Approach:

Initially, the basic and most important step was to understand the data and know the input and output. This made clear that it is a problem of regression between the range 0 and 1.

1. Pre-processing step

Data had **25** columns including the output.

1. Removing features that were not required first by manual analysis
   1. All soldier Ids and Attack Id were unique and had to be removed.
   2. Calculated total number of unique Ship IDs. **The ratio of unique ship IDs to that of total instances was 1:2** so it wasn’t right to take ship ID as an input to the model.
   3. As mentioned in the problem statement the feature “farthermostKill” could be **misleading**, so it had to be dropped.
2. Removing features which are highly correlated.
   1. Calculated **correlation coefficient** matrix for all input features which each other.
   2. “HealthLost” and “KillRank” were **highly correlated**, however later found out that eliminating any one of them decreases the accuracy.
3. Attempt to remove the outliers.
   1. Assumed, data to have a normal distribution. and went on for removing outliers. Removing datapoints which were **3 sigma** or **4 sigma** away from the mean decreased the accuracy to a great extent. Even extended this to **6 sigma** and **7 sigma**. In the later case even though the number of outliers **decreased** still it reduced accuracy and hence could not remove outliers.
   2. After the competition was over realised that not doing visualisation was a big mistake
4. Normalising the data.

Features from input data are on different scales and are thus normalised in the range **0 to 1.** This is the standard scaling of data.

1. Hybrid model.
2. First, made an vanilla **artificial neural network** architecture which had the following structure.
3. **Input layer** consisting of 20 neurons with sigmoid activation function
4. **Hidden Layer** consisting of 6 neurons with sigmoid activation function again
5. **Output layer** of one neuron which had just a linear activation function.
6. Learning rate of 0.025, batch size: 6000

Tried all possible structural and functional changes in the architecture of this neural network like changing the learning rate, batch size, activation function, number of neurons and found the above architecture as the **optimum one.**

Also tried to increase the number of hidden layers, but that only increased the training time.

1. RandomForest Regressor: This model gave good accuracies however **could not be better** than artificial neural network. The number of trees used were 10, as this was a hyper parameter it was decided earlier. The main problem with random forest regressor is it takes **average over the data points** in the leaf nodes. This could not work well with unseen data.
2. Linear regression. Even a simple linear regression line was giving a **descent fit** to the data however it was quite incorrect to entirely depend on linear regression as a high dimensional data was pretty sure to follow a **non-linear pattern**.
3. Idea for hybrid model.
   1. Since the best model (vanilla ANN) didn’t show any much effect after changing the functional and structural part of the model. So, there was a need to implement something different.
   2. As we know, in ensemble model a **majority of vote** is taken. Similarly, here in the hybrid **average of output of above 3 models** was taken. This was sure to give a better performing model than above 3 models.
   3. For any ensemble model to work well, there is necessity that the output of individual models should **not be correlated** with each other.
   4. Before deciding which models to take for hybrid, analysis was done on various models like two vanilla neural networks having different structure and functions etc . However, these models were highly correlated and an ensemble of them could not taken.
   5. Finally, 3 models were confirmed **(A. ANN with sigmoid activations, B RandomForest Regressor and C. Linear regression**). The output of these models was less correlated to each other.
   6. The final hybrid was prepared, and score was checked.
4. The hybrid model achieved **roughly the same score** as the ANN. Reason can be explained as follows:
   1. As mentioned earlier, for any hybrid model where majority of votes is taken or average of the output of individuals is taken. The output of individual models should not be correlated.
   2. The output of above 3 models were to some extent correlated to each other. And this correlation **wasn’t sufficiently less** to make our hybrid model perform good and hence it did not achieve more than the ANN.
5. Final submission was done on the vanilla ANN created and a score of **94.7** was achieved.
6. Because of unavailability of **GPU** it took a long time and patience to train different structural and functional types of ANN.