Sri Lanka Institute of Information Technology

PROJECT REGISTRATION FORM

(This form should be completed and uploaded to the Cloud space on or before XXXXXXXXX)

The purpose of this form is to allow final-year students of the B.Sc. (Hon) degree program to enlist in the final-year project group. Enlisting in a project entails specifying the project title and the details of four members in the group, the internal supervisor (compulsory), the external supervisor (may be from the industry), and indicating a brief description of the project. The description of the project entered on this form will not be considered as the formal project proposal. It should however indicate the scope of the project and provide the main potential outcome.

PROJECT TITLE (As per the accepted Topic Assessment Form)	Detecting email-based phishing websites using machine learning

RESEARCH GROUP (As per the Topic Assessment Form)	Artificial Intelligence and Machine Learning

PROJECT NUMBER		(Will be assigned by the RP Team)

PROJECT GROUP MEMBER DETAILS: (Please start with the group leader's details)

	STUDENT NAME	STUDENT NO.	CONTACT NO.	EMAIL ADDRESS
1	J.M.Lindamulage	IT20222840	0703372025	it20222840@my.sliit.lk
2	Mandira Pabasari L.	IT19966236	0702050537	it19966236@my.sliit.lk
3	Yapa S.P.J.	IT20050108	0717670212	it20050108@my.sliit.lk
4	Perera I.S.S.	IT20222468	0769251191	it20222468@my.sliit.lk

SUPERVISOR, CO_ SUPERVISOR Details

SUPERVISOR Name	CO-SUPERVISOR Name
Ms.Jenny Krishara	Ms.Madhuka Nadeeshani Koralage
Signature	Signature
Jenny 13/03/2023	13/03/2023

EXTERNAL SUPERVISOR Details (if any, may be from the industry)				
				Attach the email as Appendix 3
Name	Affiliation	Contact Address	Contact Numbers	Signature/Date

ACCEPTANCE BY CDAP MEMBER (This part will be filled by the RP team)		
Name	Signature	Date

PROJECT DETAILS

Brief Description of your Research Problem: (extract from the topic assessment form)
Phishing has become one of the most common cyberattacks today. Attackers launch phishing attacks now in a variety of ways, like spoof calls, messages, social networking, and emails, with the intention of obtaining sensitive information from victims. The most common way victims fall for phishing is by clicking on a fraudulent link in an email that was sent by an attacker. This is called email phishing. This link will take victims to websites that appear legitimate but are run by the attackers. Attackers construct illegal email accounts using real company details and send the email to victims, impersonating a real person, while making the victim click on the link to the fraud website using social engineering. If a person clicks on the link and visits the site and tries to login or perform any action, then the attackers can get passwords, login credentials, credit card information, and other sensitive information. With the COVID-19 outbreak, phishing activities have also increased. Even though there are many tools and research studies available, phishing is so widespread that no single solution can reduce all the vulnerabilities. Many people fall victim to this, even though these tools are available, due to a lack of knowledge and the inability to purchase such tools.

Main expected outcomes of the project: (extract from the topic assessment form)
The solution is to develop a web extension and a mobile application that can detect phishing sites and phishing emails. There are a few approaches that are going to be used to identify a phishing site. A content-based approach is based on website content, and it is used to identify the website. In the non-content approach, the URLs, host information is used to identify the website. In a visual-based approach, visual elements are used to identify the website. Text analysis can be used in emails that are sent by attackers with urgent requests but contain grammar and spelling errors. Email headers can be analyzed to check if the sender's address is genuine. At last, the network behavior, which can be described as traffic patterns and connections, can be used to determine if the IP address from which the connection was made is phishing or not. Using all the above methods, the final tool will be able to detect and notify the user about phishing websites based on email content, website URL, visual similarity features, and network behavior.

WORKLOAD ALLOCATION (**extract from the topic assessment form after correcting the suggestions given by the topic assessment panel.**)

(Please provide a brief description of the workload allocation)

MEMBER 1	J.M.Lindamulage IT20222840

- To develop a phishing website detection model using Convolution Neural Networks (CNN) or Graph Neural Networks (GNN)-based approach using visual similarity features between legit and phishing websites.
- To optimize inference time and model size of the developed deep learning model using optimization techniques such as model pruning and weight clustering for edge computing and integrate the optimized model into an Android mobile application.
- Find how visual features such as text content, background colors, images can be used as useful features to differentiate phishing sites and identify datasets that contain those features, such as PhishTank.
- If datasets are not sufficient, collect data by web scraping from domain names included in datasets such as the AlexaTop dataset.
- Conduct experiments with different CNN architectures, such as EfficientNet and MobileNet, and observe the relationships between features to be able to model this problem as a Graph Neural Network.
- Develop the models, compare their performances, and select the best-performing model amongst them.
- Try to optimize the size and latency caused by the selected model by applying model optimization techniques such as model pruning and weight clustering and integrating the model into the mobile application.

MEMBER 2	Mandira Pabasari L. IT19966236

- Collect and analyze the relevant network traffic features associated with a website, such as the source and destination IP addresses, protocols, and other data, to determine the site's classification as malicious or benign.
- Pre-process the collected data by cleaning and transforming it into an appropriate format for machine learning algorithms to ensure the accuracy and suitability of the analysis.
- Apply machine learning algorithms like Logistic Regression, Support Vector Machine, k-Nearest Neighbor, or Principal Component Analysis to classify websites based on their network traffic features.
- Train the machine learning model using a labeled dataset of previously classified websites to improve accuracy and effectiveness.
- Leverage the trained model to provide real-time protection against potentially malicious websites by issuing warnings or blocking access as needed. This process can provide essential protection against cyber-attacks and other online threats.

MEMBER 3	Yapa S.P.J. IT20050108

- To develop a machine learning-based tool for detecting phishing emails based on their header and content.
- To develop Convolution Neural Networks (CNN) or Long Short-Term Memory (LSTM) model that can identify potential phishing emails by analyzing the sentiments of the email content.
- Explore literature to find out what are the features and patterns which can be considered to classify the sentiment of the sentences.
- Explore what type of machine learning and deep learning techniques which has been used to recognize sentiments which indicates the email is a phishing email.
- Conduct experiments with different features with recognized deep learning models which can classify sentiments.
- Datasets that use for this are collected from Kaggle.
- Select the best models and develop pipeline to extract text from emails which user reads an email and analyze the content.
- Develop browser plugin and integrate the develop the AI pipeline.
- Combine natural language processing models with sentiment analysis to develop ml model which can analyze the sentiments of an email and classify if the email is a phishing email or not.
- Develop browser plugins which can run in the background and analyze the content of emails and provide warnings to user about any potential phishing email.

MEMBER 4	Perera I.S.S. IT20222468

- To develop a tool to identify phishing websites using URL and text content using machine learning.
- Implement a Python program to extract features from the URL.(feature extraction).
- Using a data set from the Kaggle web site (a phishing dataset for machine learning).
- Study about address bar features, HTML and Java Script features, and domain features. Study classical machine learning techniques like Random Forest, K nearest neighbors, Decision Tree, Linear SVC classifier, one class SVM classifier and wrapper-based features selection, which contains the metadata of URLs, and use the information to determine if a website is legitimate or not.
- Check the phishing, malicious URLs using machine learning techniques and block the URLs using local host file.
- Check the phishing , malicious URLs using machine learning techniques and block the URLs using local host file.

DECLARATION (Students should add the Digital Signature)

"We declare that the project would involve material prepared by the Group members and that it would not fully or partially incorporate any material prepared by other persons for a fee or free of charge or that it would include material previously submitted by a candidate for a Degree or Diploma in any other University or Institute of Higher Learning and that, to the best of our knowledge and belief, it would not incorporate any material previously published or written by another person in relation to another project except with prior written approval from the supervisor and/or the coordinator of such project and that such unauthorized reproductions will construe offences punishable under the SLIIT Regulations.

We are aware, that if we are found guilty for the above mentioned offences or any project related plagiarism, the SLIIT has right to suspend the project at any time and or to suspend us from the examination and or from the Institution for minimum period of one year".

	STUDENT NAME	**STUDENT NO.**	**Signature**
1	J.M.Lindamulage	IT20222840	Judith
2	Mandira Pabasari L.	IT19966236	
3	Yapa S.P.J.	IT20050108	
4	Perera I.S.S.	IT20222468	