**CSCI434/534 COLL 400 Capstone Project**

**Report DUE: May 4 (Sunday), 2025, at 11:59 pm**

**Overview**

The COLL 400 capstone experience will require students to engage in a research project that involves monitoring and analyzing computer network traffic, particularly from popular internet websites such as Amazon, Netflix, and Bank of America. The project will require students to apply the knowledge and tools they have learned in this class to extract new features from website traffic data and build new models to classify the specific website from which the traffic originated. The classification must be done automatically, without access to the website URL or data payload that is typically encrypted.

**Goals**

* To apply AI/ML, statistical, and/or mathematical knowledge and tools acquired from other courses in a real-world scenario.
* To enhance teamwork skills by working collaboratively in groups.
* To engage with general end users of the internet, such as non-CS students on campus, to obtain feedback on the project's ability to classify website traffic and present the findings in a way that is accessible to people outside of the class or discipline.

**Requirements**

* To work in teams of three to five and present projects using PowerPoint slides in the final week of the semester.
* A final project report is required and will be graded, accounting for 20% of the final grade. All members of the same team get the same grade for the project.
* A GitHub repository is required to host all the code used in the project along with the PowerPoint slides for the presentation.

**Project** **Guidelines**

* Objective: Develop models to classify web traffic origins (e.g., Amazon, Netflix, Bank of America) using extracted traffic data features, without access to URLs or encrypted payloads.
* Dataset: The dataset should contain at least two essential components: 1) output labels indicating the originating website (at least 3 websites as labels) and 2) input features extracted from the traffic data that are critical for determining the website's origin. You could collect data from two sources:
  + **Manual Collection**: You are encouraged to use [Wireshark](https://www.wireshark.org/) to manually collect data by visiting specific websites and gathering all corresponding traffic data (Guideline video available).
  + **Public Datasets**: You could also utilize public datasets that include above two essential components (output labels and input features).
* Method: You could start with classic classification models, such as Support Vector Machines (SVM), Decision Trees, and Convolutional Neural Networks (CNNs), to tackle the challenge of website classification. You are encouraged to modify and combine these models or develop a new model to improve accuracy and efficiency in classifying website traffic origins.
* Evaluation:
  + Split data into training, validation, and testing sets for a comprehensive evaluation.
  + Establish at least one baseline model, such as a simple linear classifier without modifications, to serve as a point of comparison.
  + Compare the baseline model's performance against your proposed method using standard metrics such as accuracy[1] and F1-Score[2].
  + Analyze the results to understand the strengths and weaknesses of your proposed method in classifying website traffic origins.

**Final Report**

The report should be 3-7 pages long, in the SIGCOMM IMC Format (Latex and Word templates available), includingtables, figures, and references. Below are the key components your report should cover:

* Abstract (5%): Provide a concise summarization of your work, briefly mentioning the motivation of your work, what was the objective you tried to solve, and the approaches you applied to solve the problem. You should also include the major evaluation results as well as their implications.
* Introduction (10%): Begin with a broad overview of your project, introducing the issue or topic you're addressing. Explain its importance and relevance, including any real-world applications. Then, succinctly describe your dataset, methods, evaluation results, and the novelty of your approach, covering all parts of your paper. Conclude this section with a clear statement of your project's scope and contributions.
* Proposed Method (35%): This is the heart of your report. Detail the methodology you employed to address your project's challenge or to achieve its goals. Describe the steps you took, your data collection methods, the models you used, and any technical specifics. Incorporate diagrams or flowcharts to aid understanding.
* Evaluation (15%): Present your project's outcomes here. Explain the tests or experiments you conducted to gauge your approach's effectiveness. Discuss the metrics you used to evaluate success and whether your project met its objectives.
* Discussion & Future Work (25%): Analyze and interpret your results in this section. Discuss the relevance of your findings in relation to the problem tackled, ensuring your explanations are accessible even to those outside your specific field or class. Address any challenges or limitations you faced, suggest potential solutions or future research directions, and underline the practical implications of your work.
* Conclusion (5%): Sum up your project succinctly, restating its importance and the main insights gained.
* References (5%): List all sources you've referenced in your report, including research papers, books, websites, etc.

**Appendix**

1. Accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined [[Link](https://en.wikipedia.org/wiki/Accuracy_and_precision)].

where TP = True positive; FP = False positive; TN = True negative; FN = False negative.

1. F1-Score computes how many times a model made a correct prediction across the entire dataset. [[Link](https://www.v7labs.com/blog/f1-score-guide)].

where TP = True positive; FP = False positive; TN = True negative; FN = False negative.