CVR COLLEGE OF ENGINEERING

Department of CSE(Data Science)
B.Tech CSE(DS) IV Year I Semester

Major Project Stage-1
Review-2, Aug.23 2025

- PRC Members
  - Dr. A. Srinivasa Reddy
  - Dr. M. Sreenu
  - Dr. Shaik Janbhasha
  - Dr. Afreen Fatima Mohammed

Explainable Intrusion Detection Using SHAP, Tree Based Ensemble Methods, and Agentic AI

22B81A6773	GOLLA DIKSHIT
22B81A67A3	BATHINI SAMUEL PRASHANTH
22B81A67B3	VULAPU SRINATH

Under the guidance of

Dr. B. Ramakrishna

Associate Professor

Domain

Domain: Artificial Intelligence and Cybersecurity

Introduction:

Artificial Intelligence (AI) is revolutionizing cybersecurity by enabling faster, smarter, and more adaptive threat detection. Traditional rule-based systems often fail to detect novel attacks, whereas machine learning models can identify complex patterns in network behavior. However, these models typically function as “black boxes,” providing little insight into their decision-making process.

This project introduces a unified approach that integrates Explainable AI (XAI) using SHAP (Shapley Additive Explanations) for model transparency and LLM(Large Language Model)-based agents for intelligent reasoning. The result is a system that not only detects threats accurately but also explains its decisions in a human-understandable manner.

Significance:

An Explainable and Agentic Intrusion Detection System (IDS) enhances both trust and clarity in cybersecurity operations. It enables security teams to understand the rationale behind each alert and take appropriate, informed action swiftly.

By combining interpretability with AI-driven automation, this system bridges the gap between human analysts and intelligent models, fostering a more transparent, adaptive, and resilient cybersecurity framework.

Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP

- **Authors & Publication:** Anonymous Authors. (2025). *Preprints.org*, 202503.1902.
- **URL:** https://www.preprints.org/manuscript/202503.1902/v1
- **Contribution:**
  Introduces a privacy-preserving and explainable Intrusion Detection System (IDS) that integrates **Federated Learning (FL)** and **SHAP** to collaboratively detect intrusions across multiple network nodes without sharing sensitive data.
- **Methodology:**
  Uses a **federated training setup** where local IDS models learn from distributed network data without sharing raw information. The central server aggregates model parameters to build a global model, and **SHAP** interprets predictions by identifying key features influencing each intrusion alert.
- **Limitations:**
  The study primarily focuses on demonstrating the feasibility of combining FL and SHAP for IDS and does not extensively explore scalability across very large networks or heterogeneous data sources. Additionally, communication overhead and synchronization delays between clients remain potential challenges for real-world deployment.

A Unified Approach for Interpreting Model Predictions

- **Authors & Publication:** Lundberg, S. M., & Lee, S. I. (2017). *Advances in Neural Information Processing Systems*, 30.
- **URL:** Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions.* Advances in Neural Information Processing Systems, 30.
- **Contribution:**
  Proposes SHAP (SHapley Additive exPlanations), a unified framework to explain predictions from any machine learning model by assigning a contribution value to each feature.
- **Methodology:**
  Uses Shapley values from game theory to fairly distribute the model's prediction among the features. This is an additive feature attribution method.
- **Limitations:**
  Exact calculation of Shapley values can be computationally expensive, especially for complex or high-dimensional models.

Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method

- **Authors & Publication:** Le, Thi-Thu-Huong, et al. (2022). *Sensors*, 22(3), 1154.
- **URL:** Le, T.-T.-H., et al. (2022). *Classification and explanation for intrusion detection system based on ensemble trees and SHAP method.* Sensors, 22(3), 1154.
- **Contribution:**
  Proposes a highly accurate and explainable Intrusion Detection System (IDS) for network traffic by combining the predictive power of ensemble tree models with the interpretability of the SHAP method.
- **Methodology:**
  Utilizes ensemble tree models (like Gradient Boosting or Random Forest) for classifying network traffic as normal or malicious. SHAP values are employed to interpret predictions, identifying which network features (e.g., duration, protocol, port) most significantly drive an attack classification.
- **Limitations:**
  Performance can be highly dependent on the quality and balance of the dataset used (e.g., NSL-KDD). Real-time application may face challenges due to the computational overhead of generating SHAP explanations for every prediction.

A Detailed Analysis of the KDD CUP 99 Data Set

- **Authors & Publication:** Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications* (pp. 1–6). IEEE.
- **URL:** Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). *A detailed analysis of the KDD Cup 99 data set.* In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (pp. 1–6). IEEE.
- **Contribution:**
  Provides a detailed analysis and critique of the KDD Cup 99 dataset, a benchmark used for intrusion detection systems. Highlights significant issues such as redundant records and imbalanced distributions, which affect IDS evaluation.
- **Methodology:**
  Performs a statistical and analytical review of the KDD Cup 99 dataset's training and testing partitions. Quantifies redundant records and analyzes sampling and distribution biases to show why using KDD Cup 99 alone can lead to inaccurate performance evaluations for intrusion detection models.
- **Limitations:**
  Focuses solely on analyzing dataset issues and does not propose a new classification model. Although the NSL-KDD dataset addresses some problems, it still retains minor limitations of the original dataset.

An Explainable and Optimized Network Intrusion Detection Model using Deep Learning

- **Authors & Publication:** MP, P. J. (2024). *International Journal of Advanced Computer Science & Applications*, 15(1).
- **URL:** MP, P. J. (2024). *An Explainable and Optimized Network Intrusion Detection Model using Deep Learning*. International Journal of Advanced Computer Science & Applications, 15(1).
- **Contribution:**
  Proposes a Network Intrusion Detection Model (NIDM) that emphasizes both optimization for performance and explainability in network security, leveraging Deep Learning techniques.
- **Methodology:**
  Employs Deep Learning as the core classification model for intrusion detection and integrates explainability mechanisms to interpret the model's predictions—enhancing transparency and trust in detecting security threats.
- **Limitations:**
  Deep Learning models require large datasets and high computational power for training. Moreover, explaining complex deep learning models is more challenging and resource-intensive compared to tree-based approaches like Random Forest.

Explainable Artificial Intelligence for Intrusion Detection System

- **Authors & Publication:** Patil, S., Varadarajan, V., Mazhar, S. M., Shahzada, A., Ahmed, N., Sinha, O., & Kotecha, K. (2022). *Electronics*, 11(19), 3079.
- **URL:** Patil, S., Varadarajan, V., Mazhar, S. M., Shahzada, A., Ahmed, N., Sinha, O., & Kotecha, K. (2022). *Explainable artificial intelligence for intrusion detection system.* Electronics, 11(19), 3079.
- **Contribution:**
  Proposes a framework for an Explainable Artificial Intelligence (XAI)-based Intrusion Detection System (IDS) aimed at enhancing the trustworthiness and usability of IDSs by making their classification decisions transparent to security analysts.
- **Methodology:**
  Incorporates XAI techniques (e.g., LIME or SHAP) with a classification model to generate post-hoc explanations for predictions. This enables analysts to identify which network features influence the detection of specific attack types.
- **Limitations:**
  XAI integration introduces computational overhead, potentially affecting real-time detection capabilities in high-speed networks. Additionally, explanations from complex models may be less precise or harder for analysts to interpret effectively.

Explaining Intrusion Detection-Based Convolutional Neural Networks Using Shapley Additive Explanations (SHAP)

- **Authors & Publication:** Younis, B., Ahmad, A., & Abu Al-Haija, Q. (2020). *Big Data and Cognitive Computing*, 4(4), 136.
- **URL:** Younis, B., Ahmad, A., & Abu Al-Haija, Q. (2020). *Explaining intrusion detection-based convolutional neural networks using Shapley additive explanations (SHAP).* Big Data and Cognitive Computing, 4(4), 136.
- **Contribution:**
  Proposes a Convolutional Neural Network (CNN)-based Intrusion Detection System (IDS) that integrates SHAP to interpret and explain deep learning model predictions, improving transparency for security analysts.
- **Methodology:**
  Implements a CNN to classify network traffic and applies SHAP to its outputs to determine the contribution of each feature, explaining why a particular instance is classified as normal or an attack.
- **Limitations:**
  CNN models demand large datasets and high computational resources for training. Applying SHAP to such models can be computationally expensive, posing challenges for real-time interpretability.

An Ensemble-Based Approach for Efficient Intrusion Detection in Network Traffic

- **Authors & Publication:** Almamun, J., Akkur, J., Monem, I. M., Almaymuni, O., Alrahaymi, M., Abdullah, E., & Al-Shana, R. (2023). *Intelligent Automation & Soft Computing*, 37(2).
- **URL:** Almamun, J., Akkur, J., Monem, I. M., Almaymuni, O., Alrahaymi, M., Abdullah, E., & Al-Shana, R. (2023). *Ensemble-based approach for efficient intrusion detection in network traffic.* Intelligent Automation & Soft Computing, 37(2).
- **Contribution:**
  Proposes an ensemble-based approach to achieve efficient and accurate intrusion detection in network traffic, supporting the use of ensemble tree models for robust IDS performance.
- **Methodology:**
  Applies ensemble learning techniques that combine multiple classifiers to form a stronger predictive model, improving accuracy and reliability compared to single-model approaches.
- **Limitations:**
  Ensemble models are more complex and require longer training times. Additionally, without explainability methods like SHAP, understanding the model's internal decision-making remains challenging.

XGBoost: A Scalable Tree Boosting System

- **Paper Title:** XGBoost: A scalable tree boosting system.
- **Authors & Publication:** Chen, T., & Guestrin, C. (2016, August). *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).
- **URL:** Chen, T., & Guestrin, C. (2016, August). *XGBoost: A scalable tree boosting system.* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
- **Contribution:**
  Introduced XGBoost (eXtreme Gradient Boosting), an optimized and highly scalable implementation of gradient boosted decision trees, offering faster training and superior predictive performance compared to traditional boosting methods.
- **Methodology:**
  Implements an enhanced Gradient Boosting framework with parallelized computation, block structure optimization, and cache-aware data access, enabling efficient handling of large-scale datasets and high-dimensional problems.
- **Limitations:**
  Despite its efficiency, XGBoost is still an ensemble of trees, making it less interpretable than single decision trees. Additionally, achieving optimal results requires extensive hyperparameter tuning.

LLM-Powered Agent Frameworks for Cyber Security

- **Authors & Publication:** Zhu, Q. (2025). *arXiv preprint arXiv:2507.10621.*
- **URL:** Zhu, Q. (2025). *Game Theory Meets LLM and Agentic AI: Reimagining Cybersecurity for the Age of Intelligent Threats.* arXiv preprint arXiv:2507.10621.
- **Contribution:**
  Introduces a novel cybersecurity framework that integrates Game Theory, Large Language Models (LLMs), and Agentic AI to counter intelligent and adaptive cyber threats by modeling strategic attacker–defender interactions.
- **Methodology:**
  Applies Game Theory to analyze adversarial strategies and employs LLMs with Agentic AI to simulate intelligent threats, model decision-making processes, and enable adaptive, automated defense mechanisms.
- **Limitations:**
  The framework is primarily theoretical and exploratory. Implementing and validating such advanced LLM/Agentic AI systems in real-world cybersecurity settings poses major technical and reliability challenges.

Challenges

1. Limited Explainability

- Most IDS models act as black boxes, providing predictions without clear justification.
- SHAP improves transparency but remains difficult for non-technical analysts to interpret effectively.

2. Static Model Adaptability

- Traditional IDS models fail to adapt to new or evolving cyberattack patterns.
- Continuous learning and adaptive updating remain limited in current IDS research.

3. Lack of Agentic Automation

- IDS systems still rely heavily on manual human intervention for analysis and response.
- Integration of explainable models with autonomous AI agents is largely unexplored.

Problem Statement

In the rapidly evolving landscape of cybersecurity, traditional Intrusion Detection Systems (IDS) face critical limitations. Most existing models behave as black boxes, offering high accuracy but little insight into *why* a network event is flagged as malicious. This lack of transparency makes it difficult for security analysts to trust and act upon the results. Moreover, many IDS models are static, struggling to adapt to new or unknown attack patterns, which leads to reduced reliability over time. Finally, current systems largely depend on manual analysis and decision-making, slowing down response times during active threats.

To overcome these challenges, this project aims to develop an Explainable and Agentic IDS that integrates tree-based ensemble models for robust detection, SHAP explainability for feature-level insights, and LLM-powered AI agents for human-readable reasoning and automated threat response. This system will enhance trust, adaptability, and decision support in modern network security.

Existing Methodologies

1. Traditional Statistical Models (e.g., Naive Bayes, Logistic Regression)

- How it works: Uses statistical probabilities and assumptions to classify network traffic as normal or malicious.
- Why it's insufficient: Struggles with complex, non-linear relationships in high-dimensional network data, resulting in poor accuracy against modern sophisticated attacks.

2. Tree-Based Models (Random Forest, Decision Trees)

- How it works: Builds ensembles of decision trees, classifying traffic based on feature splits.
- Why it's insufficient: Although accurate, these models often function as black boxes, offering limited interpretability; predictions are not easily understandable by analysts.

3. Deep Learning Approaches (CNN, LSTM)

- How it works: Learns hierarchical or temporal patterns in network traffic using neural networks.
- Why it's insufficient: High computational requirements, difficult interpretability, and need for large labeled datasets; lacks explainability for informed decision-making.

Existing Methodologies

4. Clustering and Anomaly Detection (k-Means, Isolation Forest)

- How it works: Detects deviations from normal traffic patterns without supervision.
- Why it's insufficient: Produces high false-positive rates, cannot provide detailed reasoning for detections, and struggles to adapt to evolving attack types.

5. Existing Explainable AI (SHAP, LIME applied to IDS)

- How it works: Offers feature-level explanations for model predictions.
- Why it's insufficient: Explanations are static and not integrated with actionable decision support; analysts must manually interpret results without automated guidance or recommendations.

Proposed Solution

To overcome the limitations of traditional IDS, this project proposes an Explainable and Agentic Intrusion Detection System that initially experimented with Random Forest and XGBoost models for network intrusion detection. After comparative evaluation, XGBoost was selected as the final model due to its superior accuracy, scalability, and faster training performance over Random Forest. The ensemble-based nature of XGBoost ensures robust detection of complex attack patterns while maintaining computational efficiency.

To ensure transparency, SHAP is integrated to interpret model predictions and highlight the most influential features contributing to each alert. Building upon this, an Agentic AI layer powered by Large Language Models (LLMs) converts these technical SHAP explanations into human-readable summaries and provides actionable recommendations—such as blocking suspicious IPs or flagging potential threats. The system also includes a preprocessing pipeline for handling missing values, feature scaling, and encoding, with outputs visualized through a FastAPI and Streamlit interface for real-time, interactive analysis. This comprehensive design enhances accuracy, interpretability, and decision intelligence, empowering analysts to understand and act on threats efficiently.

Proposed Method

Data Preprocessing
- Load the NSL-KDD dataset.
- Handle missing values (imputation).
- Normalize numerical features (scaling).
- Apply one-hot encoding for categorical features.

Model Training
- Train **tree-based ensemble methods** (Random Forest, Gradient Boosting).
- Build a preprocessing pipeline for reproducibility and deployment.

Explainability Layer
- Apply **SHAP (Shapley Additive Explanations)** to interpret model predictions.
- Highlight which features contribute most to detection of an attack.
- Prepare it for integration into a real Intrusion Detection System (IDS)

Agentic AI Layer

- Use LLM + agent frameworks (Agno/CrewAI) to explain predictions in human language.
- Agents suggest possible responses (e.g., blocking an IP, flagging suspicious activity).

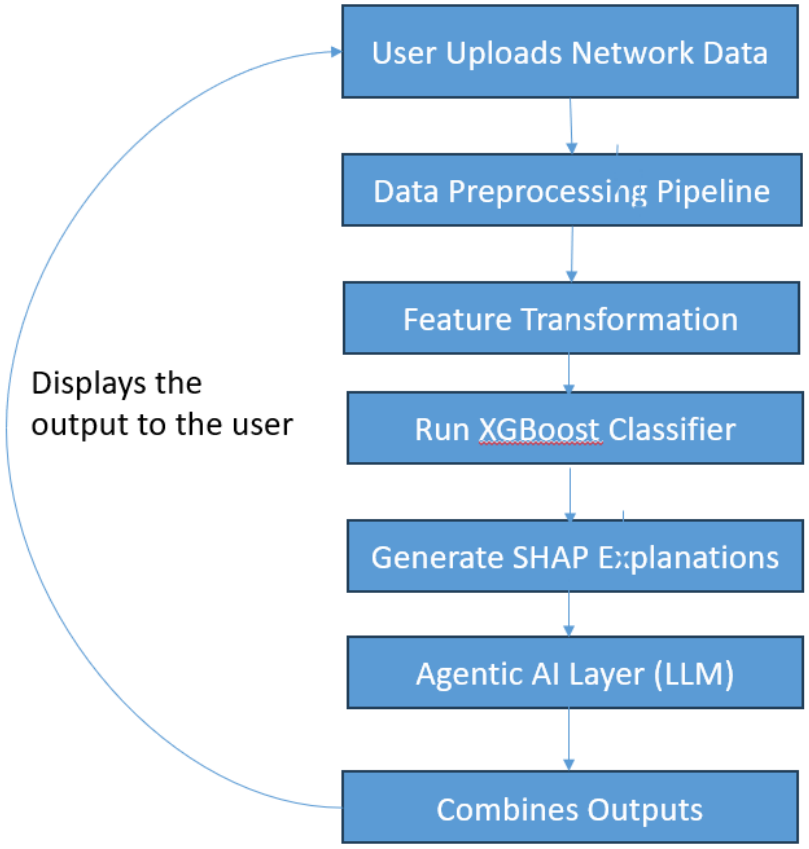
Model Evaluation

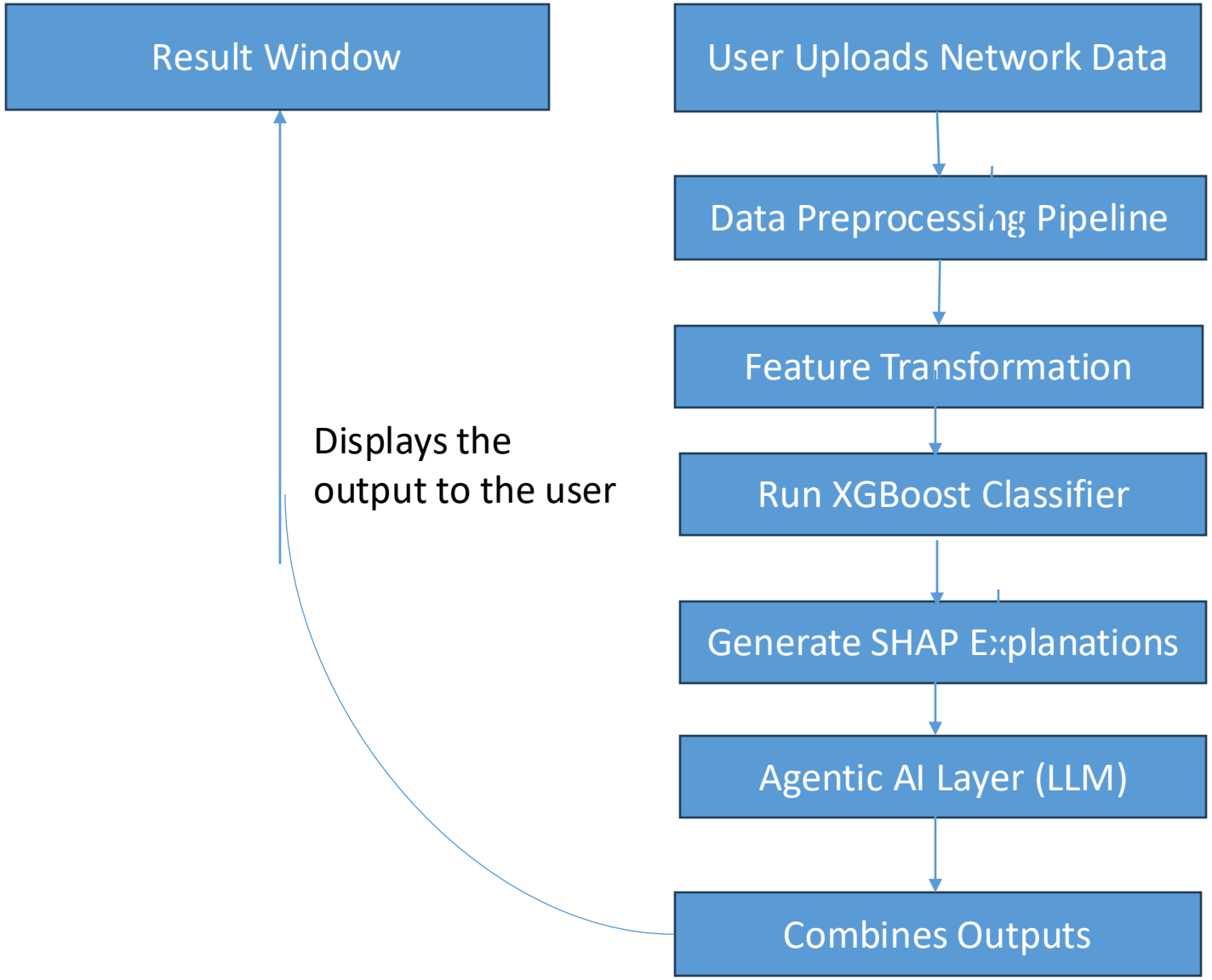
- Measure performance using Accuracy, Precision, Recall, F1-score.
- Use confusion matrix and AUC for deeper analysis.

Deployment (Optional)

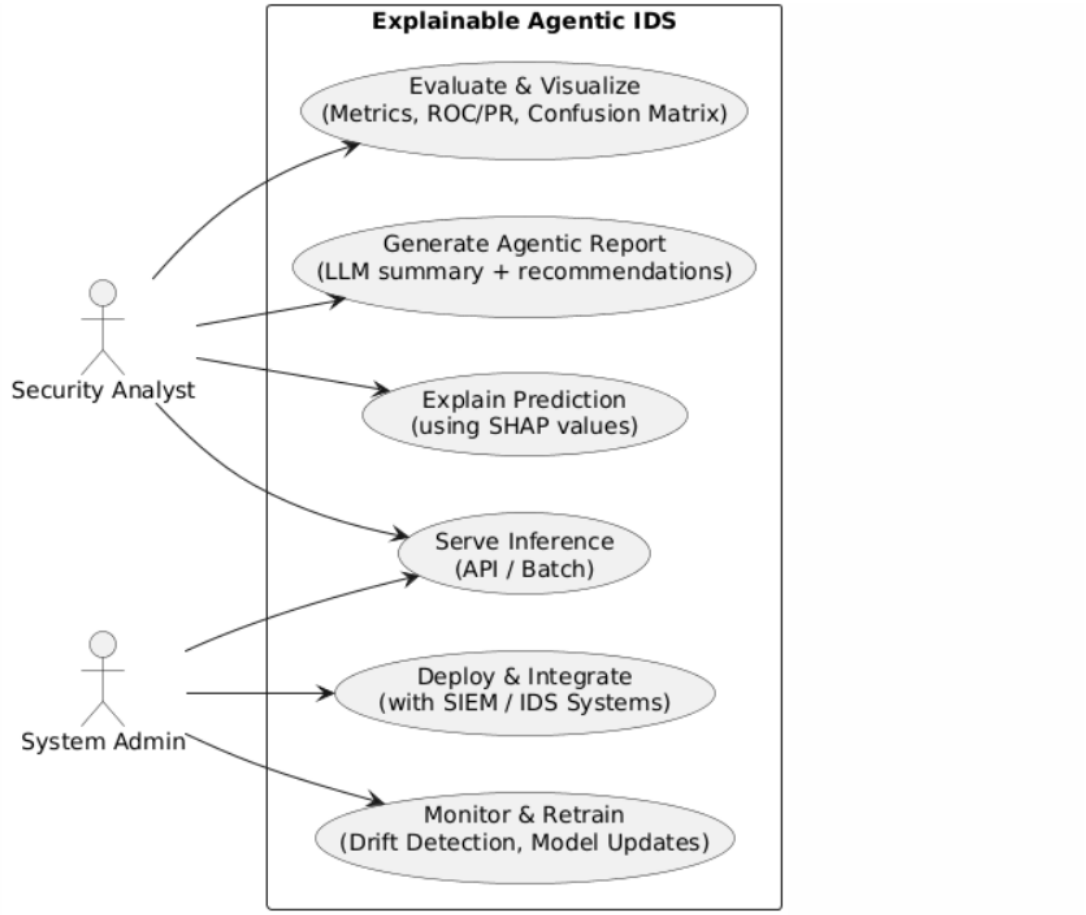
- Save the trained pipeline (preprocessing + model).

Architecture Diagram

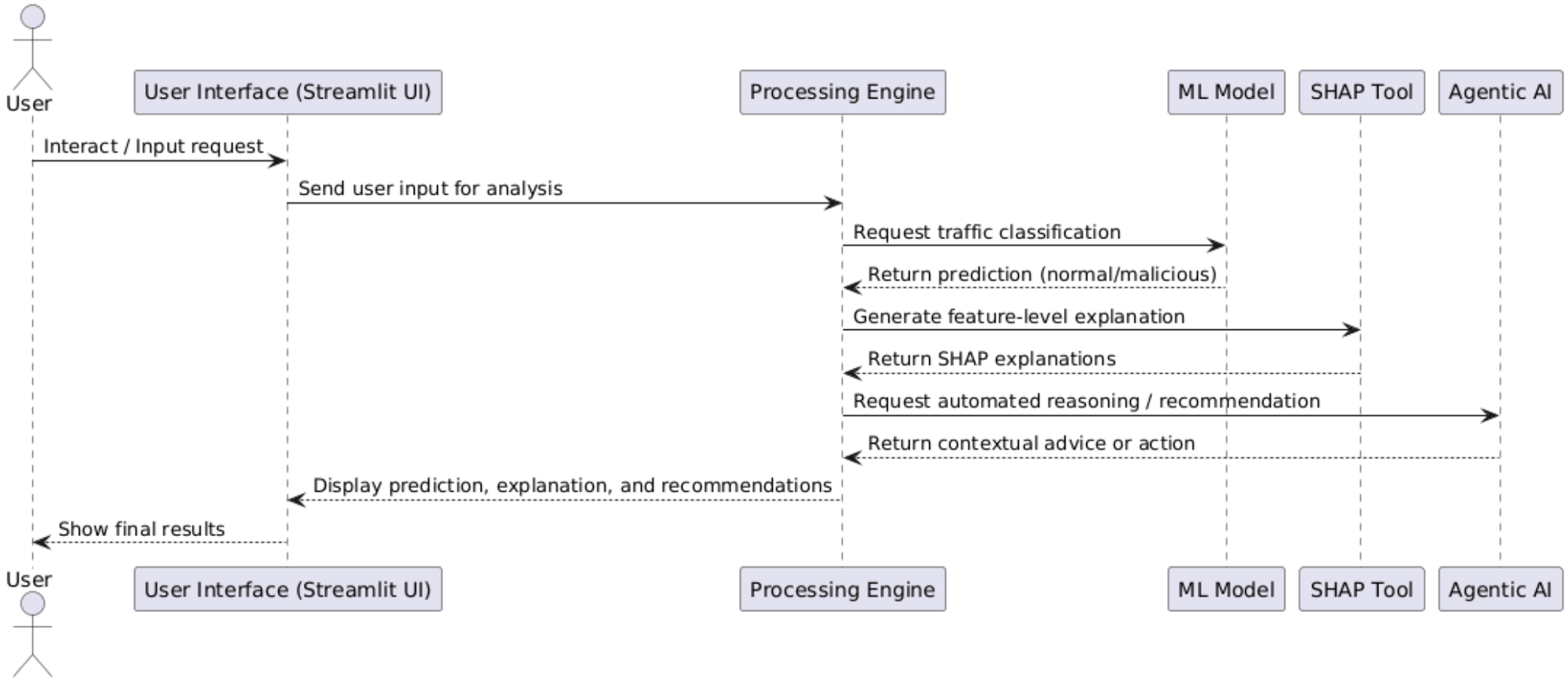



Result Window
User Uploads Network Data
Data Preprocessing Pipeline
Feature Transformation
Displays the
output to the user
Run XGBoost Classifier
Generate SHAP Explanations
Agentic AI Layer (LLM)
Combines Outputs

Use Case Diagram



Sequence Diagram



Modules Included

- Data Collection Module – Acquire and organize NSL-KDD dataset.
- Preprocessing Module – Clean, impute, scale, encode categorical and numerical features.
- Feature Engineering Module – Select and transform important features (via SHAP/importance scores).
- Modeling Module – Train ensemble ML models (Random Forest, Gradient Boosting).
- Explainability Module – Apply SHAP for interpretable predictions.
- Evaluation Module – Performance analysis (metrics, confusion matrix, ROC-AUC).
- Agentic AI Module – LLM-based assistant explains threats and suggests actions.
- Deployment Module (Optional) – Save model pipeline and enable real-time detection.

Hardware & Software Requirements

Hardware Specifications:

- CPU RAM: 16 GB
- Processor: Intel Core i7, 14th Generation
- GPU RAM: 64 GB

Software Specifications:

- Operating System: Windows
- Programming Language: Python
- Development Environment: Google Colab
- Libraries: scikit-learn, xgboost, shap, Crew AI/ Agno

Timeline for next review

TASK ID	TASK NAME	START DATE	END DATE	DURATION in days	WEEKS													
					1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Research and Analysis	28/06	18/07	21	✓	✓	✓											
2	Requirements	19/07	30/07	12	✓	✓												
3	Literature Review	31/07	20/08	21	✓	✓	✓											
4	Documentation	21/08	23/08	3	✓													
5	Review-1	23/08	23/08	1														
6	Design	25/08	29/09	35	✓	✓	✓	✓	✓									
7	Documentation	30/09	07/10	7	✓													
8	Review-2																	

References

- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
- Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.
- Le, Thi-Thu-Huong, et al. "Classification and explanation for intrusion detection system based on ensemble trees and SHAP method." *Sensors* 22.3 (2022): 1154.
- MP, P. J. (2024). An Explainable and Optimized Network Intrusion Detection Model using Deep Learning. *International Journal of Advanced Computer Science & Applications*, *15*(1).
- Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE symposium on computational intelligence for security and defense applications* (pp. 1-6). Ieee.
- Patil, S., Varadarajan, V., Mazhar, S. M., Sahibzada, A., Ahmed, N., Sinha, O., ... & Kotecha, K. (2022). Explainable artificial intelligence for intrusion detection system. *Electronics*, *11*(19), 3079.
- Younisse, R., Ahmad, A., & Abu Al-Haija, Q. (2022). Explaining intrusion detection-based convolutional neural networks using shapley additive explanations (shap). *Big Data and Cognitive Computing*, *6*(4), 126.
- Almomani, A., Akour, I., Manasrah, A. M., Almomani, O., Alauthman, M., Abdullah, E., ... & Al Sharaa, R. (2023). Ensemble-based approach for efficient intrusion detection in network traffic. *Intelligent Automation & Soft Computing*, *37*(2).
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Zhu, Q. (2025). Game Theory Meets LLM and Agentic AI: Reimagining Cybersecurity for the Age of Intelligent Threats. *arXiv preprint arXiv:2507.10621*.

Thank you