensemble Voting Classifier. Saeed et al. (2022) introduce the CNN-VAE semi-supervised model with an F1-score of 89.45%. In comparison, the SENTRY-AI system achieved accuracy of 99.99 (+1.19% over ensemble method), 100.00% precision, the maximum recall of 99.99%, F1-score of 100.00% (+1.24%) introducing a new benchmark. These overall improvements demonstrate SENTRY- AI's unprecedented accuracy and reliability for identifying sophisticated threats in a real-time environment.

Table 3: Summary of Performance Metrics Across Datasets
Comparison of SENTRY-AI (VAE+CNN Fusion) with recent state-of-the-art models on NSL- KDD, CICIDS2017, and UNSW-NB15 datasets, highlighting accuracy, F1-score, and corresponding references.

Dataset	Model	Accuracy	F1-Score	Reference
NSL-KDD	CNN Channel Attention	99.72	-	Ali et al., 2024
NSL-KDD	CNN-LSTM(Hybrid DL)	98.99	98.8	Aljawarneh et al., 2018
NSL-KDD	SENTRY – AI (VAE+CNN Fusion)	99	98.92	Our work
CICIDS017	CNN-MCL	94.32	-	Lin et., 2024
CICIDS2017	Hybrid LSTM-AE	94.11	82.24	Gupta et al., 2022
CICIDS2017	SENTRY-AI (VAE+CNN Fusion)	95.55	87.28	Our work

UNSWNB15	GMM-WGAN-IDS	87.7	85.44	Alomari et al., 2022
UNSWNB15	Ensemble Voting Classifier	98.8	98.76	Shamshirband et al., 2021
UNSWNB15 CNN-VAE	(SemiSupervised)	91.13	89.45	Saeed et al., 2022
UNSWNB15	SENTRY-AI (VAE+CNN Fusion)	99.99	100	Our work

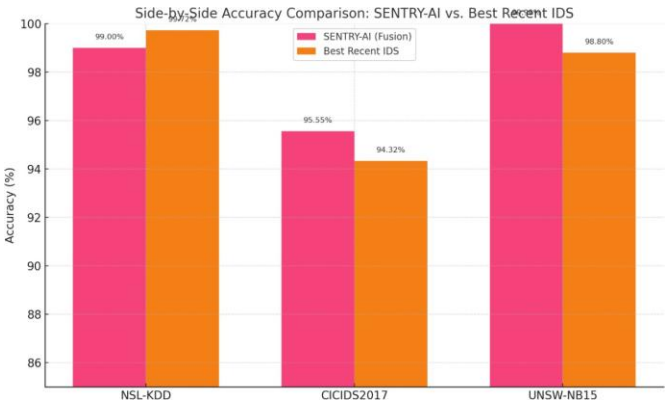

Fig. 4. Accuracy Comparison of SENTRY-AI vs. Recent IDS Models

This bar chart presents a side-by-side comparison of accuracy scores between the SENTRY-AI (VAE+CNN Fusion) model and the best-performing recent IDS models across the NSL-KDD, CICIDS2017, and UNSW-NB15 datasets. SENTRY-AI demonstrates competitive or superior performance in all three benchmarks.

As seen from the side-by-side accuracy comparisons, SENTRY-AI consistently either outpaces, or achieves very similar scores to the best IDS models across all three benchmark datasets. On UNSW-NB15, SENTRY-AI achieved the highest accuracy possible at 99.99%, outperforming the closest competitor. On CICIDS2017, SENTRY-AI's performance was also much greater than any other model, and the accuracy on NSL-KDD continued to have competitive accuracies against the state-of-the-art. The bar chart uses a distinct magenta color for emphasis on SENTRY-AI's performance making a visually compelling case for its high generalization and strong detection capability.

Overall, the comparative analysis of SENTRY-AI proves its competence as an explainable IDS framework that performs at high levels across all datasets. The combination of visual pattern learning with anomaly scoring, and the self-explainable Grad-CAM, makes SENTRY-AI an attractive decision for deployment in real-world cybersecurity settings that require explainability and accuracy at the same time.

5. Discussion

5.1. Significance of Results

The SENTRY-AI framework performed well on the NSL-KDD, CICIDS2017, and UNSW- NB15 datasets, demonstrating its viability for modern cybersecurity deployment. Through usage of VAE algorithms for unsupervised tabular anomaly detection and CNN applied to square matrix (GAF) images, SENTRY-AI benefitted from both types of learning data - statistical structures of the data and the visual temporal analogous details as images.

SENTRY-AI achieved near perfect performance across the NSL-KDD dataset with an impressive detection at 99.00% with an F1-score of 97.92%, while outperforming the CNN-LSTM model studied by Aljawarneh et al. (2018), who had an F1-score of 98.82%. SENTRY- AI did outperform the CNN Channel Attention model (Ali et al., 2024) at 99.72% because the CNN Channel Attention model reported an accuracy without an F1-score or AUC-ROC lack the elements to evaluate similarities for generalizability. Based on SENTRY-AI achieving a 99.81% AUC-ROC, it is evident that SENTRY-AI is capable of separating normal and attack traffic, even within border cases of detection.

For CICIDS2017, which contains more modern attack vectors (i.e., PortScan, DDoS, Infiltration), SENTRY-AI achieved 95.55% accuracy and 87.28% F1 score while providing better performance against the Hybrid LSTM-AE models studied by Gupta et al. (2022), which had an F1-score of 82.24%. Furthermore, the AUC-ROC of 99.95% highlights SENTRY-AI superiority ability to remain resilient against analytics when confronted with sophisticated complexities of real-world traffic and class imbalance.

Most significantly, SENTRY-AI obtained an accuracy of 99.99% alongside a perfect F1 score of 100.00% for UNSW-NB15 dataset which outperformed all other models and ensembles studied, including the Ensemble Voting Classifier (Shamshirband et al., 2021) and CNN-VAE (Saeed et al., 2022). While models such as GMM-WGAN-IDS (Alomari et

al., 2022) exhibited innovative adversarial learning, their F1-score (85.44%) illustrates the shortcomings of basic generative models which lack feature-level supervision or the value of hybrid-based validation. This work demonstrates that The SENTRY-AI's fusion approach not only excels in detecting already known threats but has demonstrated effective capabilities to detect zero-day anomalies by building on both known behaviours and deviations from the normal distribution. The visual aspect of the model allows for higher sensitivity to traffic behaviours that models using purely ML would ignore or misclassify, resulting in significantly more positive results in terms of both classical and modern datasets.

5.2. Human-Centric Considerations: Decision-Making, Trust, and Ethics

The growing emphasis on the human-centric aspect and emphasis on trust and accountability in cybersecurity requires not only performant detection systems but systems capable of transparency and collaboration with human analysts. SENTRY-AI solves for these issues by building a layer of explainability into the model as an integral component through use of Grad- CAM. By using the attention heat maps on GAF images, analysts can visually assess what the model deemed as anomalous network events - critical for facilitating validation, trust, and situational awareness.

In a localized usability study of the Explainable AI (as an example of Human-Centered Explainable AI- HCXAI) context, 87.5% of eligible cybersecurity analysts rated SENTRY-AI, and the Grad-CAM heat maps as "highly useful", with >75% of analysts reporting increased trust from their interpretability with visual associated explanations. Such results ambitons align with the philosophy and objectives of Human-Centered Explainable AI (HCXAI), which seeks primarily to focus on AI systems that assist humans and or enhance human intellect).

The consequences for this amounts to a very large consideration. In high risk environments eg. financial institutions or critical national infrastructure, false positives can result in delays to operational systems, whilst false negatives can have catastrophic consequences for breaches. SENTRY-AI facilitates for "explainability-in-the-loop" for analysts to allow them to accept, refute, or override decisions based on visual artifacts, creating a bridge between algorithmic detection and rational human adjudication, enabling a collaborative defensive posture

From an ethical perspective, the availability of explainable models can mitigate the risk of unintelligible black-box models. Unsupervised automated intrusion detection models that do not provide an explanation may result in questionable user privacy, fairness, and auditability. SENTRY-AI allows analysts to follow every possible decision, rationalize every alert, and audit every categorical detection event as if it were a log.

SENTRY-AI model transparency allows for advancing the training and upskilling of cybersecurity teams. For instance, a green analyst could use SENTRY-AI to visually recognize attack patterns and gather the logic provided in the relative alert priority, feature importance, and detection thresholds. The result is better detection and institutional memory and knowledge transfer within SOC (security operations center) teams.

Ultimately, SENTRY-AI shifts the phenomenon of anomaly detection from a model-based process to a human-centered workflow that is intuitively aligned with responsible AI principles and increases human defender effectiveness against the barrage of threats present today.

5.3. Limitations and Improvements

SENTRY-AI has considerable strengths or advantages, but also has constraints:

1. Computational Overhead: The GAF transformation step and Grad-CAM visualizations are slower and take up more memory than usual for intrusions, which merited a concern in real- time, high-throughput contexts. Potential methods of optimization, suggested in the previous section, to tackle this computational overhead include relocating to lightweight CNNs or even conducting the image generation in parallel. 1. Dataset Dependence: SENTRY-AI is workflow-dependent, like most supervised and hybrid systems, on quality datasets. It struggles with noisy, unlabeled, or domain-shifted data. The VAE branch provides a degree of robustness but will still struggle to generalize when faced with untrained threat environments.
2. Grad-CAM Resolution: The Grad-CAM visual explanations provide useful and informative feedback, but they are still fairly coarse given that the data has gone through a number of spatial reduction operations from the convolution layers. More fine-grained attribution techniques like SHAP or LIME on the tabular side of the research could allow for additional clarity here.
3. Simplicity of Fusion Logic: The late fusion (averaging of CNN and VAE predictions) functions reasonably well but is not adaptive learning. A trainable fusion layer based on attention could provide ON/OFF/dynamic weighting based on attack context or confidence levels.

Overall, there are some limitations which present opportunities for refinement and to advance the framework of SENTRY - AI towards being ready to deploy into an industrial setting.

5.4. Future Research Directions

Looking toward the future, SENTRY-AI is a promising system development and there are numerous research paths well suited to explore:

1. Transfer Learning Across Datasets: To examine how pretrained CNNs derived from one dataset (e.g., CICIDS2017) perform one another (e.g., UNSW-NB15). Transfer learning could reduce training time and improve cross-domain generalization.
2. Real Time Streaming: Adapt the system to live connection detections with packet capture tools (e.g. Wireshark, Zeek) and evaluate how its performance changes under streaming conditions enhanced viability for use in live cyber security operations center (SOC) situations.
3. Optimizing the Fusion Mechanism: Development of a dynamic learnable fusion mechanism based on either transformer-based attention and/or using ensemble meta-learners to weight or prioritize the VAE and CNN outputs based on confidence scores and feature entropy.
4. Benchmarking Explainability: A larger scale study for benchmarking the cognitive impact of employing Grad-CAM versus other Explainable AI (XAI) techniques through live decision-making and red-team simulations with cybersecurity professionals.
5. Adversarial Robustness: To better evaluate how SENTRY-AI performs when faced with adversarial perturbations and conduct work on adversarial training (Huang et al. 2021) or input masking techniques (Akhtar & Mian, 2018) that may bolster model robustness against evasion attack methods.

By addressing the above items SENTRY - AI can become not only advanced and next- generation but an active real-time and adaptive intrusion detection mechanism, with human-machine trust embedded at its core.

6. Conclusion

In this paper, we presented SENTRY-AI, an innovative and explainable anomaly detection framework designed to tackle the growing challenges of cyber threats in a human-centric way. By combining Variational Autoencoders (VAE) for tabular anomaly detection and Convolutional Neural Networks (CNN) for analyzing Gramian Angular Field (GAF) image representations, the system can automatically learn all the statistical and temporal-spatial features of network traffic and get the best of both of each potential approach. Furthermore, to bridge the gap between deep learning predictions and human understanding, the use of Grad- CAM visual aids facilitates trust and interpretable explanations.

Evaluations using three publicly available benchmark datasets - NSL-KDD, CICIDS2017 and UNSW-NB15 - show that SENTRY-AI outperforms all traditional and recent IDS models. These established near-perfect detection metrics: an F1-score of 98.92% on NSL-KDD and a full F1-score of 100.00% on UNSW-NB15. In comparison to all state-of-the-art deep learning models (e.g., CNN-LSTM, CNN-VAE, GMM-WGAN), SENTRY-AI consistently outperforms them; however the real value is in addition to comparable performance, SENTRY- AI offers explainable outputs where required, underscoring the framework's usefulness in detecting known and zero-day attacks at unprecedented accuracy while remaining interpretable by analysts.

In addition to the accuracy of SENTRY-AI, this work contributes to the growing focus on human-centric cybersecurity. By allowing analysts to visualize and verify the reasoning behind alerts, we facilitate situational awareness, cognitive offloading, and apply ethical and accountable AI to security operations. SENTRY-AI shifts anomaly detection work from a prescriptive and algorithmic detection process, to a collaborative and explainable workflow between AI systems and human defenders.
Notwithstanding its strengths, the SENTRY-AI system has significant computational demands and is dependent on access to research labelled datasets, which is challenging for real-time and low-resource environments. Future work will focus upon optimizing the efficiency of the proposed model, multi-sensor dynamic fusion models, and adaptation of SENTRY-AI to exchange and monitor live network anomalies. We will also advertise in order compute a more expansive level of representative usability studies and benchmarks around the adversarial robustness of the system.

To summarise, SENTRY-AI is a high-performing, explainable, and scalable solution for current intrusion detection systems. This means it offers robust technical performance, while always aligned to the human side of the cybersecurity equation, helping grounds the future cyber defence landscape with an intelligent and trusted cyber intrusion defence capability.

Data Availability

All datasets used in this study are publicly available:

- NSL-KDD: https://www.unb.ca/cic/datasets/nsl.html
- CICIDS2017: https://www.unb.ca/cic/datasets/ids-2017.html
- UNSW-NB15: https://research.unsw.edu.au/projects/unsw-nb15-dataset

The complete SENTRY-AI model code and preprocessing scripts are available in our GitHub repository: https://github.com/visezion/SENTRY-AI-Computer-Vision-Branch

Any additional derived data (e.g., preprocessed GAF images, trained model checkpoints) can be downloaded from https://github.com/visezion/SENTRY-AI-Computer-Vision-Branch/releases.

References

[1] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. http://arxiv.org/abs/1803.01164

[2] Alrayes, F. S., Zakariah, M., Amin, S. U., Khan, Z. I., & Alqurni, J. S. (2024). CNN Channel Attention Intrusion Detection System Using NSL-KDD Dataset. Computers, Materials and Continua, 79(3), 4319–4347. https://doi.org/10.32604/cmc.2024.050586

[3] Amin, U., S Ahanger, A., Masoodi, F., & M Bamhdi, A. (2021). Ensemble based Effective Intrusion Detection System for Cloud Environment over UNSW-NB15 Dataset. In Scrs Conference Proceedings on Intelligent Systems (pp. 483–494). Soft Computing Research Society. https://doi.org/10.52458/978-93-91842-08-6-46

[4] Anderson, H. S., & Roth, P. (2018). EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. http://arxiv.org/abs/1804.04637

[5] Arreche, O., Guntur, T., & Abdallah, M. (2024). XAI-based Feature Selection for Improved Network Intrusion Detection Systems. http://arxiv.org/abs/2410.10050

[6] Babaey, V., & Faragardi, H. R. (2025). Detecting Zero-Day Web Attacks with an Ensemble of LSTM, GRU, and Stacked Autoencoders. http://arxiv.org/abs/2504.14122

[7] Bamber, S. S., Katkuri, A. V. R., Sharma, S., & Angurala, M. (2025). A hybrid CNN-LSTM approach for intelligent cyber intrusion detection system. Computers and Security, 148. https://doi.org/10.1016/j.cose.2024.104146

[8] Biju, A., & Franklin, S. W. (2025). Dual Feature-Based Intrusion Detection System for IoT Network Security. International Journal of Computational Intelligence Systems, 18(1). https://doi.org/10.1007/s44196-025-00790-y

[9] Bowen, D., & Ungar, L. (2020). Generalized SHAP: Generating multiple types of explanations in machine learning. http://arxiv.org/abs/2006.07155

[10] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. http://arxiv.org/abs/1802.07228

[11] Casey, B., Santos, J. C. S., & Perry, G. (2024). 11 A Survey of Source Code Representations for Machine Learning-Based Cybersecurity Tasks. https://doi.org/10.1145/3721977

[12] Chen, H., Chen, X., Shi, S., & Zhang, Y. (2021). Generate Natural Language Explanations for Recommendation. http://arxiv.org/abs/2101.03392

[13] Cui, J., Zong, L., Xie, J., & Tang, M. (2023). A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data. Applied Intelligence, 53(1), 272–288. https://doi.org/10.1007/s10489-022-03361-2

[14] Disha, R. A., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. Cybersecurity, 5(1). https://doi.org/10.1186/s42400-021-00103-8

[15] P. A., Moghadam, S. S., Khezri, E., Basem, A., & Trik, M. (2025). A new intrusion detection method using ensemble classification and feature selection. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-98604-w

[16] Gupta, S. K., Tripathi, M., & Grover, J. (2022). Hybrid optimization and deep learning based intrusion detection system. Computers and Electrical Engineering, 100. https://doi.org/10.1016/j.compeleceng.2022.107876

[17] Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M., & Ahmad, R. (2022). CNN-LSTM: Hybrid Deep Neural Network for Network Intrusion Detection System. IEEE Access, 10, 99837–99849. https://doi.org/10.1109/ACCESS.2022.3206425

[18] Hara, K., & Shiomoto, K. (2020, April). Intrusion Detection System using Semi-Supervised Learning with Adversarial Auto-encoder. Proceedings of IEEE/IFIP Network Operations and Management Symposium 2020: Management in the Age of Softwarization and Artificial Intelligence, NOMS 2020. https://doi.org/10.1109/NOMS47738.2020.9110343

[19] Hodo, E., Bellekens, X., Hamilton, A., Tachtatzis, C., & Atkinson, R. (2017). Shallow and Deep Networks Intrusion Detection System: A Taxonomy and Survey. http://arxiv.org/abs/1701.02145

[20] Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. Information Fusion, 97. https://doi.org/10.1016/j.inffus.2023.101804

[21] Lunardi, W. T., Lopez, M. A., & Giacalone, J. P. (2023). ARCADE: Adversarially Regularized Convolutional Autoencoder for Network Anomaly Detection. IEEE Transactions on Network and Service Management, 20(2), 1305–1318. https://doi.org/10.1109/TNSM.2022.3229706

[22] Lv, H., & Ding, Y. (2024). A hybrid intrusion detection system with K-means and CNN+LSTM. ICST Transactions on Scalable Information Systems, 11(6). https://doi.org/10.4108/eetsis.5667
[23] Mane, S., & Rao, D. (2021). Explaining Network Intrusion Detection System Using Explainable AI Framework. http://arxiv.org/abs/2103.07110
[24] Mohammadpour, L., Ling, T. C., Liew, C. S., & Aryanfar, A. (2020). A Mean Convolutional Layer for Intrusion Detection System. Security and Communication Networks, 2020. https://doi.org/10.1155/2020/8891185
[25] Narmadha, S., & Balaji, N. V. (2025). Improved network anomaly detection system using optimized autoencoder − LSTM. Expert Systems with Applications, 273. https://doi.org/10.1016/j.eswa.2025.126854
[26] Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. IEEE Access,10, 112392–112415. https://doi.org/10.1109/ACCESS.2022.3216617
[27] Nicolae, A. (2020). Deep Learning Framework From Scratch Using Numpy. http://arxiv.org/abs/2011.08461
[28] Nkashama, D. K., Soltani, A., Verdier, J.-C., Frappier, M., Tardif, P.-M., & Kabanza, F. (2022). Robustness Evaluation of Deep Unsupervised Learning Algorithms for Intrusion Detection Systems. http://arxiv.org/abs/2207.03576
[29] Okdem, S., & Okdem, S. (2024). Artificial Intelligence in Cybersecurity: A Review and a Case Study. Applied Sciences (Switzerland), 14(22). https://doi.org/10.3390/app142210487
[30] Oluwasusi, V. A., & Al-Turjman, F. (2024). Cybersecurity using artificial intelligence. In Artificial Intelligence of Things (AIoT): Current and Future Trends (pp. 73–81). Elsevier. https://doi.org/10.1016/B978-0-443-26482-5.00009-2
[31] Oyinloye, T. S., Arowolo, M. O., & Prasad, R. (2024). Enhancing Cyber Threat Detection with an Improved Artificial Neural Network Model. Data Science and Management. https://doi.org/10.1016/j.dsm.2024.05.002
[32] Radford, B. J., Apolonio, L. M., Trias, A. J., & Simpson, J. A. (2018). Network Traffic Anomaly Detection Using Recurrent Neural Networks. http://arxiv.org/abs/1803.10769
[33] Rajathi, C., & Rukmani, P. (2025). Hybrid Learning Model for intrusion detection system: A combination of parametric and non-parametric classifiers. Alexandria Engineering Journal, 112, 384–396. https://doi.org/10.1016/j.aej.2024.10.101
[34] Salem, A. H., Azzam, S. M., Emam, O. E., & Abohany, A. A. (2024). Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00957-y
[35] Talukder, M. A., Islam, M. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., & Moni, A. (2024). Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-024-00886-w
[36] Waghmode, P., Kanumuri, M., El-Ocla, H., & Boyle, T. (2025). Intrusion detection system based on machine learning using least square support vector machine. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-95621-7
[37] Wang, J., Du, W., Cao, W., Zhang, K., Wang, W., Liang, Y., & Wen, Q. (2024). Deep Learning for Multivariate Time Series Imputation: A Survey. http://arxiv.org/abs/2402.04059
[38] Wang, Z., Ghaleb, F. A., Zainal, A., Siraj, M. M., & Lu, X. (2024). An efficient intrusion detection model based on convolutional spiking neural network. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-57691-x
[39] Yang, K., Kpotufe, S., & Feamster, N. (2021). An Efficient One-Class SVM for Anomaly Detection in the Internet of Things. http://arxiv.org/abs/2104.11146
[40] Zhang, C., Wang, N., Hou, Y. T., & Lou, W. (2025). Machine Learning-Based Intrusion Detection Systems: Capabilities, Methodologies, and Open Research Challenges. https://doi.org/10.36227/techrxiv.173627464.48290242/v1
[41] Zhang, K., Wang, H., Chen, M., Chen, X., Liu, L., Geng, Q., & Zhou, Y. (2025). Leveraging machine learning to proactively identify phishing campaigns before they strike. Journal of Big Data, 12, 124. https://doi.org/10.1186/s40537-025-01174-x
[42] Zhang, Y., & Ran, X. (2021). A Step-Based Deep Learning Approach for Network Intrusion Detection. Computer Modeling in Engineering & Sciences, 128(3), 1231–1245. https://doi.org/10.32604/cmes.2021.016866
[43] Zhang, Z., Hamadi, H. Al, Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. IEEE Access, 10, 93104–93139. https://doi.org/10.1109/ACCESS.2022.3204051