**DAEN 690 Paper Section 4.3**

**Model Training**

As mentioned previously in section 3.2, the second deliverable for our project involves more predictive analytics and machine learning, used to forecast ISS regions. Before conducting our analysis, we preprocessed our dataset to choose relevant attributes, target variables, and make sure valid entries were included. Since we are working with time dependent data, we tried multiple time series models. For time series forecasting we implemented the Autoregressive Integrated Moving Average (ARIMA) model and the Long Short-Term Memory (LSTM) model. The dataset we focused our training and testing on was on the Buffalo, New York IGRC Dataset.

Before training and testing both the models, we needed to preprocess the dataset to have the proper variables to feed into our model. We first filtered the dataset to the years 2022 and 2023, which helped to reduce the model complexity. Additionally, reducing the dataset led to more consistent data recordings from the radiosonde. Years too far back would result in much more missing data than newer years. Lastly, as for deciding which years to pick, this would be a research gap that would need to be looked into more into the future as determining the best length of time for our time series models would need to be simulated(medium). A second preprocessing task we focused on was converting the ISSC column, which was represented as a character data type, into an integer, which will allow us to feed into our model. The third preprocessing task was to make the date column the index as this was needed for our time series models. Lastly, upon some exploratory data analysis, we removed extreme outliers. We had one day that experienced 55 instances of ice supersaturated conditions, which in our initial testing threw off the models as the mean of times per day that ice supersaturated conditions were present was around 2.

After preprocessing, we went on to feature engineering. We created an ISSC volume feature which represented the number of instances per day ice supersaturated conditions occurred and that would be our dependent variable that our models would try to predict. As for our independent variables, we tested the following:

1. Average Daily Temperature – The mean temperature per day.
2. Average Daily Relative Humidity to Ice – The mean relative humidity to ice per day.
3. Rolling Average – The average value of ice supersaturated condition volume per day within a 5-day window size.
4. Exponential Smoothing – The average value of ice supersaturated condition volume per day within a 5-day window size, assigning weights to each observation. As the observation becomes older, it gets assigned a lower weight.
5. Temperature Daily volume – The number of times per day the temperature was recorded as below -42 degrees Fahrenheit.
6. Relative Humidity to Ice Daily Volume – The number of times per day that relative humidity to ice was above 100 percent.
7. Lag Features – Taking the past values and incorporating them into the current prediction.

We initially started working with the ARIMA model. We built

After some unsuccessful attempts we switched to using an LSTM model. Initially we started just by using lag variables of ‘volume’ (how many ISSCs the day had) for the predictions. We progressed through using other attributes like rolling averages and exponential smoothing of the ‘volume’ attribute, as well as averages of temperature and relative humidity to ice. This improved our predictions slightly, but we got the best results when adding ‘volume of temp’ and ‘volume of rhi’ (These variables are counts of how many times temperature and rhi were ISSC causing each day). To make sure we would get the best results we first determined how many lag variables to include and for what attributes. We then implemented another grid search to find the optimal hyper parameters such as learning rates and batch size. To visualize this better we created a heat map of the LSTM parameters, showcasing which combinations had the lower Mean Squared Error.

**Model Evaluation**

**Model Validation**