1. Contact info (e.g. email/phone). Eugene Moore moore938@regis.edu

2. Title of the project

3. High level description of the project: what question or problem are you addressing?

Since 2011[1] malware analysis is beginning to see progress with novel ways through the help of computer vision. Traditionally, security research on malware is a lengthy process resulting in error prone signature detection. In the paper “Malware image classification using one-shot learning with Siamese networks.” An attempt to classify malware similarity with limited samples is presented that might solve two problems. The first of limited sample data and second a way to spot malware “strains”, for lack of a better word.

I will use this paper as a guide to first reproduce the setup. This will provide a foundation to build on in future work. Given the time frame for this course work I believe a successful project would result in the setting up similar data processing and training networks to what is presented in the paper. Another paper, “Deep multi-task learning for malware image classification.” Presents a more advanced approach that uses multiple data points or tasks to classify malware. The use of color imaging through a custom technique, not presented, could offer better results for the one-shot approach. Time permitting, I will attempt to experiment with this option. Finally, the latest paper researched for this project presents a transfer learning approach. All techniques used to replicate these experiments have been taught and in my course work here at Regis. My main goal of this project is to leverage and demonstrate these techniques. I anticipate challenges at every phase from data selection and collection, data engineering, CNN setup, and transfer learning. I will attempt to follow the basic data science workflow taught at Regis and will no doubt have to curtail some experimentation to complete final analysis of results.

4. What type of data science task is it? (some example answers but not limited to)

This is an image classification (Computer Vision) using supervised learning. There will be some data engineering task as well to transform binary files into images.  
  
5. Data: Brief description of data. How big do you expect the data will be? Is amount of your data too big or too small? If you're web-scraping or collecting data, how long do you expect to collect the data?

Due to the choice of lab I’m attempting to replicate, I do not expect to need a large amount of data since the point of the approach for the scenario is that of limited sample data. Additionally, since the focus of the experiment is on malware, I will first use benign binaries from my local system to establish a data pipeline. Once I have a successful end to end data prep, train, and analysis established, I may use live malware as a final proof of concept. Using live malware can be problematic since it will most likely require dedicated hardware in a controlled and disconnected lab which I have access to but may not have time to set up given the course duration.

6. How will you analyze the data? What machine learning methods do you plan to use, and/or what business intelligence aspect do you plan on incorporating?

In the research I’m using as a guide I will be setting a Siamese network configuration for two CNN networks. My target setup will be transfer learning as described in the paper “IMCNN:Intelligent Malware Classification using Deep Convolution Neural Networks as Transfer learning and ensemble learning in honeypot enabled organizational network”

7. Describe any anticipated difficulties and problems. Discuss how you may overcome the problems.

I think the main difficulty will be obtaining and cautiously working with live malware. This sometimes requires gaining access to samples through a vetting process by private organizations. The next issue will be containing malware which can drive some isolated lab setup which can be time consuming task that are not directly related to the experimentation. Another challenge will be attempting color imaging for binary sample. Although the general architectures for this type of work are referenced, much of the technical implementation details will be not only challenging but perhaps time consuming. The idea for this comes from more advanced research in this area, however, the technique for color is not disclosed. From what I can tell it may be a special way of coloring different parts of the original binary that corresponds to areas of malware targets. So, my attempt at coloration, time permitting, will be purely my own experimentation which might prove rewarding.

8. Suggest a timeline for the project. This should be a weekly breakdown of what you plan on doing each week.

Wk1 Project research, write-up, schedule, and program acceptance

Wk2 Initial data collection, EDA, and data engineering

Wk3 Continued data engineering.

Wk4 CNN setup and initial classification experimentation

Wk5 Continued CNN experimentation, and deep learning approach.

Wk6 Results analysis.

Wk7 Document practicum

Wk8 Finalize presentation and presentation.

9. Create GitHub repository for your Practicum project. Add this proposal, begin a ReadMe document, and begin adding your data to your repository. Add a link to your GitHub repository to this document.

<https://github.com/gcc71/msds-692-moore>

# Reference

1 Hsiao, S.-C., Kao, D.-Y., Liu, Z.-Y., & Tso, R. (2019). Malware image classification using one-shot learning with Siamese networks. *Procedia Computer Science*, *159*, 1863–1871. https://doi.org/10.1016/j.procs.2019.09.358

2 Bensaoud, A., & Kalita, J. (2022). Deep multi-task learning for malware image classification. *Journal of Information Security and Applications*, *64*(103057), 103057. https://doi.org/10.1016/j.jisa.2021.103057

3 Kumar, S., Janet, B., & Neelakantan, S. (2024). IMCNN:Intelligent Malware Classification using Deep Convolution Neural Networks as Transfer learning and ensemble learning in honeypot enabled organizational network. *Computer Communications*, *216*, 16–33. https://doi.org/10.1016/j.comcom.2023.12.036