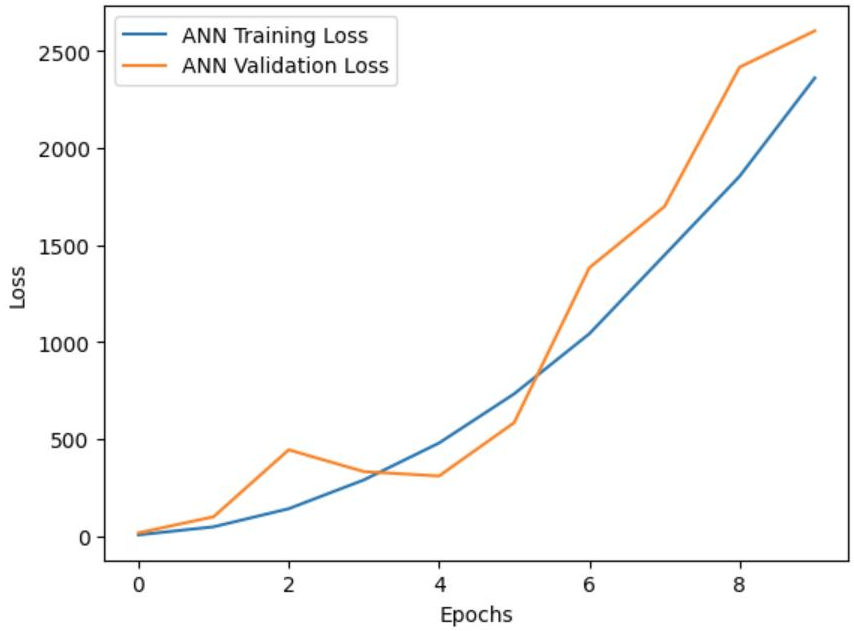
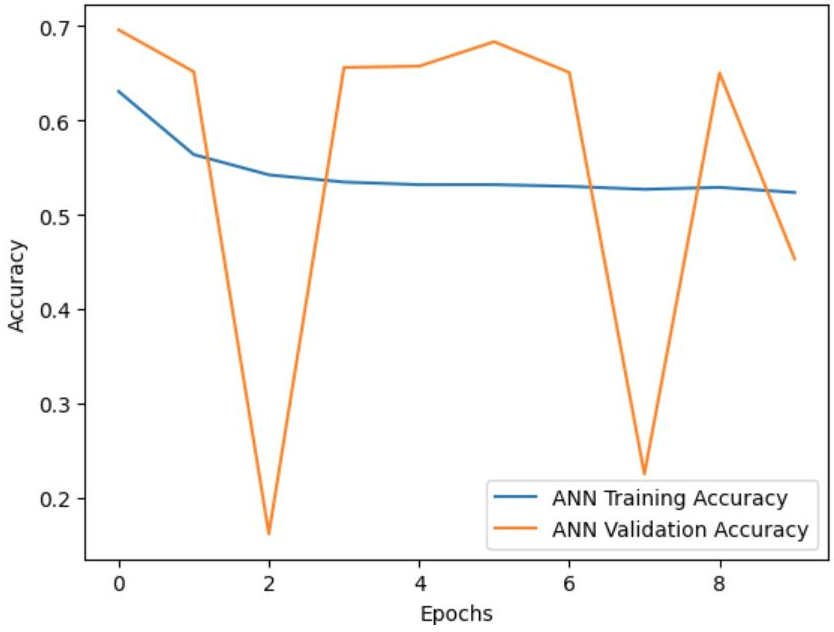
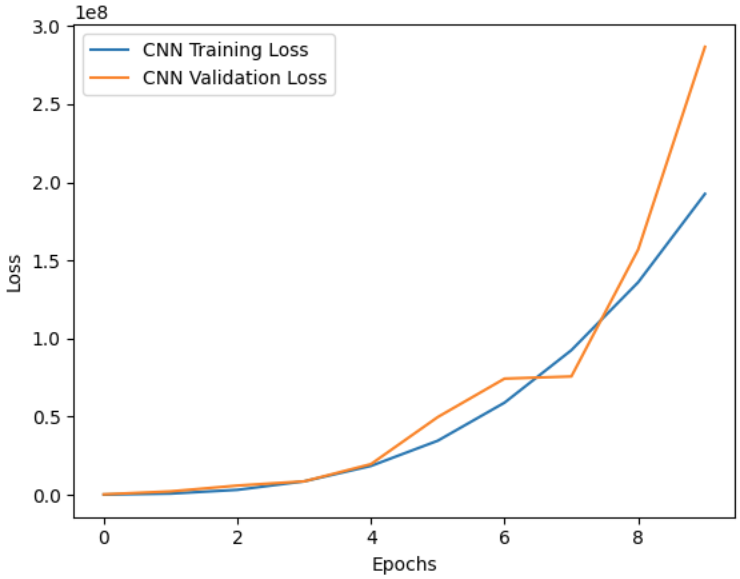
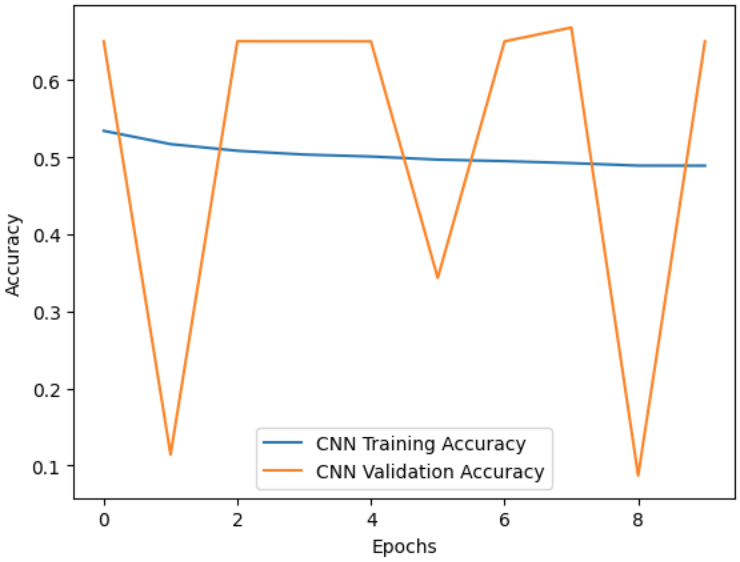
Modal Card: Multiclass Classification

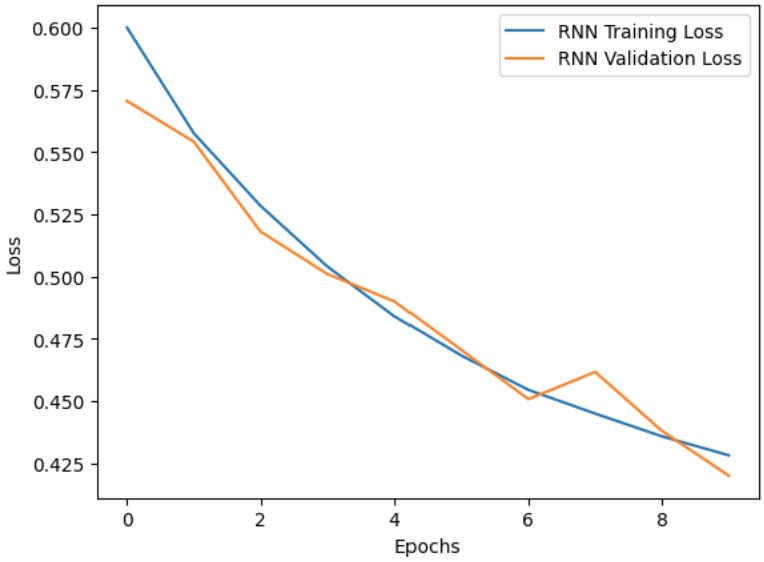
1. **Models Evaluated**
    a. ANN (Artificial Neural Network)
    b. CNN (Convolutional Neural Network)
    c. RNN (Recurrent Neural Network)
    d. RNN version (GRU Gated Recurrent Unit)
2. **Use Case**
    a. Network traffic data from packet sniffer was gathered already and provided as a clean dataset for training and testing (Please refer to data sheet) for further information.
    b. **In scope:** We perform usual EDA (exploratory data analysis) and data manipulation before we feed in the data for multiclass classification.
    c. Out of scope: Obtaining real data on network threats is a hard task in this case we were fortunate to get the synthetic and original data. (please refer to data sheet) for further information. Generating synthetic data is by itself an effort worth doing as a project.
    d. We use this dataset to do our multiclass classification analysis and produce results.
3. **Factors**
    a. ANN's are versatile and can be used for classification, regression, and pattern recognition.
    b. CNN's are typically used for images and videos due to spatial nature of the dataset.
    c. RNN's are typically used for sequential data processing often time series data that requires memory of tasks over time.
    Although each of the above neural networks perform better for their specific use case I decided to convert our data set to the required input format and perform multiclass classification
4. **Metrics**
    Primary metric for my evaluation was loss and accuracy.

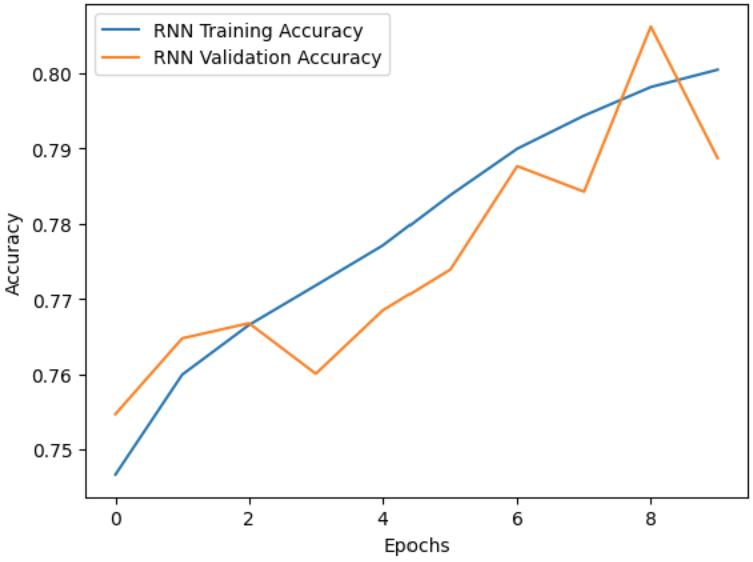

ANN Training Loss
ANN Validation Loss
2500
2000
1500
1000
500
0
Loss
0
2
4
6
8
Epochs

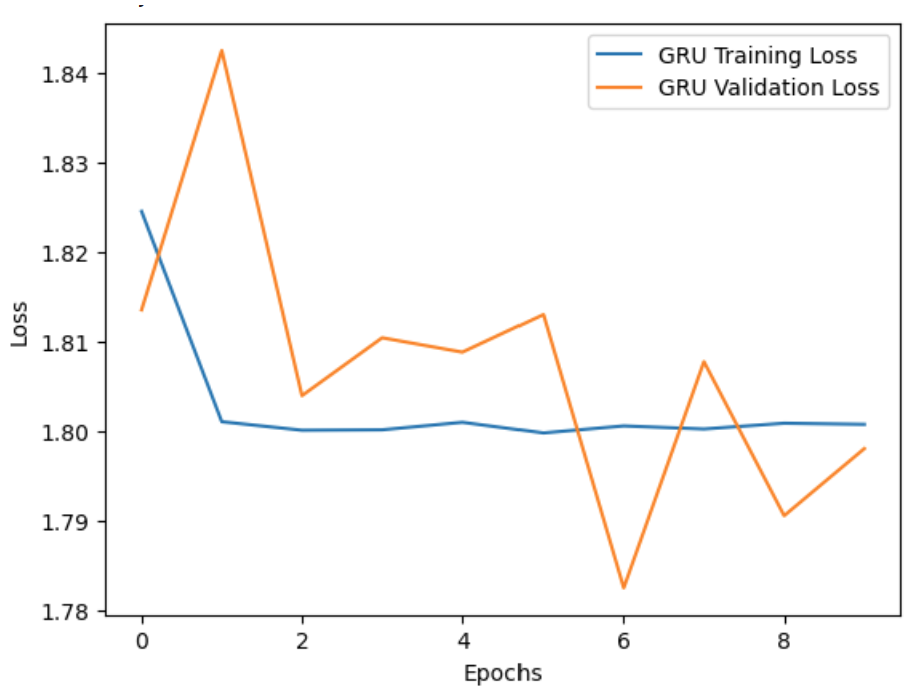

0.7
0.6
0.5
0.4
0.3
0.2
Accuracy
ANN Training Accuracy
ANN Validation Accuracy
0
2
4
6
8
Epochs

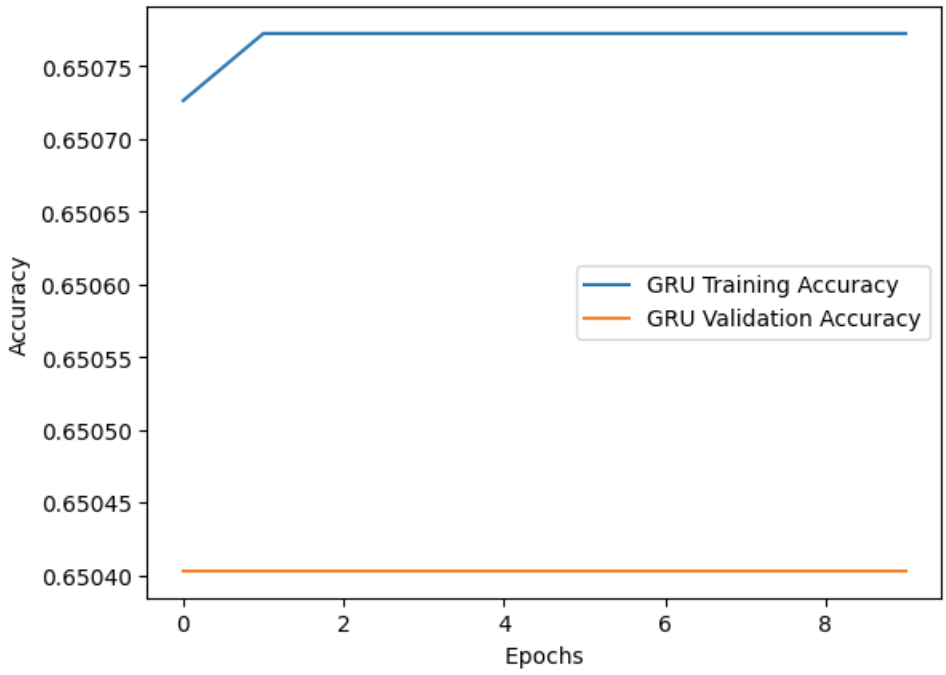

1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Loss
CNN Training Loss
CNN Validation Loss
0
2
4
6
8
Epochs


0.6
0.5
0.4
0.3
0.2
0.1
Accuracy
CNN Training Accuracy
CNN Validation Accuracy
0
2
4
6
8
Epochs

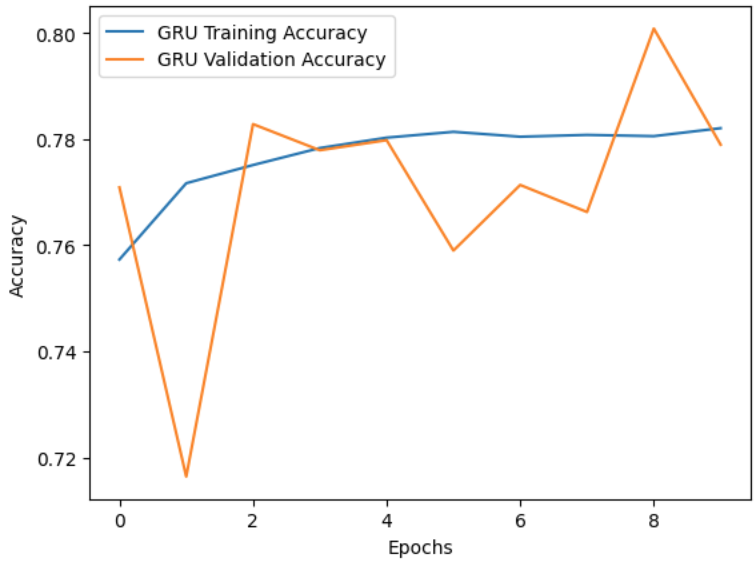

RNN Training Loss
RNN Validation Loss
0.600
0.575
0.550
0.525
0.500
0.475
0.450
0.425
Loss
0
2
4
6
8
Epochs


RNN Training Accuracy
RNN Validation Accuracy
0.80
0.79
0.78
0.77
0.76
0.75
Accuracy
0
2
4
6
8
Epochs





Based on the above observed metric it is evident simple RNN performs better than complex multi layered RNN network (or) ANN or CNN. So I decided to focus on optimizing simple GRU architecture (version of RNN) and performed GridSearchCV to optimize it.
Below are the loss and accuracy results.



Observation says for simple RNN and Simple GRU architecture loss reduce over time (epoch cycles) and accuracy approaches ~80%

5. **Dataset**

   Please refer to the data sheet to get mode details.

6. **Caveats and recommendations**

   On a positive note, I did not have to worry too much about feature reduction or picking models to train and test. Neural network takes care of it with the least effort as long as data is fed into it in the right format. RNN performs way better compared to other neural networks due to the nature of the cybersecurity dataset. Cybersecurity data set follows a sequence as time passes by. Attack happens over time with gradual escalation from exploiting weakness, compromising passwords and then exfiltrating layer after layer to approach sensitive or critical data in the network. Although my effort did not use time stamps there is an epoch time stamp field or feature that could have been used for RNN's or version of RNN (GRU) networks. My recommendation would be to start with RNN's from the start.

7. **How to use it in your organization?**
   Generate network traffic (headers from packet captures) these can often be too large to handle. Exercising caution to generate just enough not more than 1 million would be a great idea to train the model. Input data may have to be cleaned up appropriately before feeding it into the model. If this model is in production developing KPI (Key Performance Index) of the model is important to prevent model drifts. You may have to retrain the model since new threats are developed more frequently now.