SLIIT UNI
THE KNOWLEDGE UNIVERSITY

# IT4010 – Research Project - 2024

# Topic Assessment Form

**Project ID :** **R24-102**

1. Topic (12 words max)

Enhancing and Developing Adversarial Robustness of Diseases Risk Predictors Against Multiple Adversarial Attacks.

2. Research group the project belongs to

**Computing Infrastructure and Security (CIS)**

3. Research area the project belongs to

**Cyber Security (CS)**

4. If a continuation of a previous project:

| | |
|---|---|
| Project ID | |
| Year | |

5. Brief description of the research problem including references (200 – 500 words max) – references not included in word count.

Machine learning (ML) models are becoming more and more important for making critical healthcare decisions, and their vulnerability to adversarial attacks raises major concerns [1]. As ML models are rapidly being adapted for critical healthcare tasks such as disease prediction and treatment recommendation systems, ensuring the robustness and security of these models against adversarial manipulation has become a paramount concern. Adversarial attacks that cause misclassification or misdiagnosis by ML models could have severe consequences for patient health.

It is important to recognize that ML models are susceptible to adversarial attacks and unanticipated biases as they grow more and more integrated into decision-making processes. Security breaches targeting ML models can lead to compromised predictions, loss of confidential information [2], and adverse financial consequences. Therefore, ensuring the security and resilience of ML models has become a fundamental consideration in their deployment since it helps to enhance predictions' reliability, develop trust among healthcare personnel's, and protect sensitive patients including Personal Identifiable Information (PII) and Protected Health Information (PHI) data integrity.

This research aims to improve the robustness of disease risk prediction models against adversarial attacks. Specifically, it will evaluate the impact of four main attack techniques - Projected Gradient Descent [3], Carlini-Wagner [4], Brendel-Bethge [5], and a Boundary attack [6] adapted for tabular data. By exposing risk prediction models to these attacks and quantifying resulting performance losses, we intend to identify effective defense mechanisms to counter the attacks. The goal is to develop models that are resilient against perturbations to input data, feature manipulation, and other adversarial techniques that could undermine reliability in real-world deployment. Defending against a diverse set of attacks is crucial to ensuring robust and trustworthy disease risk assessments.

By investigating the attacks and their corresponding countermeasures, our research directly contributes to the development of solid and trustworthy disease risk prediction models [7].

**References**

[1] A. Qayyum, J. Qadir, M. Bilal and A. Al-Fuqaha, "Secure and Robust Machine Learning for Healthcare: A Survey," in IEEE Reviews in Biomedical Engineering, vol. 14, pp. 156-180, 2021, doi: 10.1109/RBME.2020.3013489.

[2] M. Xue, C. Yuan, H. Wu, Y. Zhang and W. Liu, "Machine Learning Security: Threats, Countermeasures, and Evaluations," in IEEE Access, vol. 8, pp. 74720-74742, 2020, doi: 10.1109/ACCESS.2020.2987435.

[3] Yan Jiang, Guisheng Yin, Ye Yuan, Qingan Da, "Project Gradient Descent Adversarial Attack against Multisource Remote Sensing Image Scene Classification", Security and Communication Networks, vol. 2021, Article ID 6663028, 13 pages, 2021. https://doi.org/10.1155/2021/6663028

[4] Pujari, Medha & Cherukuri, Bhanu & Javaid, Ahmad & Sun, Weiqing. (2022). An Approach to Improve the Robustness of Machine Learning based Intrusion Detection System Models Against the Carlini-Wagner Attack. 10.1109/CSR54599.2022.9850306.

[5] S. Y. Khamaiseh, D. Bagagem, A. Al-Alaj, M. Mancino and H. W. Alomari, "Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification," in IEEE Access, vol. 10, pp. 102266-102291, 2022, doi: 10.1109/ACCESS.2022.3208131.

[6] Alrasheedi, F. and Zhong, X. (2023) Imperceptible adversarial attack on deep neural networks from image boundary, arXiv.org. Available at: https://arxiv.org/abs/2308.15344 (Accessed: 10 January 2024).

[7] Liang, H.; He, E.; Zhao, Y.; Jia, Z.; Li, H. Adversarial Attack and Defense: A Survey. Electronics 2022, 11, 1283. https://doi.org/10.3390/electronics11081283

6. Brief description of the nature of the solution including a conceptual diagram (250 words max)

Our research focuses on assessing adversarial attacks against cardiovascular disease prediction models in the area of cyber security, specifically applying adversarial attack techniques to evaluate vulnerabilities of prediction systems." [1]. We will first be using a convolutional neural network (CNN) algorithm to train the model on a balanced **structured (Tabular data) dataset of cardiovascular data**. This model will serve as the basis for evaluating the impact of different adversarial attack methods.

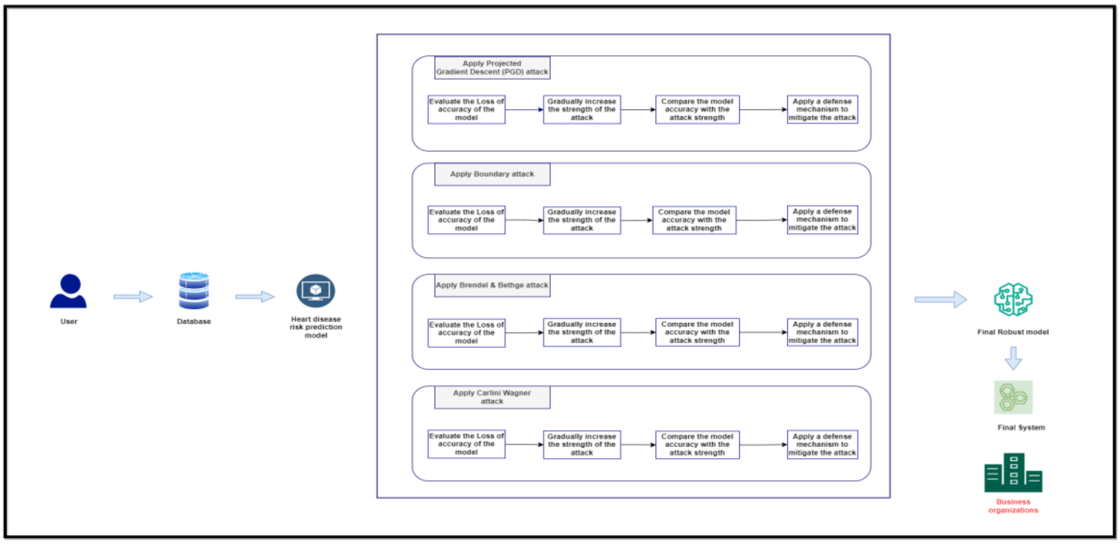
We will subject the trained CNN model to four prominent attack algorithms - Projected Gradient Descent (PGD), Carlini-Wagner (C&W), Brendel-Bethge, and Decision Boundary attacks. For each attack, we will systematically evaluate the loss in model accuracy compared to the original model. We will start with a low attack strength and then gradually increase it to analyze the impact on accuracy metrics like precision, recall, and F1-score. This comparative analysis will reveal how different attacks vary in their ability to affect model performance.

In addition, we will explore various defense strategies like adversarial training, defense distillation, and gradient regularization to mitigate these attacks. This will provide insights into effective countermeasures to strengthen model resilience. We will quantify the extent of accuracy recovery after applying defenses against each type of attack.

The key contribution of our research is developing an open-source adversarial attack evaluation framework. This tool will allow users to upload a trained model and dataset, select an attack method, and automatically evaluate the robustness by reporting accuracy metrics after attacks. Researchers can thus systematically test model vulnerability against PGD, C&W, Brendel-Bethge, and boundary attack. Our work will help build more reliable ML models for risk prediction in critical applications like healthcare.

**Reference**

[1] sedawei, O. (2023) CVD_CLEANED, Kaggle. Available at: https://www.kaggle.com/datasets/omersedawei/cvd-cleaned/data (Accessed: 02 January 2024).


User
Database
Heart disease risk prediction model
Apply Projected Gradient Descent (PGD) attack
Evaluate the Loss of accuracy of the model
Gradually increase the strength of the attack
Compare the model accuracy with the attack strength
Apply a defense mechanism to mitigate the attack
Apply Boundary attack
Evaluate the Loss of accuracy of the model
Gradually increase the strength of the attack
Compare the model accuracy with the attack strength
Apply a defense mechanism to mitigate the attack
Apply Brendel & Bethge attack
Evaluate the Loss of accuracy of the model
Gradually increase the strength of the attack
Compare the model accuracy with the attack strength
Apply a defense mechanism to mitigate the attack
Apply Carlini Wagner attack
Evaluate the Loss of accuracy of the model
Gradually increase the strength of the attack
Compare the model accuracy with the attack strength
Apply a defense mechanism to mitigate the attack
Final Robust model
Final System
Business organizations

7. Brief description of specialized domain expertise, knowledge, and data requirements (300 words max)

Expertise in cyber security is essential for this research, which examines adversarial attacks mounted against machine learning systems. Specifically, we aim to evaluate vulnerabilities in cardiovascular disease risk models that are developed using deep learning. By specializing in cyber security approaches tailored to machine learning contexts, we can thoroughly assess the robustness of prediction models to adversarial input data. A strong foundation in deep learning and neural network architectures is essential to understanding how adversarial attack methods like projected gradient descent, Carlini-Wagner, Brendel-Bethge, and decision boundary attacks fool prediction models. Expertise in generating and detecting adversarial examples is also needed to carry out robust evaluations. Additionally, domain knowledge of cardiovascular disease risk factors and modeling is critical to properly evaluate attacks against risk models and assess the clinical impact of degraded predictions.

The study necessitates access to cardiovascular disease datasets suitable for training risk prediction models, such as those containing demographic, lifestyle, and clinical biomarkers for patients. Models trained on these datasets, like neural networks, ensemble models, or other algorithms, can then be subject to adversarial attacks. Compute infrastructure for training the prediction models and generating the attacks is required.

Knowledge of model evaluation metrics is also important, including accuracy, AUC, sensitivity, specificity, and others relevant to clinical risk models. Statistical analysis skills are also beneficial to evaluate and compare the impact of different attacks. Overall, the key requirements are expertise in adversarial deep learning, cardiovascular risk modeling, software capabilities, and clinical data to conduct a rigorous comparative study of state-of-the-art adversarial threats against risk prediction models.

8. Objectives and Novelty

| Main Objective<br>**Assess the effectiveness of various adversarial attacks on cardiovascular disease (CVD) risk prediction models and attempt to increase model strength against the given attacks.** | | | |
|---|---|---|---|
| Member Name | Sub Objective | Tasks | Novelty |
| Yasuththara B.G.V | Apply the Carlini – Wagner attack against the model and identify accuracy changes. Test Defenses. | • Prepare dataset that contains input data which related to patients' behavior that indicates the prediction of Cardiovascular disease.<br>• Conduct Exploratory Data Analysis (EDA) in order to visualize and summarize the data to gain insights into its distribution, structure, patterns, and any potential issues or anomalies and perform data preprocessing and balancing before splitting up the data.<br>• Train and evaluate model accuracy.<br>• Craft targeted Carlini adversarial examples to cause misclassification of specific instances, forcing the model to wrongly predict patients to have or not have Cardiovascular disease. | **Apply Defense Distillation mechanism to the training process of a cardiovascular disease prediction model based on tabular data to reduce its vulnerability to Carlini attack.**<br>Defensive distillation hardens tabular data heart disease models against Carlini attacks. It trains a teacher model on clean data. The teacher generates soft probability labels on the training data which are used to train a |

| | | | |
|---|---|---|---|
| | | • Evaluate the global robustness of the tabular data model by attacking a large test set. Measure overall degradation in accuracy and precision/recall metrics.<br>• Understand which features/variables are most vulnerable to manipulation and distortion by the attack and why. Identify correlations. | student model, distilling the knowledge. Learning from soft probabilities makes the student's decision boundaries smoother and more robust. We test if the distilled student model has higher accuracy on new Carlini attacks versus the teacher, indicating improved defense against manipulated test data that exploits decision boundaries.<br>**Apply an Input randomization and reconstruction mechanism to the training process of a cardiovascular disease prediction model based on tabular data to reduce its vulnerability to Carlini attack.** Add an input reconstruction module that tries to reconstruct the original input from the model's internal representations. Compare the reconstructed input to the actual received input - |

| | | | |
|---|---|---|---|
| | | | significant discrepancies could indicate the input was manipulated.<br>This attack can be performed by feeding datasets similar to our chosen dataset but from various different domains to test the robustness of the model. |
| Weerakoon R. A. D. D. C | Apply the PGD attacks against the model and identify accuracy changes. Test defenses. | • Prepare dataset that contains input data which is related to patients' behavior that indicates the prediction of Cardiovascular disease.<br>• Conduct Exploratory Data Analysis (EDA) in order to visualize and summarize the data to gain insights into its distribution, structure, patterns, and any potential issues or anomalies and perform data preprocessing and balancing before splitting up the data.<br>• Train and evaluate model accuracy.<br>• Conduct baseline PGD attack.<br>• Assess attack success rate & perturbation magnitude. | **Apply the Gradient Regularization mechanism to modify the training process of a cardiovascular disease prediction model based on tabular data to reduce its vulnerability to PGD attack.**<br>Penalize strong changes in key data features when training a cardiovascular disease prediction model. This makes the model less dependent on these features, making it harder for attacks to manipulate predictions. By adjusting how the model learns |

| | | | |
|---|---|---|---|
| | | • Apply defensive strategies (e.g. adversarial training)<br>• Evaluate effectiveness against PGD attacks.<br>• Detect PGD attacks (Eg:- FGSM) | during training, we discourage overreliance on these sensitive features. Checking how well the model performs on normal and manipulated data helps us see if this technique makes the model tougher against attacks.<br>**Apply the Certified Defenses mechanism to modify the training process of a cardiovascular disease prediction model based on tabular data to reduce its vulnerability to PGD attack.**<br>Formal verification methods to prove robustness for all possible perturbations within a defined lp norm ball. This provides a certified defense guarantee for PGD within the verified epsilon radius.<br>This attack can be performed by feeding datasets similar to our chosen dataset but from |

| | | | |
|---|---|---|---|
| | | | various different domains to test the robustness of the model. |
| Karandawala D. N | Apply the Boundary attack against the model and identify accuracy changes. Test defenses. | • Prepare dataset that contains input data which related to patients’ behavior that indicates the prediction of Cardiovascular disease.<br>• Conduct Exploratory Data Analysis (EDA) in order to visualize and summarize the data to gain insights into its distribution, structure, patterns, and any potential issues or anomalies and perform data preprocessing and balancing before splitting up the data.<br>• Train and evaluate model accuracy.<br>• Apply boundary attack.<br>• Assess attack success rate & perturbation magnitude.<br>• Apply defense mechanisms tailored to tabular data.<br>• Evaluate the effectiveness of the model against Boundary attack. | **Apply Adversarial training mechanism to modify the training process of a cardiovascular disease prediction model based on tabular data to reduce its vulnerability to boundary attack.**<br>Adversarial training improves the robustness of tabular heart disease models to boundary attacks. It involves generating adversarial examples that lie at allowable input variable limits and retraining the model on these along with genuine data. Optimization on clean and boundary data makes the model's predictions less sensitive to extremes, improving its defense against attacks that exploit feature distribution edges to cause misdiagnosis. |

| | | | |
|---|---|---|---|
| | | | **Apply the Density estimation method to detect if inputs significantly diverge from the training data distribution. Identify and flag anomalous outliers that could be adversarial.**<br>This attack can be performed by feeding datasets similar to our chosen dataset but from various different domains to test the robustness of the model. |
| Bandara W H M T S | Apply the Brendel & Bethge attack against the model and identify accuracy changes. Test defenses. | • Study and research about Brendel & Bethge attack and its impact on Cardiovascular disease prediction models.<br>• Analyze existing Cardiovascular disease prediction models to identify vulnerabilities when exposed to adversarial attacks, considering Brendel & Bethge attack.<br>• Identify and document specific weaknesses in our current Cardiovascular disease prediction model when confronted with the Brendel & Bethge attack. Emphasize potential consequences and highlight areas that require improvement.<br>• Explore new features to improve how the model understands Cardiovascular disease | **Apply Adversarial Training mechanism to modify the training process of a Cardiovascular disease prediction model based on tabular data to reduce its vulnerability to Brendel & Bethge attack.**<br>Train the model on adversarial examples that are constrained to remain within realistic boundaries for the input features based on domain knowledge. This makes the |

| | | | |
|---|---|---|---|
| | | data. Think about using different types of information or changing how the data is presented to make the model work better.<br>• Investigate feature selection methods capable of reducing the model's susceptibility to adversarial perturbations. Ensure that the chosen features contribute robustly to predictions, providing a more secure foundation.<br>• Implement and meticulously test the techniques on the Cardiovascular disease dataset. Evaluate their impact on overall model performance and resilience against adversarial attacks.<br>• Design and implement tailored adversarial training strategies for our Cardiovascular disease prediction model. Account for various attack scenarios, optimizing the training process to enhance the model's robustness.<br>• Experiment with a range of adversarial attack scenarios during the training phase. Assess the model's performance and resilience under different conditions, including varying attack strengths and types, to ensure it's well-prepared for anything. | model robust to perturbations that could occur in real data.<br>**Apply Gradient obscuration mechanism to modify the training process of a Cardiovascular disease prediction model based on tabular data to reduce its vulnerability to Brendel & Bethge attack.**<br>Implement obfuscating components in the model to misguide the adversary's gradient direction used to construct perturbations.<br>This attack can be performed by feeding datasets similar to our chosen dataset but from various different domains to test the robustness of the model. |

9. Supervisor checklist

a) Does the chosen research topic possess a comprehensive scope suitable for a final-year project?

| | | | |
|---|---|---|---|
| Yes | x | No | |

b) Does the proposed topic exhibit novelty?

| | | | |
|---|---|---|---|
| Yes | x | No | |

c) Do you believe they have the capability to successfully execute the proposed project?

| | | | |
|---|---|---|---|
| Yes | x | No | |

d) Do the proposed sub-objectives reflect the students' areas of specialization?

| | | | |
|---|---|---|---|
| Yes | x | No | |

e) Supervisor's Evaluation and Recommendation for the Research topic:

Approved.

10. Supervisor details

| | Title | First Name | Last Name | Signature |
|---|---|---|---|---|
| Supervisor | Lecturer | Chethana | Liyanapathirana | Chethana (Signed) |
| Co-Supervisor | Assistant Lecturer | Deemantha | Siriwardana | Deemantha (Signed) |
| External Supervisor | | | | |
| Summary of external supervisor's (if any) experience and expertise | | | | |

**This part is to be filled by the Topic Screening Panel members.**

Acceptable: Mark/Select as necessary

| | |
|---|---|
| Topic Assessment Accepted | |
| Topic Assessment Accepted with minor changes (should be followed up by the supervisor)* | |
| Topic Assessment to be Resubmitted with major changes* | |
| Topic Assessment Rejected. Topic must be changed | |

* Detailed comments given below

## Comments

## The Review Panel Details

| Member’s Name | Signature |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |

***Important**:

1. According to the comments given by the panel, make the necessary modifications and get the approval by the **Supervisor** or the **Same Panel**.
2. If the project topic is rejected, identify a new topic, and request the RP Team for a new topic assessment.
3. The form approved by the panel must be attached to the **Project Charter Form**.