Feature Extraction Engine – Afterword
Sonia Sherpa

Ethics. What ethical questions came from this problem? And how did you address them?
There weren't many ethical issues surrounding the specifics of our project. One of the concerns we did think about was that someone might use our project to figure out what kind of cyber-attacks a company usually gets and exploit that vulnerability. Working with pcap files also poses an ethical and security issue, as they might have sensitive and confidential information sometimes, but the ones we were given by our client were safe to use and available to the public. We use this project mainly to help cybersecurity groups understand what different cyber-attacks look like. We do this by looking at certain data in computer files. But there's a risk that someone bad might use our project to figure out what kind of attacks a company usually gets.
One way to solve this problem is to make the data we look at more secure. The company could use our project to better understand the attacks and protect themselves. They could also use it to prepare for different types of attacks, making their computer systems even safer.

Learn. What did you learn about on this project?
I have always wanted to get into machine learning at some point because of how important it is in today's tech world. I'm grateful for this project opportunity as I not only learnt about datasets, machine learning, everything about what pcap files are, how they work and how they are useful to us. I also learnt a new language though out this project: Python. And working with libraries like NumPy and scapy among others. Learning how to extract statistical data from the contents of pcap files too. I have also not worked in this big scale of a project before and I'm very grateful for Dr. Warren for the guidance throughout the project and class! And to our client, James, for this opportunity!

Contribute. For each member, what was your contribution to this project?
My main role for this project was choosing which kind of data to extract from the pcap file and figuring out how to do so. Our client told us and gave us the freedom to extract any kind of data as long as its statistical as they would be running an algorithm over it anyway to find what's useful for them. I also wrote code to search if an TCP layer exists; and if it does then to extract specific data. Eventually, I ended up creating at least 65 rows of statistical data for each sample. Selecting what to extract from the packet for example: destination IP addresses, source and destination IP address lengths, and source and destination port numbers; can help provide more insights into the packet data and potentially improve the performance of any subsequent analysis or machine learning models.