


Est.
1841

YORK
ST JOHN
UNIVERSITY

Dissertation Proposal Form

Date of Submission: 1/6/2024

| | |
|-----------------------------|---|
| Name | Ayush Ale |
| Student Id | 240016074 |
| Module Code | COM7040M |
| Project Title | Prevention of cyber-attacks using Intrusion Detection System (IDS) with Machine Learning Algorithms |
| Supervisor Name | Dr. Soonleh Ling |
| Supervisor Approval | Yes |
| Supervisor Signature |  |

Section 1: Academic

This section helps Academic staff assess the viability of your project. It also helps identify the most appropriate supervisor for your proposed research. This proposal will be referred to as a point of discussion by your supervisor in seminar sessions.

| | |
|--|----------------------------------|
| NAME: Ayush Ale | STUDENT NUMBER: 240016074 |
| PROPOSED TITLE OF PROJECT: Prevention of cyber-attacks using Intrusion Detection System (IDS) with Machine Learning Algorithms | |
| BRIEFLY DESCRIBE YOUR FIELD OF STUDY: This project "The Effect of a firewall to prevent cyber-attacks in cyber security networks Using Intrusion detection using Machine Learning Algorithms " focuses on enhancing the cybersecurity measures in firewall with machine learning algorithms. This is achieved by integrating three key technologies: firewall technologies, Intrusion Detection System (IDS) and machine learning algorithms. The IDS is used to monitor network traffic for patterns indicative of cyberattacks, such as unusual login attempts, outbound connections to known malicious domains, or sudden spikes in traffic. On the other hand, firewall is a network security device or software designed to monitor and control incoming and outgoing network traffic based on predetermined security rules, and machine learning algorithms are techniques used to enable computers to learn from and make predictions or decisions based on data. This project goal is to gain knowledge about firewalls, identify the limitations of conventional firewalls in thwarting and identifying sophisticated cyber-attacks, and combine firewall technology with machine learning algorithms for enhanced performance. A detailed review of the current cybersecurity environment, firewall technologies, intrusion detection systems, and machine learning approaches is conducted in order to achieve these goals. | |
| WHAT QUESTION DOES YOUR PROJECT SEEK TO ANSWER? This project seeks to answer the following questions: 1. How effective is the integration of firewalls with IDS and machine learning algorithms in preventing cyber-attacks? By offering comprehensive threat detection, lowering false positives, and enabling proactive defense against both known and unknown cyberattacks, integrating firewalls with IDS and machine learning algorithms significantly improves cybersecurity. 2. Which machine learning algorithms provide the best performance in detecting network intrusions? Among the machine learning algorithm Random Forest, Support vector machines (SVM) and Neural network performs best for detecting network intrusion among them we are discussing Random Forest in this project. 3. What are the limitations of traditional firewalls in identifying and preventing complicated cyber-attacks? Traditional firewalls suffer for limitation in identifying and preventing complicated cyber-attacks due to their strong dependent on predefined rules and signature, lack of adaptive ability to learn and recognize new threats and vulnerabilities. | |

WHAT HYPOTHESIS ARE YOU SEEKING TO TEST?

The hypothesis for this project is: "Integrating firewalls with IDS powered by machine learning algorithms significantly enhances the detection and prevention of cyber-attacks compared to standalone firewall systems."

Here are some potential hypotheses:

Null hypothesis(H0): Integrating firewalls with IDS powered by machine learning algorithms does not significantly enhance the detection and prevention of cyber-attacks compared to standalone firewall systems.

Alternative hypothesis(H1): Integrating firewalls with IDS powered by machine learning algorithms significantly enhances the detection and prevention of cyber-attacks compared to standalone firewall systems.

WHAT ARE THE PROBABLE PROJECT OUTCOMES?

The probable project outcomes of combining machine learning with firewall and intrusion detection system are:

- Understanding the Strengths and Weaknesses of Firewalls and IDS in Cybersecurity
- Identification of best machine learning algorithm suited for Intrusion Detection.
- Creation of smart security system which combines Firewalls and IDS with machine learning.

PLEASE PROVIDE A BRIEF BIBLIOGRAPHY OF 2-4 KEY TEXTS FOR YOUR STUDY (USE HARVARD REFERENCE STYLE)

1. Alsmadi, I., Al-Hubaishi, A., Al-Mansoori, A. & Al-Hinai, A., 2017. Industrial control systems security and firewalls.. Journal of Information Security and Applications, Volume 35, pp. 15-35.
2. Aronson, J. P., Brownlee, N., Byrum, F. & Ramsey, M. S., 1997. Site Security Handbook. [Online] Available at: <https://www.ietf.org/rfc/rfc2196.txt?> [Accessed 01 November 2022].
3. Atawodi, I. S., 2019. A MACHINE LEARNING APPROACH TO NETWORK INTRUSION DETECTION SYSTEM USING K NEAREST NEIGHBOR AND RANDOM FOREST, University of Southern Mississippi: s.n.
4. Biggio, B., 2012. Poisoning Attacks against Support Vector Machines. In Proceedings of the 29th International Conference on Machine Learning.
5. Cheswick, W. R. & Bellovin, S. M., 1994. Firewalls and Internet Security: Repelling the Wily Hacker. s.l.:Addison-Wesley..
6. Debar, H., Dacier, M. & Wespi, A., 1999. Towards a taxonomy of intrusion-detection systems. Computer Networks, 31(23-24), pp. 805-822.

PLEASE NAME ANY MEMBER OF THE ACADEMIC TEAM YOU HAVE DISCUSSED THIS POTENTIAL PROJECT:

NA

(staff use only) Project Approved by Academic Team?

YES

NO

Any other Academic Staff comments

Student did not fully understand the scope of the project.

Student needs to address the following areas during the project phase:

- Detailed understanding of field of study required and project scope.
- Clear provision of project deliverables and outputs.
- Detailed description of project technical aspects is missing.
- Project timeline is unclear.

Student should conduct further research and study to fully understand the project.

Section 2: Technical

This section is designed to help the technical team ensure the appropriate equipment to support each project has been ordered. It also exists to help you fully ascertain the technical requirements of your proposed project. In filling out this section please note that we do not 'buy' major items of equipment for student projects. However, if a piece of equipment has a use to the department beyond the scope of a single project, we will consider purchasing it. Though purchasing equipment through the university is often a slow process.

PLEASE DESCRIBE YOUR PROJECT IN TECHNICAL TERMS:

The main Aim of this project is to build a prediction model that can detect network intrusion, get a greater understanding of how firewalls help prevent cybercrime on networks, and evaluate several models based on their performance and assessment criteria.

The key challenge is to detect an intrusion continuously and provide an automated response. Intrusions can occur from both internal and external sources. Internal intrusions may involve unauthorized access from friends, partners, employees, or even disgruntled clients who have gained access to the networking system. External intrusions, also known as internet attacks, involve intrusions from outside the network system.

Proposed Methodology

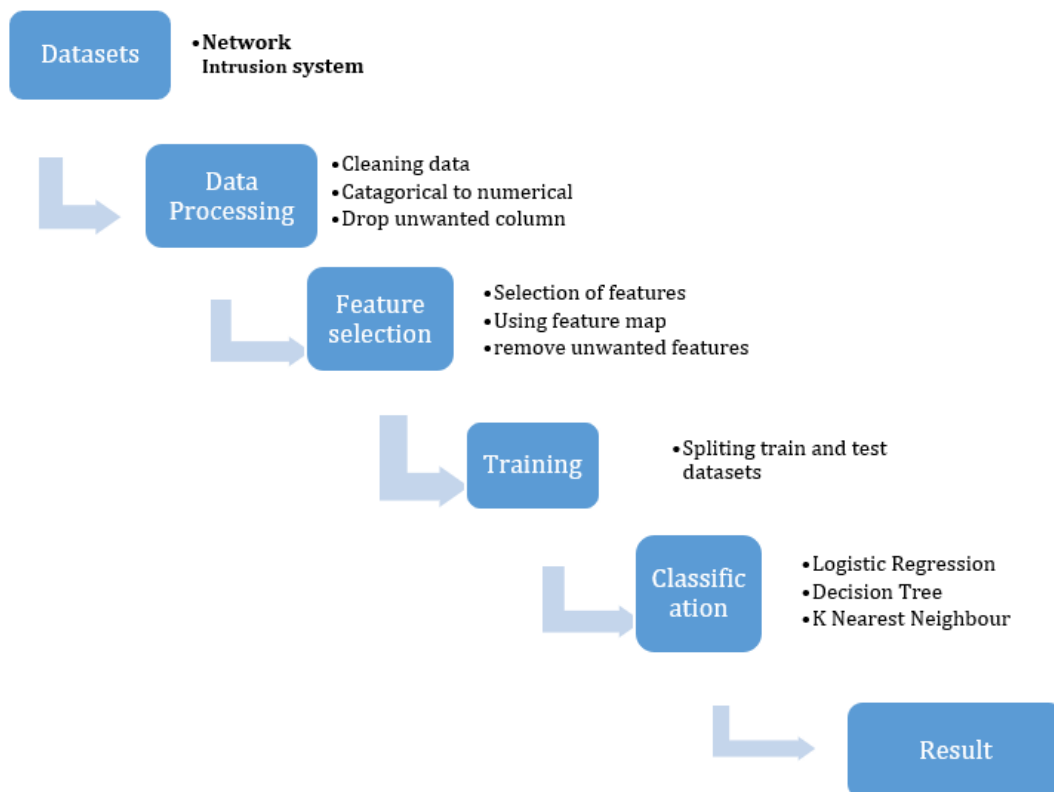


Figure 1: Steps for prediction for this research

Prediction of the attack can be done in following steps:

1. **Analysis of data:** Data is collected from open-source website KAGGLE. It consists of two types of datasets: train dataset and test dataset. The test datasets have 22544 rows and 41 columns, whereas the train datasets have 25192 rows and 42 columns. The data is evaluated based on the names, descriptions, and categories of the attributes. After which scikit-learn library's class RFE (Recursive Feature Elimination) is used to choose features. After choosing right feature with the help of Machine learning Techniques like DT (Decision Tree), LR (Logistic Regression), KNN (K-Nearest neighbour) we can increase the accuracy to find anomaly connection on the internet.
2. **Result and Analysis:** The datasets are divided and extracted according to their primary features using the feature selection approach. By cleaning, a total of 17634 rows and the top 12 columns of data were chosen for modelling or prediction. It contains 13449 normal and 11743 anomaly connections. The models K Nearest Neighbour (KNN), Decision Tree, and Logistic Regression are used to predict anomaly connections.

The following tools are used in this project:

Python:

- Purpose: Writing code for ML algorithm
- Brief Description: Python programming language is utilised in this project to construct the prediction model for the intrusion detection system.

Jupyter Notebook:

- Purpose: Development Environment for Python
- Brief Description: Jupyter notebook is a Python-based development environment, to write the code for this project

Numpy:

- Purpose: Numerical Calculations
- Brief Description: NumPy, which stands for "Numerical Python," is a powerful Python library for numerical calculation.

Matplotlib:

- Purpose: Creating Graphical representations of data
- Brief Description: Matplotlib is a well-known Python toolkit for creating static, animated, and interactive visualisations.

Pandas:

- Purpose: Handling and analysing of data
- Brief Description: A powerful Python library for handling and analysing data is called Pandas. For efficiently handling structured data, such as time series data and numerical tables, it provides data structures and procedures.

Seaborn:

- Purpose: Present statistical graphics for data
- Brief Description: Seaborn library of python for building statistical graphics. It is based on Matplotlib and strongly integrated with Panda's data structures. Exploration and understanding of data are aided by Seaborn.

Scikit Learn:

- Purpose: Tools for Machine learning Application
- Brief Description: Python's scikit-learn, sometimes known as sklearn, is a popular machine learning library. For a variety of machine learning applications, including as model selection, regression, clustering, dimensionality reduction, classification, and data preparation, it provides a wide range of tools and approaches.

Algorithm used in this project:

Random forest (RF):

Random forest is a machine-learning technique that is quite popular. It uses a collaborative learning approach, which results in the creation of multiple decision trees during the training phase of the algorithm (thus the name "forest").

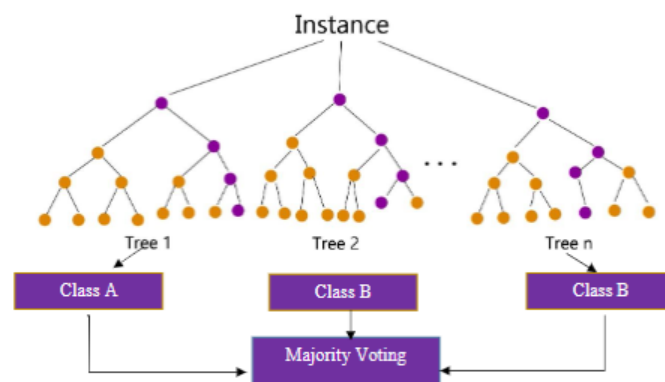


Figure 2: A flowchart illustrating the use of random forests

Decision Tree (DT):

For both classification and regression issues, a decision tree supervised machine learning method is used. It is a model containing core nodes that represent feature tests, branching nodes that represent the outcomes of decisions, and leaf nodes that represent class labels or target values.

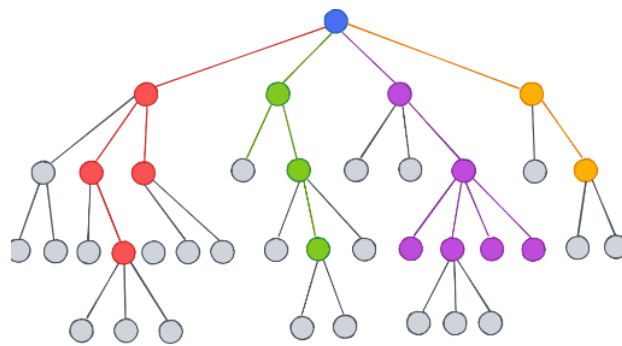


Figure 3: Decision tree

Logistic Regression (LR):

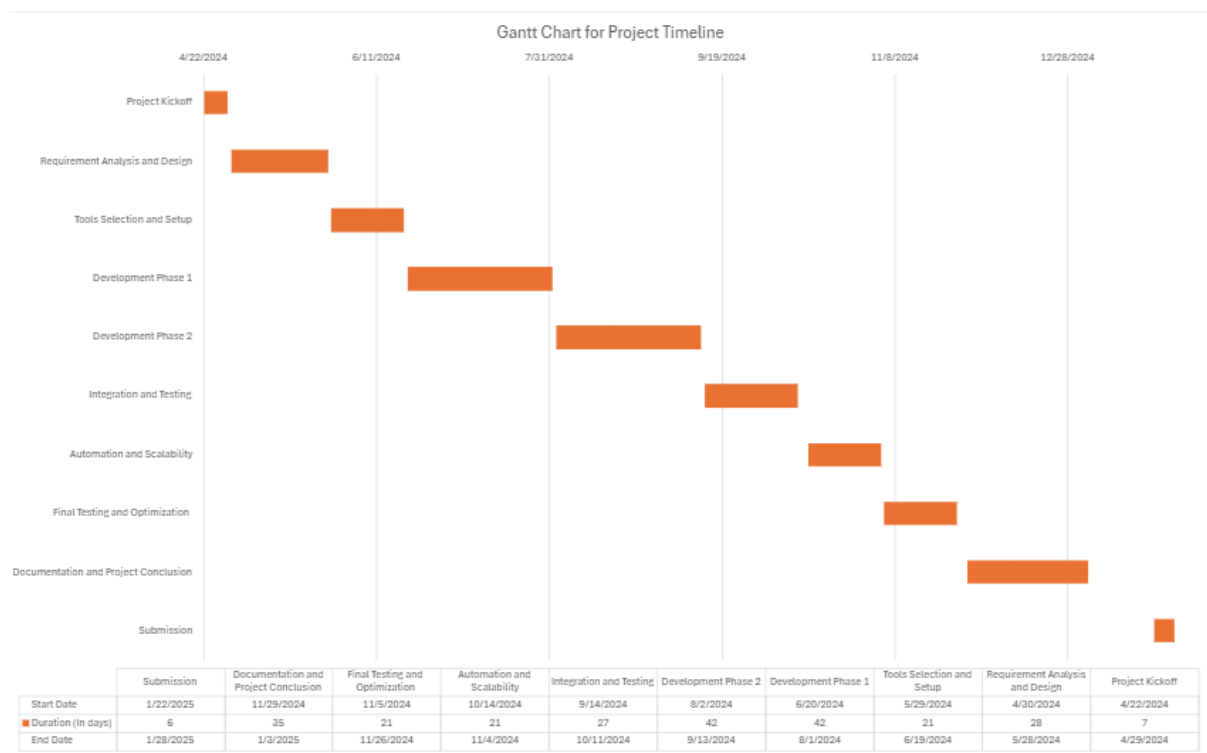
Logistic regression is a machine learning method which are widely used for binary classification applications. It's extensively used in a range of fields, including marketing, finance, healthcare, and social sciences. Logistic regression describes the relationship between a set of independent variables and a binary outcome by estimating the probability that a result will fall into one of several groups.

K-Nearest Neighbour Algorithm (KNN):

(K-NN) algorithm is a versatile and widely used machine learning algorithm that is primarily used for its simplicity and ease of implementation. KNN is a supervised learning algorithm that belongs to the domain of pattern recognition, data mining, and intrusion detection.

These tools and algorithms collectively contribute to the overall functionality of this project.

PROJECT TIMELINE



WHAT EXISTING LAB EQUIPMENT DO YOU NEED ACCESS TO UNDERTAKE YOUR PROPOSED PROJECT:

Computer systems, Virtual machines.

PLEASE LIST ANY **MINOR** EQUIPMENT YOU MUST PURCHASE TO COMPLETE YOUR RESEARCH PROJECT: (eg, switches, resistors, raspberry pi, Arduino etc)

NA

PLEASE LIST ANY **MAJOR** EQUIPMENT YOU REQUIRE TO COMPLETE YOUR RESEARCH PROJECT ALONG WITH LINKS TO WHERE IT MAY BE PURCHASED (eg a Drone, mobile phone etc).

NA

HAVE YOU DISCUSSED THE FEESIBILITY OF YOUR PROJECT WITH A MEMBER OF THE TECHNICAL TEAM? IF SO, WHO?

| | | | | |
|--|-----|--|----|--|
| NO | | | | |
| (staff use only) Project Approved by Technical Team? | YES | | NO | |
| Please comment on the Feasibility of the project: | | | | |
| | | | | |

Section 3: Ethics Approval

This section of the form will help ascertain if you need to complete and undergo the universities research ethics approval process. Please answer all questions honestly.

| Question | Yes | No |
|---|-----|-------------------------------------|
| Does your Research involve any of the following? Human participants / subjects, Human tissue, Documents | | <input checked="" type="checkbox"/> |
| Will the research require the collection of primary source material that might be considered offensive or illegal to access or hold on a computer? (e.g. studies related to state security, pornography, abuse, illegal behaviour, or terrorism). | | <input checked="" type="checkbox"/> |
| Does your research concern group which may be construed as terrorist or extremist? | | <input checked="" type="checkbox"/> |
| Will the research involve visual/vocal methods where participants may be identified? | | <input checked="" type="checkbox"/> |
| Will the research involve the use of genetic data (inherited/acquired genetic characteristics resulting from the analysis of a biological sample)? | | <input checked="" type="checkbox"/> |
| Will the study require the co-operation of a gatekeeper to give access to, or to help recruit, participants? (eg, headteacher or group leaders publicising your work) | | <input checked="" type="checkbox"/> |
| Will it be necessary for participants to take part in the study without their knowledge or consent at the time? | | <input checked="" type="checkbox"/> |
| Will the study involve recruitment of patients through the NHS? | | <input checked="" type="checkbox"/> |
| Will inducements be offered to participants? (eg the offer of being entered into a prize draw) | | <input checked="" type="checkbox"/> |
| Does the study involve participants who are particularly vulnerable or unable to give informed consent? (e.g. participants under 18. Adults with learning disabilities, the frail elderly, or anyone who may be easily coerced due to lack of capacity) | | <input checked="" type="checkbox"/> |
| Is there a possibility that the safety of the researcher may be in question? | | <input checked="" type="checkbox"/> |
| Will the study require participants to commit extensive time to the study? | | <input checked="" type="checkbox"/> |
| Are drugs, placebos, or any other substances to be administered to participants, or will the study involve invasive, intrusive, or potentially harmful procedures of any kind? | | <input checked="" type="checkbox"/> |
| If there are experimental and control groups, will being in one group disadvantage participants? | | <input checked="" type="checkbox"/> |
| Is an extensive degree of exercise or physical exertion involved? | | <input checked="" type="checkbox"/> |

| | | |
|--|--|-------------------------------------|
| Will blood or tissue samples be obtained from participants? | | <input checked="" type="checkbox"/> |
| Could the study induce psychological stress or anxiety or cause harm or negative consequences beyond the risks encountered in normal life? | | <input checked="" type="checkbox"/> |

*This part of Section 3 requires you to thoroughly **identify** and **mitigate** the ethical challenges of your research project. This is required to enable the computer Science ethics panel to properly consider if your proposed project requires you to submit a formal proposal to the university ethics panel.*

With your answers to the previous questions in mind, please describe the main ethical challenges of your research project and how you propose to mitigate them. Your discussion may include material not covered in the above questions. Please be as thorough as possible:

N/A