**Project Introduction Form**

**Team Members:**

– Hadeer Amr Fawzy (22010450) cybersecurity

– Farida Ahmed (2206160) cybersecurity

– Muhammed Salah (22010448) cybersecurity

–Marwan Gaber Ramdan (2206167) cybersecurity

– Omar Ibrahim Ahmed (2206209) cybersecurity

– Youssef Tamer Muhammed Ali (2206208) cybersecurity

– Ahmed Saber Ahmed (2203185) ai

**Project Title**: **AI-Based Phishing Detection A Machine Learning Approach for Cybersecurity**

**Abstract**

**Background**: Phishing attacks remain a major cybersecurity threat, tricking users into revealing sensitive data. Traditional detection methods (e.g., blacklists) fail against evolving tactics.

**Purpose**: Develop an ML-based classifier to detect phishing websites in real time.

**Methods**: Use supervised learning (Random Forest, SVM) on a dataset of legitimate and phishing URLs.

**Expected Results**: High accuracy (~95%) in classification.

**Expected Conclusion**: ML can effectively enhance phishing detection compared to rule-based methods.

**Research Problem:**

How can machine learning improve the detection of phishing websites compared to traditional methods? **Research Question:**

Which machine learning model achieves the highest accuracy in classifying phishing websites? **Project Literature Review:**

1. <https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms>
2. <https://www.researchgate.net/publication/379263290_PHISHING_DETECTION_USING_MACHINE_LEARNING-A_MODEL_DEVELOPMENT_AND_INTEGRATION/link/6646163e0b0d284574374927/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19>
3. <https://www.researchgate.net/publication/371538674_Comparison_of_Support_Vector_Machine_SVM_and_Random_Forest_Algorithm_for_Detection_of_Negative_Content_on_Websites>

**Your hypothesis which you will contribute with proving or disproving it:**

A machine learning model (specifically Random Forest) trained on a combination of URL structural features (length, special characters, domain age) and content-based features (JavaScript obfuscation, presence of login forms) can achieve >95% accuracy in detecting phishing websites, outperforming traditional blacklist-based methods by at least 20% in identifying zero-day phishing attacks.

**The domain you will be working on:**

**(Domain Four)**

Cyber defense using AI

**List datasets you will be using:**

1. phishing dataset from UCI https://archive.ics.uci.edu/dataset/327/phishing+websites
2. OpenPhish live feed for testing <https://www.openphish.com/>
3. [Mendeley Phishing Dataset](https://data.mendeley.com/datasets/h3cgnj8hft/1" \t "_blank) (adds more features like JavaScript tricks)
4. PhishStorm https://research.aalto.fi/en/datasets/phishstorm-phishing-legitimate-url-dataset

**List the AI methods you will be using:**

1. Random Forest
2. SVM (Support Vector Machine)
3. Logistic Regression (for baseline comparison)

**List type of AI cyber-attack or regular cyber-attacks you will be dealing with:**

1. Phishing (URL-based deception)
2. URL Spoofing as paypa1.com instead of paypal.com
3. Fake Login Pages stealing passwords

**List all tools/software programs/operating system you will be using:**

1. Python (Scikit-learn, Pandas)
2. Jupyter Notebook
3. Wireshark (network traffic analysis)