Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

e.INS
Ecosystem of Innovation for Next Generation Sardinia

Project funded under the NATIONAL RECOVERY AND RESILIENCE PLAN (PNRR) - MISSION 4 COMPONENT 2, "From research to business" INVESTMENT 1.5, "Creation and strengthening of Ecosystems of innovation" and construction of "Territorial R&D Leaders"

Enhancing IDS with Ensemble LSTM Networks Using Real and GAN Data

Roberto Saia, Salvatore Carta, Gianni Fenu, Alessandro Sebastian Podda, and Livio Pompianu

Department of Mathematics and Computer Science
University of Cagliari, Italy

AIBD LAB
Artificial Intelligence and Big Data Laboratory
https://aibd.unica.it/

ITNAC 2024, 34th IEEE International Telecommunication Networks and Applications Conference
27÷29 November 2024 – Sydney, Australia

Outline

- **Overview**
  Introduction ▪ State of the Art
- **Intuition**
  Idea ▪ Architecture
- **Formalization**
  Data Preprocessing ▪ GAN Generation ▪ Classification
- **Results**
  Experiments ▪ Conclusions and Future Directions

OVERVIEW

Introduction

- Today's computer **system security** is critical at every operational level and device, as the compromise of a single element can **propagate** through other connected network elements, causing unpredictable and dangerous effects.

- To **face unauthorized access** and evolving malicious strategies, researchers have intensified efforts to develop effective **Intrusion Detection Systems** (IDSs) that monitor and analyze network traffic to detect illegitimate activities.

Introduction

- A **significant challenge** for IDSs is the volume and complexity of the involved data since modern networks generate vast amounts of traffic data.
- Another **crucial challenge** is the high dimensionality of network data then often reduces the algorithm effectiveness.
- In addition, the **dynamic nature** of cyber threats continually evolves as attackers experiment with new methods to overcome existing security measures, rendering static detection models less effective.

State of the Art

- **Multiple Techniques and Strategies**: Various methods have been designed to enhance IDS performance, utilizing a range of advanced techniques to address security challenges.
- **Adapting to Evolving Threats**: Due to the rapidly evolving nature of cyber-attacks, existing systems must be continuously updated and tuned.
- **Ensuring Effective Defense**: Regular improvements are essential to maintain high performance and ensure robust protection against emerging threats.

State of the Art

- In this context, we propose the **HYDRA-LNNE** (*Hybrid Data Real and Artificial LSTM Neural Network Ensemble*) approach (hereafter referred to as **HYDRA**).

- It **enhances** intrusion detection by combining **feature selection** and **quantization**, followed by an **ensemble** of Long Short-Term Memory (**LSTM**) neural networks trained on both **real** and **synthetic** data.

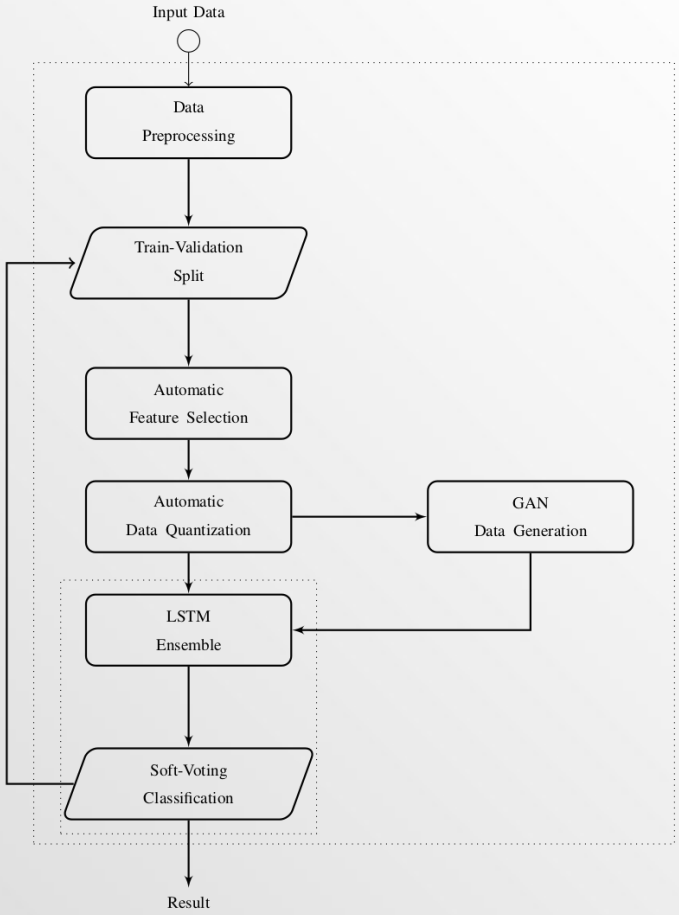
INTUITION

Idea

- Identifying **relevant features** for intrusion detection.
- Converting continuous data into discrete categories, **reducing dimensionality** and **computational complexity**.
- Training three **LSTM networks** on **real data**, GAN-generated **synthetic data**, and a **mix of both**.
- Validate the approach using the **NSW-NB15 dataset** that offers an extensive and different representation of modern network traffic and cyber threats.

Architecture

High-level architecture of the **HYDRA** approach:



FORMALIZATION

Data Preprocessing

Data Cleaning:
The dataset is processed to handle missing and anomalous values, convert numerical fields, replace missing values using the most frequent criterion, and clip excessively large values.

Feature Selection:
The Boruta technique is applied to identify the most relevant features by comparing the importance of each original feature with the maximum importance of its shadow features.

Data Quantization:
Continuous numerical features are transformed into discrete categorical values through quantization, where the optimal number of subdivisions is determined automatically.

Data Generation with GAN

Method: Synthetic data is generated using a Generative Adversarial Network (GAN) that involves iterative training of a generator and a discriminator .

Objective: The goal is to produce synthetic samples that reflect the statistical characteristics of real data.

- **Objective Function**:

$$\min_G \max_D V(D, G)$$

- **Value Function**:

$$V(D, G) = E_{x \sim p_{\text{data}}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

Data Classification

Method: Network events classification utilizes an ensemble of Artificial Neural Networks (ANNs) trained on both real and synthetic data.

Data Handling: A Train-Validation-Split criterion ensures no data leakage between training and testing datasets.

Model Training: Three LSTM models are trained using real data, synthetic data, and a combination of both. An early stopping mechanism is implemented to prevent overfitting.

Prediction Aggregation: Predictions are aggregated using the *Soft Voting* criterion, resulting in the final classification output:

$$\hat{y} = \arg\max_j \left(\frac{1}{N} \sum_{i=1}^{N} P_{ij} \right)$$

where $\hat{y}$ is the predicted class, P_{ij} represents the predicted probability of the i-th instance being in class j, and N is the number of models in the ensemble.

HYDRA Algorithm

The aforementioned steps lead to the following **algorithm**:

Input: Raw Dataset $\mathcal{D}$
Output: Predicted class labels for $\mathcal{D}$
$\mathcal{D}_{\text{pre}} \leftarrow \text{Preprocess}(\mathcal{D})$
$\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}} \leftarrow \text{Split}(\mathcal{D}_{\text{pre}})$
$\mathcal{F} \leftarrow \text{SelectFeatures}(\mathcal{D}_{\text{train}})$
$\mathcal{Q} \leftarrow \text{Quantize}(\mathcal{D}_{\text{train}})$
$\mathcal{G} \leftarrow \text{GenerateGAN}(\mathcal{D}_{\text{train}})$
$\mathcal{M}_{\text{real}} \leftarrow \text{TrainLSTM}(\mathcal{D}_{\text{train}})$
$\mathcal{M}_{\text{gan}} \leftarrow \text{TrainLSTM}(\mathcal{G})$
$\mathcal{M}_{\text{combined}} \leftarrow \text{TrainLSTM}(\mathcal{D}_{\text{train}}, \mathcal{G})$
$\hat{y}_{\text{real}} \leftarrow \text{Predict}(\mathcal{M}_{\text{real}}, \mathcal{D}_{\text{test}})$
$\hat{y}_{\text{gan}} \leftarrow \text{Predict}(\mathcal{M}_{\text{gan}}, \mathcal{D}_{\text{test}})$
$\hat{y}_{\text{combined}} \leftarrow \text{Predict}(\mathcal{M}_{\text{combined}}, \mathcal{D}_{\text{test}})$
$\hat{y}_{\text{final}} \leftarrow \text{SoftVote}(\hat{y}_{\text{real}}, \hat{y}_{\text{gan}}, \hat{y}_{\text{combined}})$
return $\hat{y}_{\text{final}}$

RESULTS

Experimental Strategy

1) **Comparison with Recent Works**: Validated the HYDRA approach by comparing it with recent studies using the same UNSW-NB15 training and test sets.

2) **Train-Validation-Split**: Implemented a stratified shuffle split to maintain class distribution in training and validation sets. This ensures model generalization and prevents overfitting, allowing for effective evaluation and hyperparameter tuning without contaminating the validation process with test data.

3) **Automatic Optimization**: Automated key procedures such as feature selection (Boruta), data quantization, and GAN-based synthetic data generation to optimize the HYDRA approach. This identifies optimal parameters on the first run and saves process data for future use.

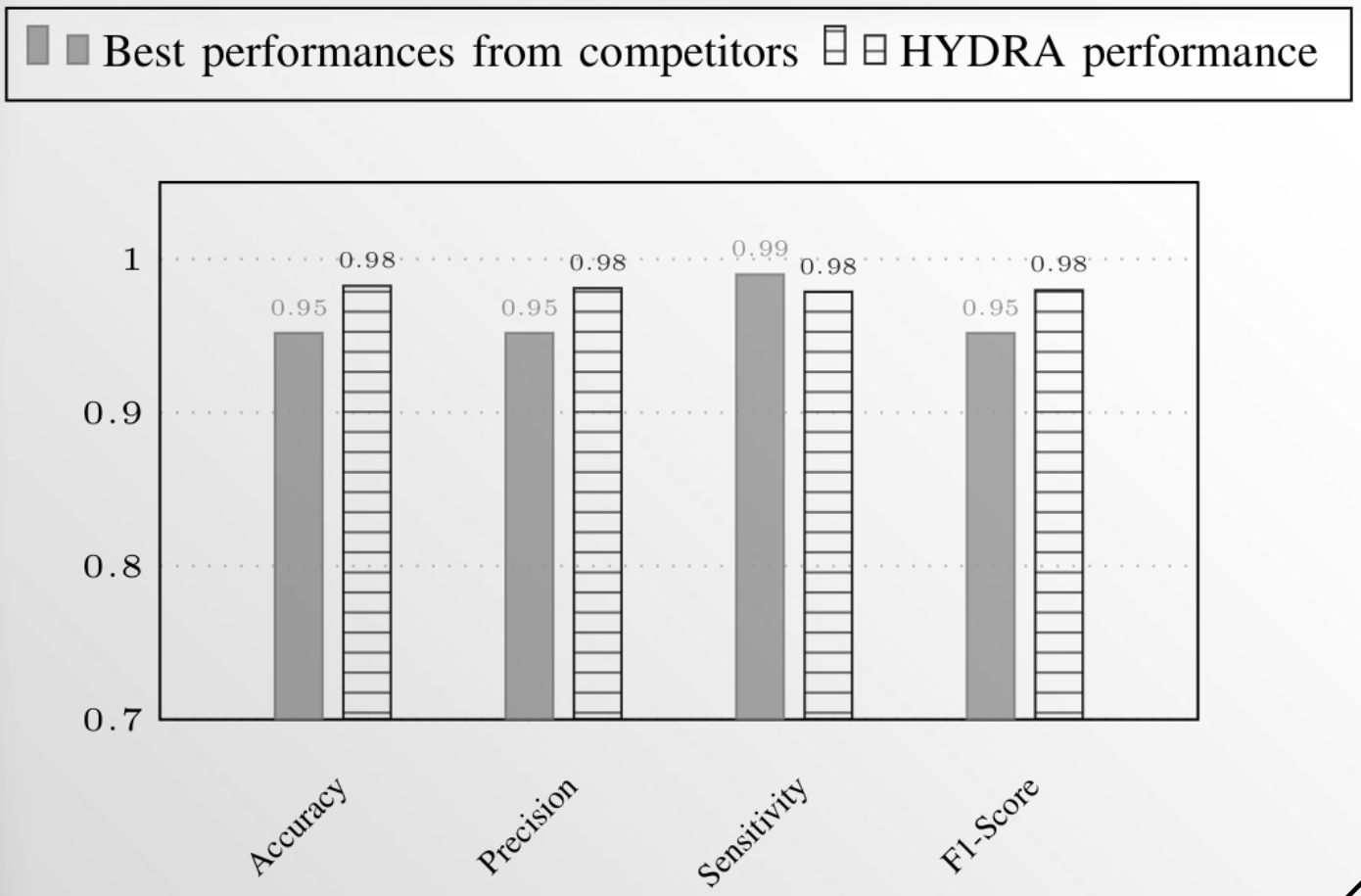
4) **Evaluation Metrics**: Assessed model performance using various metrics: **Accuracy**, **Precision**, **Sensitivity**, **Fallout**, **F1-Score**, Area Under the Curve (**AUC**), Mean Absolute Error (**MAE**), Relative Absolute Error (**RAE**), Root Mean Squared Error (**RMSE**), and **Kappa Statistic**.

Experimental Results

Ref.	Approach	Accuracy	Precision	Sensitivity	Fallout	F1-Score	AUC	MAE	RAE	RMSE	Kappa
[32]	LR	0.8964	0.9031	0.8964	NC	0.8930	NC	0.1036	0.2246	0.3219	0.7641
	KNN	0.9171	0.9172	0.9171	NC	0.9171	NC	0.0829	0.1798	0.2880	0.8205
	RF	0.9518	0.9518	0.9518	NC	0.9518	NC	0.0482	0.1045	0.2196	0.8956
	DT	0.9370	0.9370	0.9370	NC	0.9370	NC	0.0630	0.1366	0.2511	0.8634
	AdaBoost	0.9229	0.9230	0.9229	NC	0.9230	NC	0.0771	0.1670	0.2776	0.8330
	Bagging	0.9471	0.9470	0.9471	NC	0.9470	NC	0.0529	0.1146	0.2299	0.8850
	J48	0.9388	0.9388	0.9388	NC	0.9388	NC	0.0612	0.1326	0.2474	0.8674
[33]	ANN	0.8862	0.8491	0.9647	NC	0.8647	NC	NC	NC	NC	NC
	DT	0.8727	0.8244	0.9768	NC	0.8941	NC	NC	NC	NC	NC
	RF	0.8740	0.8209	0.9863	NC	0.8960	NC	NC	NC	NC	NC
	SVM	0.8404	0.7903	0.9666	NC	0.8696	NC	NC	NC	NC	NC
	LR	0.8184	0.7585	0.9831	NC	0.8563	NC	NC	NC	NC	NC
	KNN	0.8711	0.8570	0.9192	NC	0.8870	NC	NC	NC	NC	NC
[34]	DT	0.9276	0.9096	0.9642	0.1172	0.9361	NC	NC	NC	NC	NC
	RF	0.8616	0.8040	**0.9900**	0.2957	0.8873	NC	NC	NC	NC	NC
	GBT	0.8803	0.8355	0.9743	0.2348	0.8996	NC	NC	NC	NC	NC
	MLP	0.8491	0.7770	0.9676	0.3230	0.8545	NC	NC	NC	NC	NC
[35]	ANN	0.8439	0.7856	0.9853	NC	0.8742	NC	NC	NC	NC	NC
	LR	0.7764	0.7318	0.9374	NC	0.8220	NC	NC	NC	NC	NC
	KNN	0.8446	0.8031	0.9509	NC	0.8708	NC	NC	NC	NC	NC
	SVM	0.6089	0.5889	0.9588	NC	0.7297	NC	NC	NC	NC	NC
	DT	0.9085	0.8033	0.9838	NC	0.8845	NC	NC	NC	NC	NC
	HYDRA	**0.9825**	**0.9810**	0.9787	**0.0213**	**0.9798**	**0.9988**	**0.0175**	**0.0402**	**0.1323**	**0.9597**

The column labeled '**Ref.**' refers to the papers containing the state-of-the-art approaches with which we compared our work.

Results Overview



Conclusions

- We presented **HYDRA**, a novel Hybrid Data Real and Artificial LSTM Neural Network Ensemble approach designed for intrusion detection in network security.
- It leverages **advanced machine learning techniques** to improve IDS performance by integrating real and synthetic data.
- The validation results demonstrate that HYDRA **outperforms** both traditional and complex algorithms in various key metrics.
- As **future research directions**, we aim to explore scalability for larger datasets and investigate adaptive learning techniques to enhance the model's capability against emerging threats.

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

UNICA UNIVERSITÀ DEGLI STUDI DI CAGLIARI

e.INS
Ecosystem of Innovation for Next Generation Sardinia

Project funded under the NATIONAL RECOVERY AND RESILIENCE PLAN (PNRR) - MISSION 4 COMPONENT 2, "From research to business" INVESTMENT 1.5, "Creation and strengthening of Ecosystems of innovation" and construction of "Territorial R&D Leaders"

Enhancing IDS with Ensemble LSTM Networks Using Real and GAN Data

THANK YOU FOR YOUR ATTENTION

AIBD LAB
Artificial Intelligence and Big Data Laboratory
https://aibd.unica.it/

ITNAC 2024, 34th IEEE International Telecommunication Networks and Applications Conference
27÷29 November 2024 – Sydney, Australia