Program Profile		
Program	Program name	Computer Science and Engineering
	Category	B3

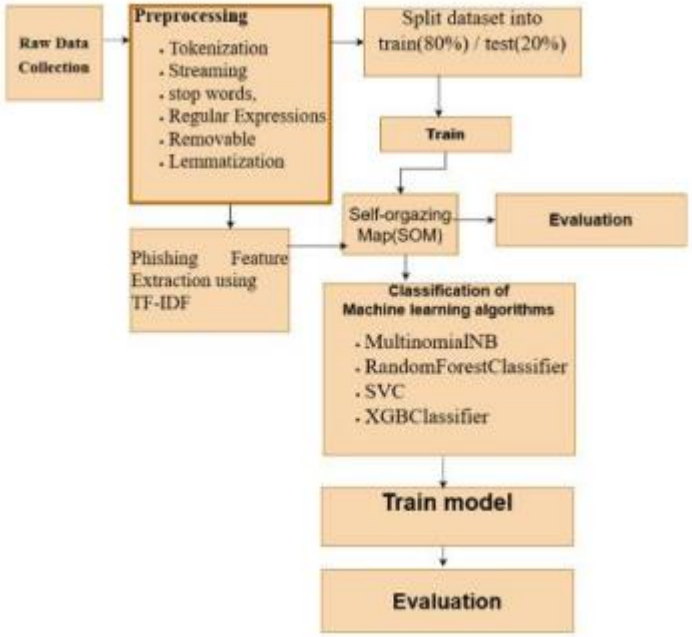
Summary of Program	
Program Name	Computer Science and Engineering
Category	B1
Abstract of Program	Empirical study attempts to analyze major trends of phishing Uniform Resource Locator detection, emphasized on improvement of 80% training data on feature extraction using term frequency-inverse document frequency (TFIDF)based on self-organizing map with classification machine learning algorithm. To analyze characteristics of phishing Uniform Resource Locator term frequency-inverse document frequency (TFIDF) is implemented for phishing feature extraction incorporated with self-organizing mapping (SOM). Significance of a phishing features in URL are subjected to data frame depending on quantity and quality of malicious attacks. For advanced classification techniques support vector classifier,Multinomial naïve Bayes, Random Forest Classifier, Extreme Gradient Boosting(XGB) Classifiers are compared using python programming language to determine best performance with proposed model. Essential comparison performance data used for risk prediction in order to select best classification model incorporated with proposed model. Proposed framework is able to achieve a superior accuracy with Kappa performance indicators of score 0.686 which considered as moderate value. Suggested method is based on URL features containing entropy, address bar based features, IP-based features, and URL-based features. Investigation provides significance mean accuracy of 97.5% using Support Vector Classifier (SVC). Proposed model benchmark 10 datasets each containing 11430 URL (50% legitimate and 50% phishing) based on 87 features. Other classifiers produced an accuracy rate of Multinomial naïve Bayes (84% accuracy), Random Forest Classifier (93%accuracy), Extreme Gradient Boosting (XGB) Classifier (96% accuracy). Study limits the use of more phishing features in the dataset

Details of Program		
Planning		
Objectives	Long-term Goals	Phishing websites pose significant cybersecurity risks and evolve constantly. Existing rule-based models often struggle to adapt to new attack vectors. This study investigates the role of TF-IDF-based textual analysis and SOM-based

		clustering in understanding and classifying phishing URL structures.
	Short-term Targets	To create a scalable and accurate model for phishing URL detection using SOM and TF-IDF. The model enhances real-time threat identification using AI-driven techniques and enables proactive defense against evolving cyber threats.
	Rationale	Team lead
Subject (Leader)	Initiator(s)	Md.Nazmus Sakib
	Champion(s)	Sakib,Md.Nazmus
	Major team member(s)	Md Nazmus Sakib, Kazi Hassan Robin, Ayesha Siddika, and Shamsun Nahar
Environment	Nature/Society	Relevant to financial institutions, e-commerce, and IT service providers vulnerable to phishing scams.
	Industry/Market	IT industry
	Citizen/Government	Citizen
Resources	Human resources	Yes
	Financial resources	No
	Technological resources	yes
Mechanism	Strategy (Weight/Sequence)	Weight
	Organization	World University of Bangladesh
	Culture	This is supported by university culture
Doing		
Launch date		12/12/2026
Responsible organization		World University of Bangladesh
Program content and process		To successfully develop, implement, and scale the phishing URL detection model, a multidisciplinary team is essential: Cybersecurity Specialists • Experts in phishing techniques, threat intelligence, and malware behavior. 3 • Provide domain knowledge for selecting relevant phishing features and validating model predictions. Machine Learning Engineers • Design and optimize classification algorithms such as SVC, Random Forest, and XGBoost.

	• Responsible for feature engineering using TF-IDF and SOM integration, hyperparameter tuning, and model evaluation. Data Scientists • Handle data preprocessing, regular expression cleaning, tokenization, lemmatization, and visualization. • Conduct statistical analysis and interpret confusion matrices, accuracy scores, and Kappa metrics. Software Developers / System Integrators • Build API endpoints, dashboards, or browser plugins to deploy the model for real-time phishing detection. • Ensure compatibility with cloud environments and scalability for real-time streaming data. DevOps/Cloud Engineers • Set up the infrastructure for real-time model deployment and continuous integration/continuous delivery (CI/CD). • Manage cloud-based compute resources for model training and inference. Research Assistants / Interns • Support data collection from phishing repositories and assist in maintaining up-to-date datasets. • Help in documentation, literature review, and benchmarking against recent models. Project Manager / Coordinator • Coordinate efforts across disciplines and ensure timely completion of milestones. • Facilitate collaboration with academic, industry, and policy partners.
Key highlights of the content/process	Feature extraction using TF-IDF • SOM training • Classification model training • Validation using confusion matrix • Accuracy and Kappa evaluation
Differences from traditional approaches	Traditional models rely on static blacklists. This model learns dynamic phishing patterns and adapts in real time using streaming and AI.
Progress as of today	Presently work has been partially completed and published
Problems in implementation	• Dataset availability • High model complexity • Real-time processing latency
Approaches to solve the problems	Using balanced datasets • Leveraging SOM for clustering • Incorporating streaming architecture for continuous learning
Completion date, if completed	Completed in October 2023 https://doi.org/10.1007/978-981-97-0126-1_23
Seeing	
Impacts on students	Enhanced skill development in cybersecurity, data mining, and machine learning.
Impacts on professors	The study equips professors with advanced phishing detection knowledge, enabling them to enrich cybersecurity curricula and guide student research on applied machine learning in security.

Impacts on university administration	The framework helps university administration strengthen institutional cybersecurity policies and defenses against phishing attacks, reducing risks to academic data and resources.
Responses from industry/market	The industry is likely to adopt the proposed model as a cost-effective cybersecurity solution to improve phishing detection accuracy, thereby enhancing consumer trust and reducing fraud losses.
Responses from citizen/government	Citizens benefit through safer digital experiences, while governments may endorse or integrate such models into national cybersecurity initiatives to strengthen public defense against phishing.
Measurable output (revenues)	The measurable output could include revenue growth for cybersecurity firms through commercial deployment of the model in enterprise security systems
Measurable input (expenses)	Expenses primarily involve computational resources for training/testing models, data acquisition, software tools, and expert manpower for deployment and monitoring.
Cost-benefit analysis for effectiveness	The framework demonstrates high accuracy (97.5% with SVC), meaning the benefits of reduced phishing attacks and enhanced data security significantly outweigh the relatively moderate costs of implementation and maintenance.
Future Planning	
Where does the project go from here?	**Expanded Feature Engineering** Incorporate additional phishing indicators such as DNS-based features, WHOIS data, SSL certificate analysis, and behavioral clickstream data to improve robustness. **Larger & Real-Time Datasets** Move beyond the current 10 benchmark datasets to real-world, continuously updated URL streams, ensuring adaptability to evolving phishing tactics. **Hybrid & Deep Learning Models** Explore integration with deep learning architectures (e.g., CNNs, RNNs, Transformers) for automatic feature learning and improved classification performance. **Deployment in Practical Systems** Develop prototypes or browser plug-ins, email filters, or security dashboards that apply the model in real-time to protect end-users and institutions. **Policy & Collaboration** Partner with government agencies, universities, and cybersecurity firms to integrate the model into national awareness programs and institutional cybersecurity frameworks. **Cost-Benefit & Scalability Studies** Conduct longitudinal studies measuring cost savings and attack prevention rates when the system is deployed at organizational or enterprise levels.

Addendum	
Exhibits, pictures, diagrams, etc.	
Reports, mimeos, monographs, books, etc.	**Peer-Reviewed Research Articles** • Published papers on phishing detection using machine learning (e.g., TF-IDF feature extraction, SOM, and SVC) that align with and validate the methodology. • Comparative studies on classification models (SVM, Random Forest, XGBoost, Naïve Bayes) in cybersecurity contexts. **Technical Reports & Memos** • Internal project reports documenting dataset preparation, feature engineering processes, and classification results. • Validation reports comparing accuracy, kappa scores, and error rates across multiple datasets. **Monographs** • Extended academic works on phishing trends, machine learning in cybersecurity, and real-time URL threat detection. • University-published monographs that consolidate empirical findings into broader cybersecurity frameworks. **Books** • Reference texts on **Machine Learning in Cybersecurity**, **Data Mining for Security Applications**, and **Phishing Attack Mitigation** to substantiate the theoretical underpinnings of the project.

	• Handbooks on applied artificial intelligence and digital risk management for higher education and industry contexts.
Others which may help explain the program (including website links)	program can be further explained and validated through i, open-source phishing datasets such as uci https://archive.ics.uci.edu/dataset/327/phishing+websites