A KAN-Driven Threat Detection Pipeline for Network Intrusion Detection

Emmanuelli El-Mahmoud[1], Pablo Rivas[2], and Erika Leal[1]

[1] Department of Computer Science, Baylor University, Waco, Texas, USA
{Emmanuelli_Elmahmou1,Erika_Leal}@Baylor.edu
[2] Dept. of Computing Technology, Marist University, Poughkeepsie, New York, USA
Pablo.Rivas@Marist.edu

Abstract. The rapid growth of interconnected digital ecosystems, with IoT networks at the forefront, has introduced heightened vulnerability to complex cyber threats, underscoring the critical importance of developing advanced and scalable Network Intrusion Detection Systems. While hybrid deep learning models successfully capture spatiotemporal dependencies in network traffic, their reliance on opaque Multi-Layer Perceptrons creates a "black-box" dilemma that hinders operational trust and interpretability. Furthermore, these conventional models often suffer from spectral bias, leading to high false-positive rates and poor detection of minority attack vectors in highly imbalanced real-world data. To address these challenges, we propose a novel, explainable hybrid architecture: the CNN-BiGRU-KAN. This framework integrates one-dimensional Convolutional Neural Networks and Bidirectional Gated Recurrent Units for hierarchical feature extraction, replacing the traditional MLP classification head with a Kolmogorov-Arnold Network. By leveraging learnable univariate spline functions on network edges rather than static activation functions, the KAN layer dynamically adapts to complex, non-linear boundaries. Coupled with an L_1 sparsification regime and a specialized class-imbalance pipeline utilizing SMOTE and Asymmetric Focal Loss, the architecture inherently promotes human-readable, symbolic explainability. Evaluated on the benchmark CIC-IDS-2018 dataset, the CNN-BiGRU-KAN significantly outperforms the standard baseline, achieving 94.42% accuracy in binary classification and 96.89% in multi-class categorization, with a multi-class Macro F1 score of 0.9611. Our findings demonstrate that integrating KANs into NIDS pipelines successfully bridges the gap between high-dimensional predictive accuracy and structural transparency.

Keywords: Kolmogorov-Arnold Networks · Network Intrusion Detection · Explainability · Spatiotemporal Analysis.

1 Introduction

The rapid expansion of interconnected digital infrastructure has fundamentally transformed modern computing ecosystems. Emerging paradigms such as the internet of things (IoT), cyber-physical systems, edge computing and cloud-native

architectures have enabled large-scale interconnection across domains including healthcare, smart cities, industrial control systems, and intelligent transportation. While these technologies significantly enhance the convenience, efficiency and accessibility of data, they simultaneously introduce a larger attack surface of modern networks, leading to the development of new sophisticated cyber threats. Cyber adversaries increasingly exploit the heterogeneity, scale, and dynamic nature of distributed computing to launch complex attacks such as distributed denial-of-service (DDoS), botnet propagation, data exfiltration, and advanced persistent threats (APTs) [19,20]. Consequently the development of robust and intelligent cybersecurity mechanisms has become a critical priority for safeguarding digital infrastructures.

Intrusion Detection Systems (IDS) play a central role in network defense by continuously monitoring system activities and network traffic to detect malicious behavior. Traditional intrusion detection techniques are primarily implemented using signature- or rule-based mechanisms that rely on predefined attack patterns, IDS, and expert-defined rules [2]. Despite the effectiveness of these methods in identifying known threats, they exhibit weakness when it comes to adapting for previously unseen attacks or dynamic threats. Furthermore, maintaining a comprehensive signature database becomes increasingly difficult as cyber threats evolve rapidly in scale and complexity. These limitations have led to the direction of research efforts toward the adoption of Machine Learning techniques, which enable network intrusion detection systems to learn patterns directly from data and detect anomalies indicative of attack activities.

Over the past decade, machine learning and deep learning approaches have demonstrated significant improvements in intrusion detection performance. Traditional machine learning models such as support vector machines, Random Forests, and Gradient Boosting Algorithms have been widely used for anomaly-based detection. Recently, deep neural architectures have emerged as promising solutions to network intrusion detection because of their ability to model complex spatial and temporal dependencies in network traffic packets. Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Graph Neural Networks (GNNs), and transformer models have been proposed to perform network traffic analysis and anomaly detection. These models can learn high-dimensional feature representations and achieve high detection accuracy on benchmark datasets [**?**]. However, despite their success, deep-learning IDS models face several critical challenges that hinder their prospective deployment in real-world security systems. Deep-learning IDS often function as "black boxes", providing high precision predictions without the requisite transparency for security researchers, analysts, and practitioners to validate findings or trace the provenance of an alert. This lack of interpretability is a significant barrier in high-stakes environments where a false positive can lead to costly downtime, and a false negative can result in a catastrophic breach.

To address the dual challenges of achieving high predictive accuracy on highly imbalanced network data and maintaining structural transparency, this paper

proposes a novel, explainable hybrid architecture: the CNN-BiGRU-KAN. By integrating one-dimensional Convolutional Neural Networks (CNNs) for spatial feature extraction, Bidirectional Gated Recurrent Units (BiGRU) for temporal sequence modeling, and a Kolmogorov-Arnold Network (KAN) as the classification head, we move beyond the rigid, opaque constraints of standard Multi-Layer Perceptrons (MLPs). The proposed architecture leverages the Kolmogorov-Arnold representation theorem to learn complex, non-linear mappings via univariate spline functions directly on the network edges, thereby intrinsically promoting interpretability and generalization. The main contributions of this work are:

- Novel Hybrid Architecture: We introduce a state-of-the-art NIDS framework that utilizes the grid extension capabilities and adaptive non-linearity of a KAN layer to accurately isolate complex, high-dimensional cyber threats , significantly outperforming traditional CNN-BiGRU baselines.
- Robust Imbalance Remediation: We deploy a highly specialized, multi-stage resampling pipeline, combining dynamically targeted SMOTE and Asymmetric Focal Loss, that effectively addresses the severe class imbalance of the CIC-IDS-2018 dataset, ensuring exceptional recall even for critically rare attack vectors.
- Inherent Explainability: Rather than relying on post-hoc interpretation methods, our model relies on a specialized L_1 sparsification regime , yielding a mathematically auditable, human-readable symbolic representation of the model's decision-making process.

2 Related Works

Recent advances in the field of anomaly detection in network traffic have transitioned toward complex deep learning architectures to address the evolving nature of cyber threats. In 2025, Abdulganiyu et al. proposed XIDINTFL-VAE, a hybrid framework utilizing variational autoencoders and XGBoost to handle class imbalance in high dimensional traffic [3]. Their model boasted a high performance of 99.67% precision and F1 score of 94.74%. The authors emphasized a reduction in false positive rates while maintaining high detection performance. In research conducted in study [4], authors presented a NIDS framework featuring a hybrid deep learning architecture named STL-HDL network. They used an STL ensemble algorithm that included SMOTE and Tomek links to balance the data set. The architecture of their main model was a CNN-LSTM network used for network traffic classification. The model reports an accuracy of 99.83% on CIDDS-001 for multiclass and 99.17% on UNSW-NB15 for binary classification.

Suresh et al. perform a comparative study on the performance evaluation of BiGRU and multi head attention architectures against convolutional neural networks [6]. Their experiments confirm that both models produce exceptional detection performance on the CSE-CIC-IDS 2018 dataset [15] with the BiGRU + MHA model achieving 99.65% accuracy and the CNN achieving 98.85% accuracy on multi-class attack classification. In their research, Rahma et al [10] propose a hybrid CNN-BiGRU for spatiotemporal feature extraction in a network intrusion

detection setting. Their approach employs a genetic algorithm approach to tackle class imbalance. They perform a comparative analysis with the standard CNN-BiGRU and noted significant performance differences with the proposed GA-CNN-BiGRU achieving 98.88% accuracy with the class balancing applied.

A significant advancement in the integration of KANs for cybersecurity is the development of the Convolutional Kolmogorov-Arnold Network (CKAN), as proposed by Abdel-Basset et al. (2024) [1]. This architecture innovates by replacing traditional Multi-Layer Perceptron (MLP) layers with KAN layers within a CNN framework, specifically tailored for IoT intrusion detection. By substituting static weight matrices with learnable univariate functions, the CKAN model achieves state-of-the-art results on benchmark datasets such as NSL-KDD, CICIoT2023, and TON_IoT, reaching accuracies as high as 99.93% for binary classification. Crucially, the authors demonstrate that CKANs maintain superior performance metrics (precision, recall, and F1-score) while utilizing significantly fewer parameters than standard CNNs, RNNs, or Autoencoders. In addition to hybrid architectural innovations, recent literature has explored the use of KANs as a standalone solution for multi-vector threat detection. Mostafa and Hamad introduced KAN-MID, a specialized framework designed for the joint detection of malicious URLs and network intrusions within IoT ecosystems [14]. Their work emphasizes the capacity of KANs to model high-dimensional, non-linear feature spaces, such as those found in lexical URL patterns and diverse IoT protocols, using a fraction of the computational parameters required by traditional MLP-based classifiers. By leveraging the Kolmogorov-Arnold representation theorem to learn complex mapping functions directly on the network edges, the KAN-MID framework demonstrated high robustness against adversarial perturbations commonly used to evade detection in resource-constrained environments. They reported 99.81% accuracy in detecting spam URLs and realized an average accuracy of 98.32% across the CIC-IDS2018 and ICSX-URL2016 datasets.

Authors in [8] propose a hybrid 1D-CNN architecture for explainable network intrusion detection in IoT environments, integrating model-agnostic techniques such as SHAP and LIME to provide post hoc interpretability of model predictions, and achieving an accuracy of 99.95% on the TON-IoT dataset, thereby demonstrating both high predictive performance and enhanced transparency. Addressing the broader challenge of the "black-box" nature of deep learning models, [7] introduce a quantum-inspired KAN framework that incorporates quantum variational activation functions to learn discriminative representations of benign network traffic, improving anomaly detection effectiveness; their proposed architecture, QKAN-AE, advances explainable network intrusion detection systems by promoting both interpretability and generalization, marking a significant step toward transparent and robust AI-driven security solutions.

3 Background

The evolution of Network Intrusion Detection Systems (NIDS) has transitioned from rigid, rule-based heuristics to increasingly complex deep learning architec-

tures. However, as the field targets high-stakes deployment, a fundamental tension has emerged between predictive performance and structural transparency. This section situates our work within the intersections of Kolmogorov-Arnold Networks, hybrid spatial-temporal architectures, and the current landscape of Explainable AI (XAI) in cybersecurity.

3.1 Kolmogorov Arnold Networks

The emergence of Kolmogorov-Arnold Networks has catalyzed a fundamental re-evaluation of neural architecture design, challenging the long-standing dominance of the Multi-Layer Perceptron. By replacing static activation functions at nodes with learnable univariate functions on edges, KANs move beyond the rigid constraints of fixed-weight linear transformations to offer a more flexible, mathematically expressive primitive for deep learning [13,17].

The Kolmogorov-Arnold Representation Theorem: KANs are grounded in the Kolmogorov–Arnold representation theorem, which establishes that any continuous multivariate function can be decomposed into a finite sum of continuous univariate functions composed with addition [11]:

$$f(x_1, ..., x_n) = \sum_{q=1}^{2n+1} \Phi \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right). \tag{1}$$

In this equation, $\phi_{q,p}$ are the univariate functions that map to each input dimension to an intermediate space. Φ represents the outer univariate functions in the representation theorem.

KAN Layer Formulation In a KAN, there is no conventional weight matrix W. Instead, each "weight" is a learnable function ϕ. For a layer with n_{in} inputs and n_{out} outputs, the transformation is defined by a matrix of functions [13,18]:

$$x_{out} = \Phi(x_{in}), \Phi = \{\phi_{i,j}\}. \tag{2}$$

Each activation function $\phi(x)$ is typically parameterized as a combination of a basis function that maintains numerical stability and a **B-spline**:

$$\phi(x) = w \cdot (base(x) + spline(x)). \tag{3}$$

In (3), the spline component is a linear combination of the B-spline basis functions $B_i(x)$ defined as follows:

$$spline(x) = \sum_i c_i B_i(x), \tag{4}$$

where c_i represents the learnable coefficients, G represents the grid size which is a critical hyperparameter when it comes to training KANs. A larger G allows the KAN to provide more flexibility and precision, allowing the spline function to fit more complex, high-frequency and non-conventional functions, lending to the effectiveness of KANs in high-dimensional use cases.

The Grid Extension The grid extension property of KANs refer to their ability to increase the resolution of the underlying splines, effectively moving from a coarse grid without necessitating a complete retraining of the model. Mathematically, if a spline function $f(x)$ is represented on a grid with G_1 intervals, it can be re-parameterized onto a finer grid with G_2 intervals ($G_2 > G_1$) such that the new function $\tilde{f}(x)$ approximates $f(x)$ with high fidelity [21].

The grid extension capability of KANs allow it to offer unique practical implications and benefits over traditional MLPs:

- Hierarchical Optimization: KANs allow for a coarse-to-fine training strategy. A model can be trained on a coarse grid to capture the global landscape and low-frequency components of the target function [21].
- Resolution Adaptability: In the context of anomaly detection, the underlying data distribution may require varying levels of granularity, grid extensions provide a method to scale the model capacity dynamically.
- Mitigation of Spectral Bias: While MLPs often suffer from spectral bias, the grid extension offers KANs a mechanism to increase local frequency capacity, enabling the network to represent complex, non-linear boundaries efficiently.

Interpretability and Sparsification One of the promising features of KANs, particularly for high-stakes domains like NIDS, is their inherent interpretability which offers additional model insights over the post-hoc explanation methods, such as SHAP or LIME typically applied to MLPs [12]. Since KANs learn univariate functions $\Phi_{i,j}$ on edges rather than static weights, the model's decision-making process can be visualized as a set of 1D curves. To move from a dense network of functions to human-readable symbolic representation, the architecture relies on a specialized sparsification regime. During training, an L_1 regularization penalty is applied not to weights, but to the functional magnitude of the activation functions. The total loss function L_{total} is defined as:

$$\ell_{total} = \ell_{data} + \lambda \left(\mu_1 \sum_{l=0}^{L-1} |\Phi_l|_1 + \mu_2 \sum_{l=0}^{L-1} S(\Phi_l) \right), \tag{5}$$

where $|\phi|_1$ represents the functional L_1 norm, calculated as the average absolute value of the function across N_p inputs, denoted as:

$$|\phi|_1 = \frac{1}{N_p} \sum_{s=1}^{N_p} |\phi(x^s)|. \tag{6}$$

3.2 Hybrid CNN Network Intrusion Detection Models

The evolution of NIDS models has shifted from shallow machine learning to deep hybrid architectures. The rationale behind hybridization is the heterogeneous nature of network traffic: packet headers contain spatial and topological information, while flow sequences contain temporal dependencies. A single-model approach often suffers from a feature-representation trade-off, where optimizing for one dimension leads to information loss in the other.

Hybrid Feature Learning Hybrid deep learning models operate on the principle of hierarchical representation learning. In these architectures, the initial layers act as a spatial filter, reducing the dimensionality of raw traffic or flow-based statistics, while subsequent layers behave as a temporal aggregator or integrator. Mathematically, if $X \in \mathbb{R}^{T \times F}$ represents a network flow with T time steps and F features [16], the hybrid model seeks to learn the following mapping:

$$\mathcal{H}(X) = \mathcal{F}_{temporal}(\mathcal{F}_{spatial}(X)). \tag{7}$$

Research indicates that these combined models consistently outperform standalone CNNs or RNNs by reducing both False Positive Rates (FPR) and the impact of vanishing gradients in long-sequence flow analysis [5,9].

CNN-BiGRU Model The Convolutional Neural Network-Bidirectional Gated Recurrent Unit serves as the baseline for this study, representing one of the current state-of-the-art models used in sequential traffic analysis. This hybrid can tackle two challenging aspects of network traffic analysis:

- Spatial Correlation: The CNN component utilizes 1D or 2D convolutional kernels to identify local correlations that exist between adjacent features in a packet header. This stage performs automated feature extraction, transforming the input X into a high-level feature map M:

$$M = \sigma(W_{conv} * X + b), \tag{8}$$

  where $*$ denotes the convolution operation and σ is typically a ReLU or Leaky ReLU activation.
- Bidirectional Temporal Context: Unlike standard GRUs, the BiGRU can process the feature map M in both the forward ($\overrightarrow{h}$) and backward ($\overleftarrow{h}$) directions. This is critical for NIDS because of the intent of an attack often becomes clear when both preceding and succeeding packets in a flow are analyzed.

4 Methodology

This study proposes a hybrid deep learning inspired architecture for network intrusion detection, evaluated on the benchmark CIC-IDS-2018 dataset. The proposed model integrates 1D CNNs, BiGRUs, and KANs into a unified pipeline.

4.1 Dataset and Preprocessing

The dataset used as the experimental benchmark in this research is the CIC-IDS2018 developed by the Canadian Institute for Cybersecurity. The dataset comprises network traffic captures spanning multiple days and encapsulates 14 distinct attack categories alongside benign traffic, resulting in approximately 16 million raw flow records encoded across 80 statistical features. These features include flow duration, inter-arrival time statistics, packet length distributions, flag counts, and flow metric rates derived from bidirectional network flows using the CICFLowMeter tool.

4.2 Class Imbalance Remediation

CIC-IDS-2018 is severely imbalanced: benign traffic constitutes $\sim 75\%$ of samples, while critical attack classes (e.g., SQLi) fall below 0.01%, yielding deceptively high accuracy but negligible minority recall. We address this via a training-only, multi-stage resampling pipeline. For each minority class, a dynamic target is defined as the geometric mean of its size and the majority count, capped by a configurable ceiling to ensure proportional amplification without overinflating moderate classes. SMOTE is applied on the raw feature matrix prior to sequence construction to preserve temporal coherence; to avoid prohibitive k-NN cost, oversampling is restricted to minority subsets, excluding benign data. The neighborhood parameter is set to $\min(5, n_{\min} - 1)$, for stability in rare classes.

Residual imbalance is addressed at the objective level using Asymmetric Focal Loss, with elevated γ for Infiltration and SQLInjection to emphasize hard, minority examples. Class weights derived from the resampled distribution further penalize misclassification of underrepresented classes.

4.3 Proposed Model Architecture

The proposed architecture, depicted in Fig. 1 processes network flow sequences through three sequential components: a CNN block for local feature extraction, a BiGRU network for temporal dependency modelling, and a KAN layer for complex non-linear fusion of the previous two layers. The output of the KAN is then evaluated on binary and multi classification head.

The CNN block consists of two successive one-dimensional convolutional layers with kernel size of 3 and a padding of one, producing 64 and 128 feature maps respectively. Each convolutional layer is followed by batch normalization

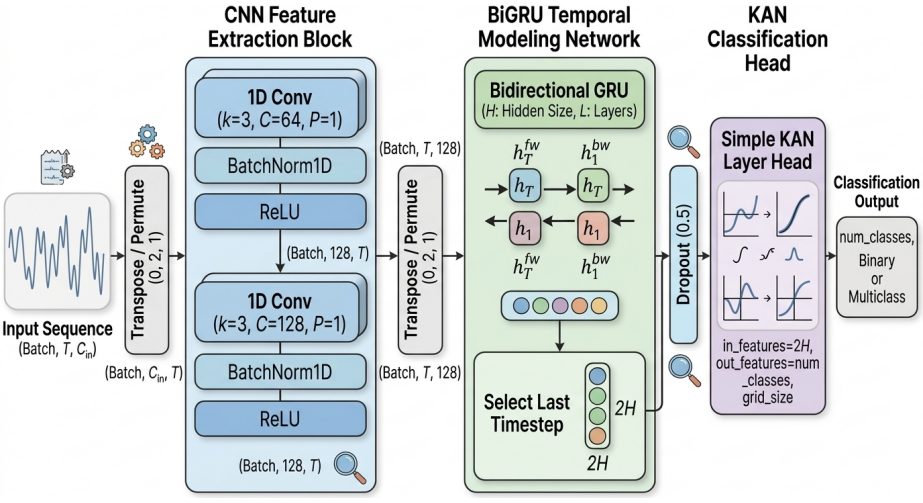

Fig. 1: Proposed CNN-BiGRU-KAN architecture.

and ReLU activation. The convolutional layers operate over the feature dimension of each temporal step, learning local correlations among flow statistics that are nearby in the latent space. Dropout regularization is applied after the second convolutional block to mitigate overfitting in the learned feature representations.

The last layer is an important contribution. After the BiGRU, the model employs a KAN layer, replacing the conventional MLP used in standard architectures. KAN layers learn univariate spline functions on each input dimension rather than fixed linear transformations, endowing the classified with adaptive non-linearity that does not rely on fixed activation functions. The KAN layer maps the output of the BiGRU to the number of classes being considered.

4.4 Training Procedure

The model was trained using the Adam optimizer with an initial learning rate of 10^{-3}, selected in accordance with the linear scaling rule given a batch size of 8192. A linear warmup schedule was applied for the first five epochs, ramping the learning rate from 10% of its target value to the full value to allow batch normalization to stabilize before large gradient updates are applied. Following warmup, a ReduceLROnPlateau scheduler halved the learning rate when validation loss failed to improve for seven consecutive epochs, with a minimum learning rate floor of 10^{-6}. Gradient norms were clipped to a maximum value of 5.0 at each update step to prevent gradient explosion in the recurrent component. Training was performed on 2 NVIDIA A5000 GPUs with model parallelism.

5 Results

To rigorously evaluate the efficacy of the proposed CNN-BiGRU-KAN architecture, we conducted extensive comparative experiments against a state-of-the-art hybrid baseline, the CNN-BiGRU. The models were evaluated on the benchmark CIC-IDS-2018 dataset, assessing both binary anomaly detection and granular multi-class threat categorization.

5.1 Binary Classification Performance

The primary objective of the binary classification task is to distinguish malicious network intrusions from benign traffic. As depicted in Fig. 2 and detailed in Table 1, the integration of the KAN layer yields a substantial performance improvement over the traditional MLP classification head used in the baseline.

The proposed CNN-BiGRU-KAN architecture achieved an overall accuracy of 0.9442, representing an 8.92% absolute improvement over the CNN-BiGRU baseline (0.855). Most notably, the precision of the proposed model surged to 0.9618 compared to the baseline's 0.7064. This drastic reduction in false positive rates is critical for real-world deployment, where alert fatigue remains a primary challenge for Security Operations Centers. While recall saw a negligible decrease from 0.8425 to 0.8378, the Macro F1 score, which provides a harmonic mean of

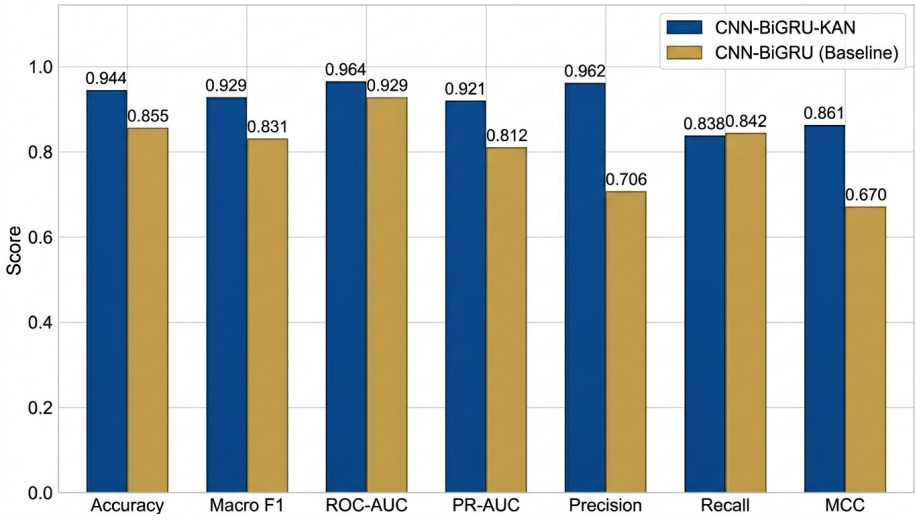

Fig. 2: Binary Classification Performance Metrics.

Table 1: Evaluation Metrics on CIC-IDS 2018 Binary-Classification

Model Type	Accuracy	Precision	Recall	Macro F1 Score
CNN-BiGRU	0.8550	0.7064	0.8425	0.8315
CNN-BiGRU-KAN	0.9442	0.9618	0.8378	0.9287

precision and recall robust to class imbalance, increased significantly from 0.8315 to 0.9287. These findings suggest that the learnable univariate functions on the KAN edges successfully map the complex, high-dimensional spatiotemporal features extracted by the CNN-BiGRU block into a more separable space.

5.2 Multi-Class Threat Categorization

Evaluating the architecture on the multi-class task provides deeper insights into its capacity to isolate specific attack vectors (e.g., Botnet, BruteForce, DDoS, DoS, Infiltration, and SQLInjection) amidst highly imbalanced data distributions. As shown in Table 2, the proposed model demonstrates overwhelming superiority in the multi-class paradigm. The CNN-BiGRU-KAN achieved an accuracy of 0.9689, significantly outperforming the baseline's 0.7883. The limitations of the traditional CNN-BiGRU become glaringly apparent in its recall (0.7742), indicating an inability to accurately retrieve minority attack classes. Conversely, the proposed model achieved a recall of 0.9651 and a near-perfect precision of 0.9831, resulting in a Macro F1 score of 0.9611 (compared to the baseline's 0.8621).

This performance delta is further corroborated by the normalized confusion matrices, see Fig. 3 and Fig. 4. The baseline matrix reveals significant inter-class

Table 2: Evaluation Metrics on CIC-IDS 2018 Multi-Classification

Model Type	Accuracy	Precision	Recall	Macro F1 Score
CNN-BiGRU	0.7883	0.9321	0.7742	0.8621
CNN-BiGRU-KAN	0.9689	0.9831	0.9651	0.9611

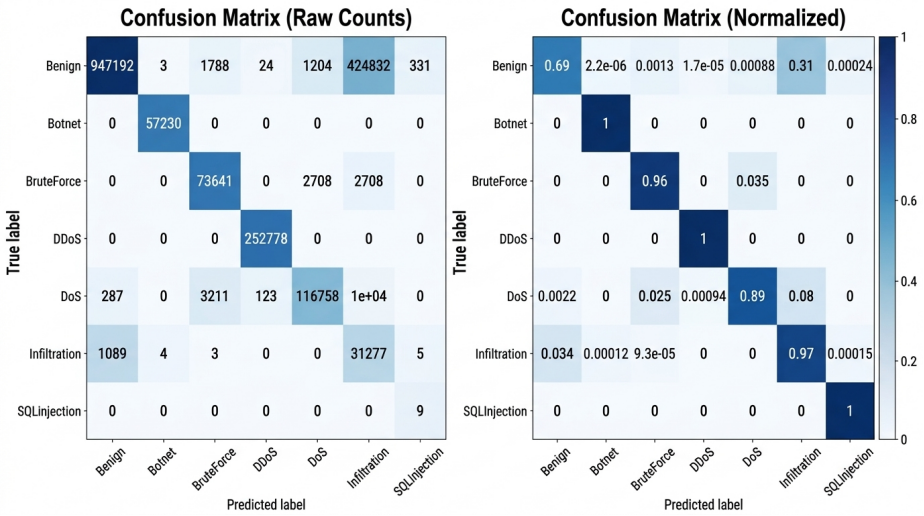

Fig. 3: Baseline CNN-BiGRU Multi-Classification Confusion Matrix.

confusion, particularly misclassifying instances of DoS and Infiltration attacks as benign traffic. In contrast, the confusion matrix for the CNN-BiGRU-KAN exhibits rigid diagonal dominance. The grid extension and dynamic resolution capabilities of the KAN layer appear to heavily mitigate the spectral bias typically suffered by standard MLPs, allowing the network to establish highly non-linear decision boundaries that successfully isolate even sophisticated, low-volume infiltration attempts.

6 Discussion

The empirical evaluation of the CNN-BiGRU-KAN architecture demonstrates a definitive performance advantage over the conventional CNN-BiGRU baseline, particularly in domains traditionally hostile to deep learning models: high false positive rates and severe class imbalance. The integration of a KAN layer as the classification head fundamentally alters how the model synthesizes spatiotemporal features, bridging the gap between high-dimensional feature extraction and highly non-linear decision boundaries.

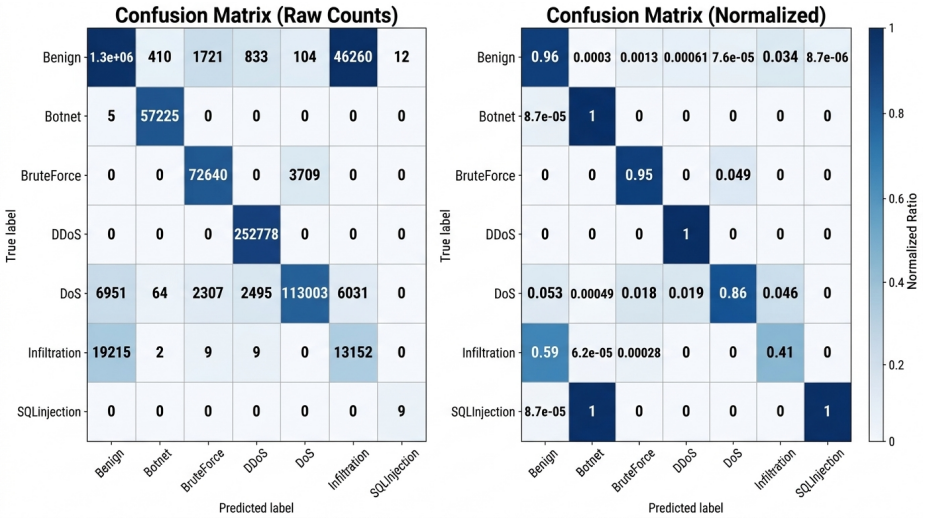

Fig. 4: Proposed Method Multi-Classification Confusion Matrix.

6.1 Adaptive Non-Linearity and the Reduction of False Positives

A critical bottleneck in the operational deployment of anomaly-based NIDS is the high incidence of false positives, which induces alert fatigue among security analysts and can lead to costly operational downtime. In the binary classification task, the CNN-BiGRU-KAN achieved a remarkable precision of 0.9618, an absolute improvement of over 25% compared to the baseline's 0.7064.

We attribute this surge in precision to the fundamental architecture of the KAN layer. Traditional Multi-Layer Perceptrons (MLPs) apply fixed linear transformations followed by static activation functions, which often struggle to establish tight, complex decision boundaries around benign traffic clusters. By operating on the Kolmogorov-Arnold representation theorem, KANs replace these static weight matrices with learnable univariate spline functions on the network edges. This allows the network to dynamically adapt its functional capacity. The resulting architecture does not simply draw hyperplanes through the BiGRU's temporal feature space; rather, it shapes bespoke, highly non-linear manifolds that tightly encompass benign behavior, strictly isolating anomalous vectors and drastically reducing false-positive misclassifications.

6.2 Mitigating Spectral Bias in Multi-Class Threat Detection

The superiority of the CNN-BiGRU-KAN is most pronounced in the multi-class classification paradigm. The baseline model exhibited a severe degradation in recall (0.7742), indicating an inability to reliably identify minority attack classes buried within the imbalanced CIC-IDS-2018 dataset. Conversely, the proposed architecture maintained a recall of 0.9651 and an accuracy of 0.9689. This

performance delta highlights a known limitation of standard deep neural networks: spectral bias. MLPs are notoriously biased toward learning low-frequency functions, making them prone to ignoring the high-frequency, nuanced feature variations that characterize sophisticated, low-volume attacks. KANs, however, possess a unique grid extension property. By enabling the parameterization of underlying splines onto finer grids during optimization, the model dynamically scales its resolution. This explicit mechanism allows the CNN-BiGRU-KAN to increase local frequency capacity precisely where the feature space is most convoluted, enabling the capture of complex, non-linear boundaries. When combined with our asymmetric focal loss and SMOTE pipeline, the KAN layer successfully isolates transient temporal anomalies that the baseline simply averages out.

6.3 Towards Inherent Explainability in NIDS

The proposed CNN-BiGRU-KAN architecture inherently advances the interpretability of network traffic analysis. Because the KAN learns univariate functions $\phi_{i,j}$ directly on the edges, the model's decision-making process can be mathematically visualized as a set of 1D curves rather than opaque weight matrices. Furthermore, the application of an L_1 regularization penalty to the functional magnitude of the activation functions drives the network toward a sparsification regime. This forces the architecture to prune unnecessary functional pathways, distilling the complex spatiotemporal mapping learned by the CNN and BiGRU blocks into a more transparent, human-readable symbolic representation. Consequently, security practitioners are not just provided with an alert; they are provided with a structurally auditable pathway detailing how the spatiotemporal features of a packet flow triggered a specific malicious classification.

7 Conclusion

In this paper, we proposed a novel, explainable hybrid deep learning architecture, the CNN-BiGRU-KAN to address the escalating complexity and frequency of cyber threats in interconnected computing ecosystems. While traditional deep neural networks have demonstrated proficiency in anomaly detection, their reliance on opaque MLP layers has historically forced a compromise between predictive accuracy and model transparency. Our framework resolves this tension by replacing the static linear transformations of standard MLPs with a KAN layer, which utilizes learnable univariate spline functions on network edges to dynamically map high-dimensional spatiotemporal features.

Our empirical evaluation on the highly imbalanced CIC-IDS-2018 benchmark dataset demonstrates that the proposed architecture achieves remarkable performance improvements over state-of-the-art hybrid baselines. The CNN-BiGRU-KAN framework reached an accuracy of 94.42% in binary classification and 96.89% in multi-class categorization, exhibiting expressive improvements in precision, recall, and Macro F1 scores. Crucially, the grid extension capabilities and L_1 sparsification regime inherent to the KAN layer not only mitigated the

spectral bias that causes high false-positive rates in conventional models but also provided a mathematically interpretable, human-readable structural pathway for alert provenance.

Acknowledgments. Part of this work was funded by the National Science Foundation under grant CNS-2136961, and the Department of Education under grant P116Z230151. The authors thank the Rivas.AI Lab (`https://lab.rivas.ai`) for the support and helpful feedback throughout this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abd Elaziz, M., Fares, I.A., Aseeri, A.O.: Ckan: Convolutional kolmogorov–arnold networks model for intrusion detection in iot environment. IEEE Access **12**, 134837–134851 (2024). `https://doi.org/10.1109/access.2024.3462297`
2. Abdulganiyu, O.H., Ait Tchakoucht, T., Saheed, Y.K.: A systematic literature review for network intrusion detection system (ids). International journal of information security **22**(5), 1125–1162 (2023). `https://doi.org/10.1007/s10207-023-00682-2`
3. Abdulganiyu, O.H., Tchakoucht, T.A., Saheed, Y.K., Ahmed, H.A.: Xidintfl-vae: Xgboost-based intrusion detection of imbalance network traffic via class-wise focal loss variational autoencoder. The Journal of Supercomputing **81**(1), 16 (2025). `https://doi.org/10.1007/s11227-024-06552-5`
4. Al, S., Dener, M.: Stl-hdl: A new hybrid network intrusion detection system for imbalanced dataset on big data environment. Computers & Security **110**, 102435 (2021). `https://doi.org/10.1016/j.cose.2021.102435`
5. Alqahtani, H., Abdullah, M.: A taxonomy of ids in iots: Ml classifiers, feature selection models, datasets and future directions. International Journal of Advanced Computer Science and Applications **15**(6) (2024). `https://doi.org/10.14569/ijacsa.2024.0150682`
6. Balasubramanian, S.K., Perumal, S.: Comparative study of bigru with multi-head attention and cnn for network intrusion detection using a cleaned and balanced cse-cic-ids 2018 dataset. Turkish Journal of Engineering **9**(4), 725–737 (2025). `https://doi.org/10.31127/tuje.1695208`
7. Do, P.H.: Beyond the black box: A generalizable and interpretable approach to network intrusion detection using quantum-inspired kans. SSRN Electronic Journal (2025). `https://doi.org/10.2139/ssrn.5935984`
8. Ebrahimi, F., Javidan, R., Akbari, R., Hosseini, Y.: Intrusion detection in the internet of things using convolutional neural networks: an explainable ai approach. Cybersecurity **8**(1), 66 (2025). `https://doi.org/10.1186/s42400-025-00369-2`
9. Ghosh, K.P., Hasan, M., Robin, M.T.I., Hossain, M.A., Islam, M.S.: A novel deep learning framework with temporal attention convolutional networks for intrusion detection in iot and iiot networks. Scientific Reports **15**(1) (2025). `https://doi.org/10.1038/s41598-025-32697-1`
10. Jablaoui, R., Liouane, N.: Ga-cnn-bigru-ids: A robust framework for intrusion detection system based on ga for data augmentation and hybrid cnn-bigru model for spatiotemporal feature extraction. Computers & Electrical Engineering **130**, 110900 (2026). `https://doi.org/10.1016/j.compeleceng.2025.110900`

11. Kolmogorov, A.N.: On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. American Mathematical Society Translations **28**, 55–59 (1963). https://doi.org/10.1090/trans2/028/04
12. Liu, Z., Ma, P., Wang, Y., Matusik, W., Tegmark, M.: Kan 2.0: Kolmogorov-arnold networks meet science. arXiv preprint arXiv:2408.10205 (2024). https://doi.org/10.48550/arXiv.2408.10205
13. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024). https://doi.org/10.48550/arXiv.2404.19756
14. Rizk, F., Rizk, R., Rizk, D., Rizk, P., Chu, C.H.H.: Kan-mid: A kolmogorov-arnold networks-based framework for malicious url and intrusion detection in iot systems. IEEE Access **13**, 160855–160873 (2025). https://doi.org/10.1109/access.2025.3605171
15. Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A., et al.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP). pp. 108–116. SCITEPRESS - Science and Technology Publications (2018). https://doi.org/10.5220/0006639801080116, https://doi.org/10.5220/0006639801080116
16. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems 28. vol. 28 (2015). https://doi.org/10.48550/arXiv.1506.04214, https://doi.org/10.48550/arXiv.1506.04214
17. Somvanshi, S., Javed, S.A., Islam, M.M., Pandit, D., Das, S.: A survey on kolmogorov-arnold network. ACM Computing Surveys **58**(2), 1–35 (2025). https://doi.org/10.1145/3743128
18. Ta, H.T., Tran, A.: Af-kan: Activation function-based kolmogorov-arnold networks for efficient representation learning. arXiv preprint arXiv:2503.06112 (2025). https://doi.org/10.48550/arXiv.2503.06112
19. Vanin, P., Newe, T., Dhirani, L.L., O'Connell, E., O'Shea, D., Lee, B., Rao, M.: A study of network intrusion detection systems using artificial intelligence/machine learning. Applied Sciences **12**(22), 11752 (2022). https://doi.org/10.3390/app122211752
20. Walling, S., Lodh, S.: An extensive review of machine learning and deep learning techniques on network intrusion detection for iot. Transactions on Emerging Telecommunications Technologies **36**(2), e70064 (2025). https://doi.org/10.1002/ett.70064
21. Wang, Y., Siegel, J.W., Liu, Z., Hou, T.Y.: On the expressiveness and spectral bias of kans. arXiv preprint arXiv:2410.01803 (2024). https://doi.org/10.48550/arXiv.2410.01803