Date: 03/13/2024

PREPARED BY: GROUP F-33

Adit Sandeep Virkar

Arnav Raviraj

Shivanjay Vilas Wagh

Vinay Kantilal Chavhan

Project Plan

5G-NIDD: A Comprehensive Network Intrusion Detection Dataset Generated over 5G Wireless Network

Table of Contents

[Project Description 2](#_heading=h.gjdgxs)

[Overview: 2](#_heading=h.30j0zll)

[Key Requirements:](#_heading=h.1fob9te) 3

[Deliverables:](#_heading=h.3znysh7) 4

[Acronyms and Abbreviations :](#_heading=h.2et92p0) 5

[Design and Architecture](#_heading=h.tyjcwt) 6

[Implementation Strategy](#_heading=h.3dy6vkm) 10

[High-level Work Breakdown Structure :](#_heading=h.1t3h5sf) 10

[Schedule / Timeline :](#_heading=h.4d34og8) 18

[Required Hardware :](#_heading=h.2s8eyo1) 19

[Required Software :](#_heading=h.17dp8vu) 19

[Third Party Content / Libraries :](#_heading=h.3rdcrjn) 20

[Quality:](#_heading=h.26in1rg) 21

[Other Special Considerations:](#_heading=h.lnxbz9) 21

[Process](#_heading=h.35nkun2) 22

[Process Description and justification](#_heading=h.1ksv4uv) 22

[Tools :](#_heading=h.44sinio) 22

[Roles and Responsibilities:](#_heading=h.2jxsxqh) 23

[Location of Project Artifacts:](#_heading=h.z337ya) 24

[Sponsor Communications:](#_heading=h.3j2qqm3) 24

# **Project Description**

## **Overview:**

* **Statement of work:** Our topic for the capstone project is “ Detecting intrusion in softwarized 5G networks using machine learning”. Our sponsor, Dr. Abdallah Moubayed, assigned our team a research paper, “5D-NIDD: A comprehensive network Intrusion detection dataset generated over 5G Wireless network”, which provides a detailed exposition of the creation, characteristics, and analysis of a novel dataset, 5G-NIDD, developed for network intrusion detection research in 5G environments. The dataset is distinguished by its generation from a real, operational 5G test network, offering a unique and realistic benchmark for evaluating AI/ML-based security solutions. This dataset addresses the scarcity of comprehensive, current, and realistic datasets for network security research, especially in the context of the emerging 5G networks, characterized by their complexity, scale, and diverse service offerings.
* **Final Deliverables**: Our object is to perform various data processing techniques, such as preprocessing the 5G-NIDD dataset by removing all the null and zero columns to improve the accuracy and performance of the machine learning models. With this preprocessed data, we must apply various correlation techniques to extract the relevant features and normalize the dataset with extracted features. After this, we feed our processed dataset to the machine-learning models and analyze their performance and accuracy. We have to achieve a better result than what is presented in the assigned paper.
* **Business Domain:** The business domain of our research is network security within the context of 5G wireless technology. We must address the critical need for sophisticated and proactive security mechanisms to identify and mitigate intelligent attacks on the complex and heterogeneous infrastructure of 5G networks. The increasing reliance on 5G technology across various sectors, including critical infrastructure, healthcare, and IoT, underscores the importance of this research domain
* **Technical Platform:** As our work is primarily based on a dataset, 5G-NIDD, which is generated from a real 5G test network located in Oulu, Finland, which is part of the 5G Test Network Finland (5GTN). This network provides a realistic environment for generating network traffic, both benign and malicious, under controlled conditions. The dataset includes network flows derived from various attack scenarios, such as Distributed Denial of Service (DDoS) attacks, ICMP floods, UDP floods, and HTTP floods, among others, alongside normal traffic. Tools and technologies include Argus for flow data generation, Tracewrangler for data processing, and various attack simulation tools like hping3 and GoldenEye. As per our requirement, we only have to utilize the generated dataset and feed it to the machine learning models after processing it.
* **Motivation of work:** Building upon the research described in "5G-NIDD: A Comprehensive Network Intrusion Detection Dataset Generated over 5G Wireless Network," our motivation of work focuses on several vital enhancements and expansions to further the field of network security, particularly within the rapidly evolving 5G and future 6G environments such as the expansion of threat models, dynamic and real-time data labeling, incorporation of AI/ML for active defense mechanisms, distributed detection and federated learning, testbed and dataset as a service and research beyond 5G security research.

## **Key Requirements:**

* **Functional Requirements:**
* Data Preprocessing: The solution must be capable of preprocessing the 5G-NIDD dataset by
  + Identifying and removing all null and zero-value columns to ensure data quality.
  + Applying correlation techniques to identify and extract relevant features from the dataset.
  + Normalizing the dataset based on the extracted features prepares it for machine learning analysis.
* Machine Learning Model Integration: Implement machine learning models that can be trained on the preprocessed 5G-NIDD dataset to detect network intrusions.
  + The models must support classification tasks to distinguish between benign and malicious network traffic.
  + The solution should enable model performance evaluation using accuracy, precision, recall, and F1 score metrics.
* Performance Analysis: The system must include functionality to analyze and report the performance of the machine learning models.
  + Comparison against baseline results presented in the assigned research paper to determine improvement.
  + Detailed model performance analysis under various attack scenarios mentioned in the dataset, such as DDoS, ICMP floods, UDP floods, and HTTP floods.
* **Technical Requirement:**
* Dataset Handling: The technical platform must support efficient handling of the 5G-NIDD dataset, including:
  + Secure storage and access mechanisms for the dataset.
  + Capability to process large volumes of network flow data generated from an actual 5G test network.
* Machine Learning Tools and Libraries: Utilize appropriate machine learning libraries and frameworks compatible with the dataset's characteristics and the project's goals.
  + The chosen libraries must support the preprocessing techniques, feature extraction, normalization, and model evaluation criteria outlined in the functional requirements.
* Scalability and Extensibility: Design the solution with scalability in mind to accommodate future enhancements, such as
  + Expanding the dataset with additional attack scenarios or more complex network environments.
  + Integrating additional machine learning models or deep learning approaches for enhanced detection capabilities.
* **Performance Requirement:**
  + Accuracy Improvement: The machine learning models must achieve a higher accuracy in detecting network intrusions than the baseline presented in the research paper.
  + Specific accuracy targets may be set based on preliminary analysis and the performance of existing models documented in the research paper.
  + Processing Time: The preprocessing and analysis processes should be optimized for efficiency.
  + Define maximum acceptable processing times for data preprocessing and model training phases.
  + Resource Utilization: Ensure that the solution is optimized for minimal and efficient use of computational resources, considering the complexity and volume of the 5G-NIDD dataset.

## **Deliverables:**

* Sprint 1:
  + Develop an in-depth understanding of the assigned research paper, “5G-NIDD: A Comprehensive Network Intrusion Detection Dataset Generated over 5G Wireless Network.”
  + Understanding the reference papers on 5G and 6G intrusion detection systems will help you understand the requirements to improve the performance and accuracy of machine learning models.
* Sprint 2:
  + Perform Data Preprocessing on the 5G-NIDD dataset, such as finding correlation coefficient, input sampling, and data normalization.
  + Apply boosting algorithms to the processed data and analyze the accuracy.
* Sprint 3:
  + Exploration of Feature selection on the provided 5G-NIDD dataset.
  + Improve Boosting model performance by feeding processed data into the boosting algorithm.
  + Implement ANN (Artificial Neural Networks) based model to analyze the accuracy and performance.
  + Implement a GAN (Generative Adversarial Networks) based model to analyze the accuracy and performance.
* Sprint 4 (till date):
  + Implement feature selection and input sampling techniques on the dataset.
  + Explore and Implement various boosting algorithms using multi-classification.
  + Implement a GAN-based model by applying multi-classification.
  + Implement ANN models using TensorFlow and modify the metrics.
* Final Deliverables:
  + This is a comprehensive report detailing the data preprocessing steps, feature extraction, and normalization techniques, machine learning models used, their performance analysis, and a comparison with the baseline results.
  + A software solution that includes all the code for data preprocessing, machine learning model implementation, performance analysis, and documentation for setup, configuration, and usage.

## 

## **Acronyms and Abbreviations:**

|  |  |
| --- | --- |
| **Glossary of Abbreviations and Acronyms** | **Definition** |
| 5G-NIDD | 5th Generation Network Intrusion Detection System |
| 5GTN | 5G Test Network |
| AdaBoost | Adaptive Boosting |
| AI | Artificial Intelligence |
| ANN | Artificial Neural Networks |
| CatBoost | Categorical Boosting |
| DoS | Denial of Service |
| GAN | Generative Adversarial Networks |
| IDS | Intrusion Detection System |
| ML | Machine Learning |
| ReLU | Rectified Linear Unit |
| XgBoost | Extreme Gradient Boosting |

Table 1: Acronyms and Abbreviations

## Process

## Process Description and justification:

 For our project, we have chosen an adapted Agile development process specifically tailored to our team's and the project's requirements. This process involves using Taiga as our primary project management tool, supplemented by Google Drive for document storage and YouTube for sharing sprint review meetings.

 The adapted Agile process is justified due to its flexibility, iterative nature, and focus on collaboration, which aligns with the dynamic requirements of our machine learning project. We can effectively plan and track our progress throughout each sprint by breaking down our work into manageable user stories and tasks within Taiga. Additionally, Taiga's support for Agile methodologies ensures that our team can easily prioritize tasks, adjust workflows as needed, and respond to project scope or requirements changes.

 Furthermore, integrating Google Drive allows us to organize and share project documentation seamlessly, enhancing collaboration and ensuring accessibility for all team members. This aligns with the sponsor's requirement for regular updates.

 The chosen development process meets the sponsor's requirements by facilitating efficient project management, fostering collaboration, and enabling transparent communication. It suits our team's needs, allowing us to adapt to changing requirements and deliver high-quality results within the project timeline.

## Tools:

For the project focused on improving ML models for network intrusion detection, several non-hardware tools are essential for development and software process management:

* Python Programming Language: Python is the primary programming language for developing machine learning models due to its extensive libraries and frameworks, such as scikit-learn, TensorFlow, PyTorch, and XGBoost.
* Jupyter Notebooks and Google Colabs: Jupyter Notebooks and Google Collabs are invaluable for exploratory data analysis, prototyping models, and documenting the development process. They provide an interactive environment for running code, visualizing data, and writing explanatory text.
* Data Analysis and Visualization Libraries: Pandas, NumPy, sckitlearn, and matplotlib are indispensable for data preprocessing, analysis, and visualization. They facilitate tasks like data cleaning, feature engineering, and generating visualizations to gain insights into the dataset.
* XGBoost, AdaBoost, CatBoost, and GAN Libraries: Specific libraries like XGBoost and GAN frameworks (e.g., TensorFlow or PyTorch) are crucial for implementing and fine-tuning the XGBoost algorithm and generative adversarial networks (GANs) for improving model accuracy and robustness.
* Version Control Systems: Tools like Git and GitHub are essential for version control, collaboration, and managing codebase changes across team members. They enable tracking of project history, branching, merging, and resolving conflicts.
* Integrated Development Environments (IDEs): IDEs such as PyCharm, VS Code, SublimeText, or JupyterLab provide sophisticated development environments with features like code autocompletion, syntax highlighting, debugging, and project management, enhancing developer productivity.
* Taiga: Taiga is chosen for project management due to its intuitive interface and support for Agile methodologies. It facilitates efficient collaboration and task tracking for the team's iterative development process. Its flexibility and integration capabilities further enhance productivity and streamline project workflows.
* Zoom and YouTube: Zoom and YouTube are used for recording meetings due to Zoom's ease of use and comprehensive recording features. YouTube provides a platform for sharing recorded meetings, ensuring accessibility and transparency for stakeholders. One meeting is uploaded per sprint to maintain an organized and accessible archive of project progress.

By effectively leveraging these non-hardware tools, the development team can streamline the development process, improve collaboration, and deliver high-quality ML models for network intrusion detection that meet the project objectives and stakeholder requirements.

## Roles and Responsibilities :

In our project development process, we have defined specific roles and responsibilities for each team member to ensure the smooth execution of tasks. While roles may rotate throughout the project, this transition will be managed effectively to maintain continuity and efficiency. It's crucial to note that every team member, regardless of their assigned role, contributes significantly to developing software deliverables. These roles provide clarity and allow individuals to take ownership of specific aspects of the project, fostering accountability and teamwork.

**Sprint 1:**

Roles:

* Scrum Master: Arnav Raviraj
* Product Advocate: Dr. Abdallah Moubayed
* Developers: Adit Sandeep Virkar, Vinay Kantilal Chavhan, Shivanjay Vilas Wagh

**Sprint 2:**

Roles:

* Scrum Master: Vinay Kantilal Chavhan
* Product Advocate: Dr. Abdallah Moubayed
* Developers: Adit Sandeep Virkar, Arnav Raviraj, Shivanjay Vilas Wagh

**Sprint 3:**

Roles:

* Scrum Master: Shivanjay Vilas Wagh
* Product Advocate: Dr. Abdallah Moubayed
* Developers: Adit Sandeep Virkar, Arnav Raviraj, Vinay Kantilal Chavhan

**Sprint 4:**

Roles:

* Scrum Master: Adit Sandeep Virkar
* Product Advocate: Dr. Abdallah Moubayed
* Developers: Shivanjay Vilas Wagh, Arnav Raviraj, Vinay Kantilal Chavhan

## Location of Project Artifacts :

Regarding project organization, our choice encompasses a combination of platforms to optimize usability, reliability, and change management. GitHub is our primary repository for version control and collaborative coding efforts, offering robust features for code management, issue tracking, and pull requests. To access our project materials, including codebase and documentation, please click <https://github.com/Araviraj8/SER_517_F_33>.

Additionally, we utilize Google Drive to store and share project documentation, meetings, and other non-code assets. Its intuitive interface and real-time collaboration capabilities enhance usability and ensure easy access to essential project resources for all team members. For supplementary project documents and materials, please access our Google Drive folder: <https://drive.google.com/drive/folders/17fXt0y-jq5i_r6ynpcHEslSKF3EHksBE>.

We rely on Taiga to facilitate project planning, task management, and agile workflows. This platform enables us to organize user stories, track tasks, and prioritize features, fostering efficient project execution and seamless change management. For project management-related tasks and updates, please visit our Taiga project board at <https://tree.taiga.io/project/araviraj8-ser517-f-33/backlog>.

Overall, this multifaceted approach to project organization ensures that our team can effectively collaborate, maintain reliability in our deliverables, and adapt to changes throughout the project lifecycle.

## Sponsor Communications :

Our sponsor meetings are scheduled bi-weekly on Tuesdays at the Polytechnic campus. Due to the spring break, one of our meetings has been rescheduled to the following Tuesday. Below are the dates and details of the sponsor meetings:

**Sponsor Meeting 1:**

Date: Jan 23, Tuesday, 1:30 pm

Our sponsor gave us a research paper entitled '5G-NIDD: A Comprehensive Network Intrusion Detection Dataset Generated over 5G Wireless Network'. They introduced the problem statement, described the dataset's generation process, and highlighted its importance. Our task was to investigate new machine-learning models. In the following meeting, we were assigned to conduct a literature review, familiarize ourselves with the project and the paper, and so forth. The overall objective was to enhance the models examined in the paper and conduct an analysis.

**Sponsor Meeting 2:**

Date: Feb 6, Tuesday, 1:30 pm

We handed in our literature review and received feedback. During our discussion, we focused on the XGBoost model, known for its superior performance compared to Random Forest (RF) and Artificial Neural Networks (ANN) examined in the original paper. Our suggestion involves leveraging XGBoost to enhance the model's performance across various metrics such as accuracy, inference time, and training time. Additionally, we delved into topics like adversarial machine learning and federated learning (FL), the latter being a distributed approach for deploying intrusion detection models. Furthermore, we addressed concerns regarding noise attacks and poisoning attacks during the conversation.

**Sponsor Meeting 3:**

Date: Feb 20, Tuesday, 1:30 pm

During the meeting, we engaged in a comprehensive discussion with the Sponsor regarding the application of our model. We presented our analysis, focusing on the utilization of boosting models. The Sponsor suggested exploring the potential of GAN-based and ANN-based algorithms, prompting us to consider various classification techniques for the test set and to experiment with different data processing methods.

Furthermore, we discussed the possibility of applying alternative data processing techniques to the raw data, including various feature selection strategies, to enhance performance. The Sponsor guided us in prioritizing fixing one model and implementing ANN and GAN-based models before our next meeting, scheduled after spring break.

**Sponsor Meeting 4:**

Date: March 12, Tuesday, 1:30 pm

We analyzed various boosting, ANN, and GAN-based models for binary classification. The sponsor has recommended expanding the analysis from binary to multi-class classification, broadening the scope of the study to handle more than two classes in classification tasks. Additionally, they suggest employing input sampling techniques to introduce variations in the training data. Hyperparameter tuning is also emphasized to fine-tune the model's parameters for optimal performance, with techniques such as grid search or random search being helpful in this context. Finally, the sponsor suggests exploring various evaluation metrics to assess model performance accurately in multi-class classification scenarios, including but not limited to accuracy, precision, recall, and F1-score.

# 

# 