Programme Name:	**BSc (Hons) Computing [Top-up]**
Partner University Name:	**Newcastle College Group**
Semester 1 or 2:	**Semester 1**
Module Code:	**CMP600**
Name of Lecturer:	**Ashfaque E. Alam**
Due Date:	**22.12.2025**
Assessment Type:	**Proposal report**

Declaration of Authenticity (please check and sign):	
By submitting this assignment, I declare that this work is fully my own, and the work of others (including internet sources) is acknowledged by quotations and appropriate citing and referencing.	✓
I declare that this work has not made use of the work of any other student(s) past or present at this or any other educational institution or source.	✓
I declare that I have not commissioned another person to complete this work. This could include the use of professional essay-writing services, essay banks, or ghost-writing services.	✓
I declare that I have not asked another person to complete this work for me. This could include the help of friends, classmates, other students, or family members.	✓
I declare I have not used any AI software to help write this work outside of the assessment brief/guidelines (this includes translation tools, paraphrasing tools, or content creation websites such as 'ChatGPT'), other than for proofreading.	✓
I acknowledge that I have read all the texts listed in the bibliography or references list. I confirm all sources cited in the work appear in the bibliography or references list at the end.	✓

Signed: Sergiu Ionut Pascaru

Areas for development identified at draft stages by feedback given by tutors	**How I have addressed this feedback in this final submission**
1.	
2.	
3.	

Declaration of AI use in this academic work

It is important to acknowledge the use of AI tools in your work and to be clear about how you have used them. It is generally considered poor academic practice to include any content generated by GenAI tools in your work, unless you are explicitly reflecting on, analysing and evaluating the output of the software.

You should put an '**x**' against any statement in the list below that you know is TRUE to honestly declare how you have used AI in your work:

Statements

[ ] I can confirm that I have not used any AI tools within this piece of work.

[x] I have used AI tools to help me with developing ideas.

[x] I have used AI tools to help formulate a plan or structure for my assessment.

[x] I have used AI I tools to assist me with my research/gathering information.

[ ] I have used AI tools to help me understand key theories and concepts.

[x] I have used AI tools to identify themes as part of my data analysis for my work.

[ ] I have used AI tools to give me feedback on a draft piece of work.

[ ] I have used AI tools to proofread my work for correct spelling and grammar.

[ ] I have used AI tools to generate tables, figures or diagrams.

A Machine Learning Approach to Real-Time Phishing URL Detection: A Comparative Analysis of Classification Models

Student Name: Sergiu Ionut Pascaru
Student ID: 2310-111729
Module Code: CMP600
Module Title: Dissertation
Submission Date: 15.12.2025

Table of Contents

1. Abstract

The primary purpose of this study is to enhance the cybersecurity defence mechanism by developing a real-time system capable of detecting phishing URLs with a high level of accuracy. The blacklists have been lagging behind the advanced phishing techniques. This limitation demonstrates the necessity of automated detection systems that would be able to adjust to phishing methods developed. The purpose of this proposal is to describe a quantitative study that will contrast the performance of three machine learning classifiers, in this case, Logistic Regression, Random Forest and Support Vector Machine (SVM) using lightweight lexical features. The process was founded on Cross-Industry Standard Process of Data Mining (CRISP-DM) framework and the secondary data analysis. The study will also aim at identifying the most optimal model that will provide the most appropriate trade-off between high detection rates and low False Positive rate to make sure that the user has confidence in the model. Agile principles that will be managed with the help of Trello will guarantee orderly development and on-time completion of the project.

2. Introduction

One of the most common threats to online security is phishing, and the global economy incurs billions of dollars in costs to address this particular issue. The drawbacks of the traditional security measures are further highlighted by the fact that the blacklist-based type of filters can have high rates of latency and, consequently, users can become victims of a so-called zero-day phishing sites between the deployment and detection. With this proposal in place, it is now evident that the existing detection systems lead to a computational overhead problem because most of them rely on slow content-based analysis. The purpose of the project is to design and test a phishing URL detector software with the help of a well-organized experimental design and detailed technical documentation. It targets at a scientifically valid comparison of the classification algorithms and providing the one that would suit best in a real-time and client-side, browser environment where the speed is of the essence and the other factors of accuracy are secondary.

3. Background Study

This paper will be set against the context of artificial intelligence and network protection. The available evidence indicates that machine learning can be very efficient at URL classification, although most models rely on the so-called heavy features, i.e. the extraction of HTML content, or WHOIS lookups, which are not tolerable in real-time applications. The reason why this study is necessary is because there is the need to maximize the rate at which detection is made without compromising the accuracy. Despite the experimentation of different algorithms separately, the recent literature does not provide a direct comparison between a linear model and a non-linear model, which is constrained to lexical attributes (attributes obtained out of the URL string alone). This research will fill this gap by providing an empirical analysis of the Logistic Regression, Random Forrest, and SVM with the following limitations, which can result in faster, more privacy security products.

4. Research Questions

The research question in this study will be as follows:

To which extent are machine learning classifiers (Logistic Regression, Random Forest, and SVM) with only lexical URL features helpful to distinguish between valid and phishing URLs in a real-time detection scenario?

To answer this question, the following sub-questions are developed:

What is the best F1-score and True Positive Rate of which of the three models?

Which one has the minimum False Positive Rate (FPR), thereby producing a minimum level of disturbance to the legitimate user activity?

How does the exclusion of host based features affect the computational time, which is required to produce classification?

5. Aims and Objectives

5.1 Aim

This project will design and test a low-latency phishing URL detection system that will solely be based on URL lexical features. The emphasis on lightweight characteristics can be seen in practical limitation: the ability to detect in real-time without the need to be connected to external networks through queries or webpages content analysis.

5.2 Objectives

The measurable and the particular objectives of this project are:

1. **Data Acquisition:** The data to be acquired and pre-processed is Malicious URLs dataset (Siddhartha, 2024) in Kaggle. This data set has more than 650,000 URLs classified as benign, phishing, defacement and malware, which is a good and versatile source of data to train and test the model.
2. **Feature Engineering:** To obtain a Python based extraction engine that will compute lexical attributes (e.g., entropy, path length, special characters) of raw URL strings.
3. **Model Development:** : To develop, train, and hyperparameter-optimize three different classifiers: Logistic Regression, Random Forest, and SVM.
4. **Evaluation:** In order to evaluate the model performance statistically using Confusion Matrices, ROC curves and F1-scores.
5. **Project Management:** To apply Agile project management skills using Trello to schedule sprints, work and conduct retrospectives.

6. Scope of the Study

6.1 Specific Requirements and Expected deliverables.

As specified in the assessment brief, this project will deliver:

- **Project Documentation:** Single Word/PDF Documentation of contextual research, contextual design, methodology, security issues, and evaluation (around 7,000 words).
- **Project Product:** An executable software prototype (uploaded to some repository like GitHub).
- **Viva Voce:** A presentation to justify the project findings.

6.2 Limitations

The research is limited by the following restrictions:

- **Secondary Data:** The research will be based on an already prepared anonymized data, it will not capture any real network traffic.
- **Feature Constraint**: Lexical features are analysed to prevent analysis during deep content analysis and DNS features are analysed to perform in real-time.
- **Model Selection:** Only three algorithms have been found to compare them.

7. Research Methodology

7.1 Methodology Overview

The research design that will be used in the study will involve a quantitative research design that will involve studying secondary data and the Cross-Industry Standard Process of Data Mining (CRISP-DM) paradigm. The methodology of the research is outlined on the six stages of CRISP-DM:

1. **Business Understanding:** There is a need to establish the purpose of low-latency detection.
2. **Data Understanding**: Data exploration of the Kaggle dataset.
3. **Data Preparation:** Cleaning /Extracting features (changing text URLs to numeric vectors).
4. **Modelling:** This involves training of the classifiers that are chosen.
5. **Comparison**: Statistically comparing test data
6. **Deployment:** Prototype and documentation.

   For consistency with the research objectives, the multi-class labels in the dataset will be converted into a binary classification problem, where benign URLs form the negative class and phishing, defacement, and malware URLs are grouped into a single malicious class. This change is based on real-life deployment environment where the main focus is to distinguish between safe and potentially malicious URLs.

7.2 Justification

In order to objectively compare performance measures (Accuracy, Precision, Recall), a quantitative approach must be used. The reason as to why CRISP-DM is chosen is

because it is the industry standard of data mining, which ensures that there is an organised lifecycle. In addition, the implementation of the project will use an Agile system, which is based on the Trello tool, as the assessment brief would demand it to ensure an iterative development and reflective practice. To represent realistic real-time deployment conditions, runtime performance will be quantified as an average time needed to access one URL, both feature extraction and model inference.

8. Ethics

8.1 Data Sources and Collection

The dataset that is used in the study is Malicious URLs dataset (Siddhartha, 2024) available on Kaggle. The dataset is more than 650,000 labelled URL records, which is made available as an open-access, anonymised secondary data source to be used in research and education. As the dataset does not contain any personal or identifiable information and does not involve direct interaction with human participants, the project does not require participant recruitment, informed consent, or primary data collection.

8.2 Data Security and Analysis

Despite the fact that the dataset contains a mix of malicious URLs, the risk can be assessed as low.

- **Security:** URLs will be treated as plain text in a Python offline system.
- **Analysis:** It is a pure statistical analysis. No personal data will be processed and the NCG Research Ethics Policy and Procedures will be applied.

9. Research Plan

9.1 Project Management Approach

Project management approach entails employing project management methods to supervise and control project tasks to ensure delivery of both quality and quantity outputs

- **Planning Cards:** To draw the mapping of work in the start of each sprint.

- **Task Cards:** Tasks that are under the backlog.
- **Retrospective Cards:** To check the progress at the end of each sprint.

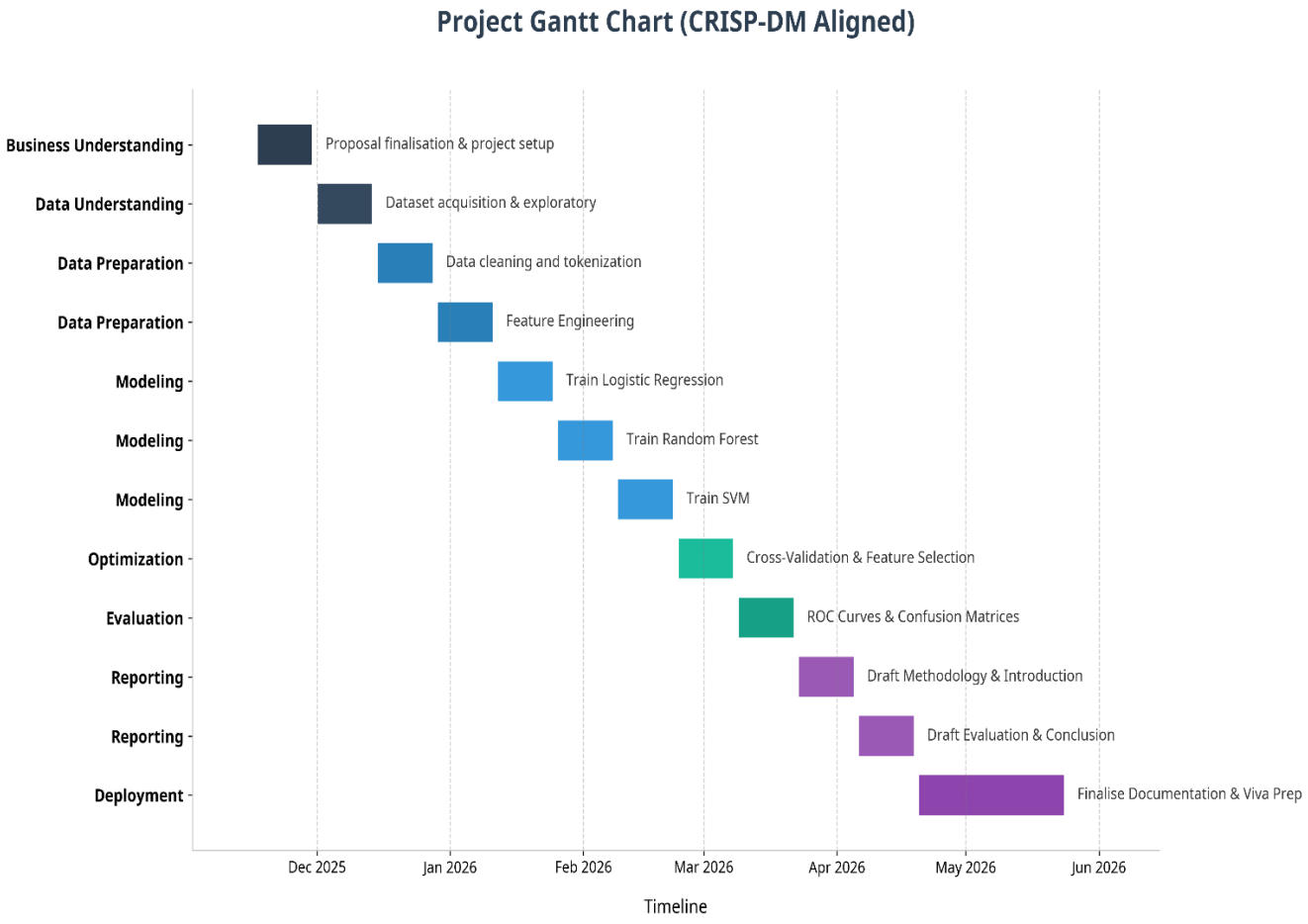
9.2 Sprint Schedule

The operational period runs from mid-November 2025 to May 24, 2026.

Sprint	Dates	Phase (CRISP-DM)	Objectives & Supervisor Interaction
Sprint 1	Nov 17 – Nov 30	Business Understanding	**Obj:** Setup GitHub/Trello. Finalize Proposal. **Sup:** Tutorial 1 (Proposal Review).
Sprint 2	Dec 01 – Dec 14	Data Understanding	**Obj:** Acquire Kaggle dataset. Perform Exploratory Data Analysis (EDA).
Sprint 3	Dec 15 – Dec 28	Data Preparation	**Obj:** Data cleaning and tokenization. **Sup:** Tutorial 2 (Data Check).
Sprint 4	Dec 29 – Jan 11	Data Preparation	**Obj:** Feature Engineering (Lexical extraction: Entropy, Length, Special Chars).
Sprint 5	Jan 12 – Jan 25	Modeling	**Obj:** Train Logistic Regression (Baseline). **Sup:** Tutorial 3 (Methodology Review).
Sprint 6	Jan 26 – Feb 08	Modeling	**Obj:** Train Random Forest (Ensemble tuning).
Sprint 7	Feb 09 – Feb 22	Modeling	**Obj:** Train SVM (Kernel optimization). **Sup:** Tutorial 4 (Model Progress).
Sprint 8	Feb 23 – Mar 08	Optimization	**Obj:** Cross-Validation (k-fold) and Feature Selection.
Sprint 9	Mar 09 – Mar 22	Evaluation	**Obj:** Generate ROC Curves, Confusion Matrices. Analyse FPR. **Sup:** Tutorial 5 (Results Analysis).

Sprint	Dates	Phase (CRISP-DM)	Objectives & Supervisor Interaction
Sprint 10	Mar 23 – Apr 05	Reporting	**Obj:** Draft Methodology and Introduction chapters.
Sprint 11	Apr 06 – Apr 19	Reporting	**Obj:** Draft Literature Review and Results chapters.
Sprint 12	Apr 20 – May 03	Reporting	**Obj:** Draft Discussion and Conclusion. **Sup:** Tutorial 6 (Final Draft Review).
Sprint 13	May 04 – May 17	Reporting	**Obj:** Final proofreading (7,000 words), Formatting, Turnitin pre-check.
Sprint 14	May 18 – May 24	Deployment	**Obj:** Final Submission (Doc + Product). Viva Prep.

Figure 1:



The project schedule can be viewed in Figure 1 where CRISP-DM methodology is used to organize it. The first thing on the list will be business understanding and proposal finalisation, then data acquisition and exploratory analysis will occur to verify that a person is familiar with the data before feature engineering. The development of the model is done stepwise, beginning with a baseline Logistic Regression model, and followed by more sophisticated classifiers, to enable the evaluation of performance to be done in stages. The issues of hyperparameter optimisation and evaluation are planned once all the models are implemented to provide equal comparison under equal conditions. The last stage provides enough time to write the reports in a structured manner, revise them, and format after which they are finally submitted and viva prepared. This performance methodology facilitates cyclic development without deviating on academic and delivery needs.

10. Bibliography

1. Chapman, P. et al. (2000) *CRISP-DM 1.0: Step-by-step data mining guide*. CRISP-DM Consortium.

2. Ma, J., Saul, L.K., Savage, S. and Voelker, G.M. (2009) 'Beyond blacklists: Learning to detect malicious web sites from suspicious URLs', in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1245–1254.

3. NCG (2025) *CMP600 – Dissertation Assessment Brief*. Newcastle College Group.

4. Sahingoz, O.K., Buber, M., Demir, O. and Diri, B. (2019) 'Machine learning based phishing detection from URLs', *Expert Systems with Applications*, 117, pp. 227–237.

5. Siddhartha, M. (2024) *Malicious URLs dataset: Classifying URLs as benign or malicious*. Kaggle. Available at: https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset (Accessed: 22 December 2025).

11. Appendices

11.1 Ethical Form Content

RESEARCH ETHICS APPROVAL FORM

Researcher	Sergiu Ionut Pascaru
Email address	2310-111723@elizabethschool.com
School	Newcastle College Group
Project Title	A Machine Learning Approach to Real-Time Phishing URL Detection: A Comparative Analysis of Classification Models
Names and email addresses of other researchers	N/A.
Proposed timescale of the project (MM/YY – MM/YY)	10/25 – 05/26

Aims of the project
To design, develop and test a real time machine learning system that can detect phishing URLs using lexical features. The target of the project is to compare the work of three classifiers (Logistic Regression, Random Forest, SVM) with a certain emphasis on the reduction of the False Positive Rate to increase the user experience and confidences in automated security tools.
Project design overview
The research is a secondary data analysis research that uses the Cross- Industry Standard Process of Data Mining (CRISP -DM) methodology to the quantitative data. The dataset employed by the study is the Malicious URLs dataset (Siddhartha, 2024) that is hosted on Kaggle and contains anonymised records of the URLs that are publicly accessible to study and educate the consumers. The data will be trained and tested supervised machine learning models on offline Python environment. The project will deal with feature engineering, model training, hyperparameter optimisation, and statistical performance evaluation. Neither does it involve live user monitoring, network traffic interception, or primary data collection of human subjects.

Total number of participants (individuals taking part in the research)	0(Zero)

Selection method for participants
N/A - Secondary Data Analysis of an existing dataset.

State the nature and source of any methodological advice revised in the planning of the project	Peer-reviewed literature about machine learning in the context of cybersecurity (e.g., Sahingoz et al., 2019) and the CRISP-DM guide to data mining project were used as a methodology guide. The appropriateness of the secondary data that is pre-validated was confirmed by supervisor guidance.
Procedures carried out on the participants	N/A.
Potential risks to participants and measures taken to address them	None. The data is completely anonymized and located in a public place. It does not involve any human subjects.
Potential risks to researchers and measures taken to address them	The dataset I will be working with will contain malicious URLs. Mitigation: 1.There will be no clicking of the URLs and visiting the URLs in a browser. 2. The entire analysis will be performed in an isolated Python environment. 3. The malicious domains will not be connected to any active network without permission.
Potential risks to bystanders and measures taken to address them	None.
Will informed consent be received from all participants prior to their involvement? Yes ☐ No ☒	
If not, please explain	The study is based on the publicly available and open-source dataset (Kaggle) in which data is already collected, anonymized, and can be used publicly/research. There is no recruiting or involvement of new human subjects.
Outline how records of consent will be stored	N/A - Secondary Data.
List other institutions or bodies involved in the project	None.
Outline any payments to be made to participants	None.
Identify any funding sources external to	None.

the institution, including any associated conditions/restrictions	
Outline how the findings of your research will be disseminated	The findings will be submitted as a dissertation report to Newcastle College Group

Supporting Documents

Document	Tick the box if applicable
Research proposal	☒
Information provided to participants	☐
Consent forms (including approval from parents/guardians/gatekeepers)	☐
Participant recruitment materials	☐
Draft data collection tools (e.g. questionnaires, surveys, interview plans)	☐
Supporting evidence from other institutional ethics committees	☐
Risk assessment form	☒
CRB reference number	☐

Requirements

1. I have completed all of the required research ethics training ☒
2. I have read the NCG Research Ethics Policy and Procedures ☒
3. I am aware of my legal, professional and statutory obligations and understand where to seek further guidance if required ☒
4. I will notify the SEC if the scope of my project changes significantly during its operational phase ☒
5. I understand that my research cannot commence until I have received full ethical approval ☒

Researcher signature: Sergiu Ionut Pascaru

Date: 22.12.2025

Checklist

Will the research project involve…	Yes	No
Contact with human participants, either directly or indirectly	☐	☒
Consent being sought from the National Research Ethics Service	☐	☒
Vulnerable groups (e.g. children, those with cognitive impairments)	☐	☒
Involvement of gatekeepers for access to participants (e.g. school children, care home residents)	☐	☒
Colleagues or staff as participants	☐	☒
Participants being studied in any manner without their knowledge or consent	☐	☒
Sensitive topics (e.g. alcohol or drug use, sexual behaviour)	☐	☒
Significant risk to either the researcher, participants or bystanders	☐	☒
Potential for psychological stress, anxiety or loss of social standing as a result of the research or its dissemination	☐	☒
Prolonged or repetitive testing	☐	☒
Members of the public in a data collection capacity	☐	☒
Access to existing data which requires the consent of an external body before use	☐	☒
Potential for participants to be identified during dissemination by compromising anonymity or confidentiality	☐	☒
External research partners	☐	☒
External funding	☐	☒
Payments to participants other than for reasonable expenses	☐	☒
Sharing of data beyond the scope of the initial consent from participants	☐	☒
CRB check	☐	☒

To be completed during the ethical approval process

STAGE 1	
Approved	☐
Approved, subject to specified amendments	☐
Referred to the SEC	☐
Declined	☐
I have informed the researcher in writing of the outcome of this process (please attach a copy)	
Name: Signed: Position: Supervisor/Chair of the SEC (please circle) Date:	

STAGE 2	
Approved	☐
Provisional approval	☐
Approved, subject to specified amendments	☐
Referred to the REC	☐
Declined	☐
I have informed the researcher in writing of the outcome of this process (please attach a copy)	
Name: Signed: Position: Chair of the SEC / Chair of the REC (please circle) Date	