**Evaluation of models**

For each pair of heuristic algorithm and scaler type, we fetched two sets of independent results and evaluated them together.

This is because although heuristic algorithm’s objective is to return good results, it does not guarantee returning best result for a given problem every time its executed. Moreover, it also does not guarantee returning same results for a given problem after each execution.

In real-world scenario, we may want to run and test the algorithm over multiple number of times in order to get a good sample size of results for making inference, but due to limitations of time and compute resources, we restricted to two results and evaluated the same.

Summary of results fetched by evaluating models based on each heuristic algorithm

**Evaluation of results obtained using Artificial Bee Colony algorithm**

1. Binary classification using Standard Scaler

2. Multi-class classification using Standard Scaler

3. Binary classification using Robust Scaler

4. Multi-class classification using Robust Scaler

**Evaluation of results obtained using Flower Pollination Algorithm**

1. Binary classification using Standard Scaler

2. Multi-class classification using Standard Scaler

3. Binary classification using Robust Scaler

4. Multi-class classification using Robust Scaler

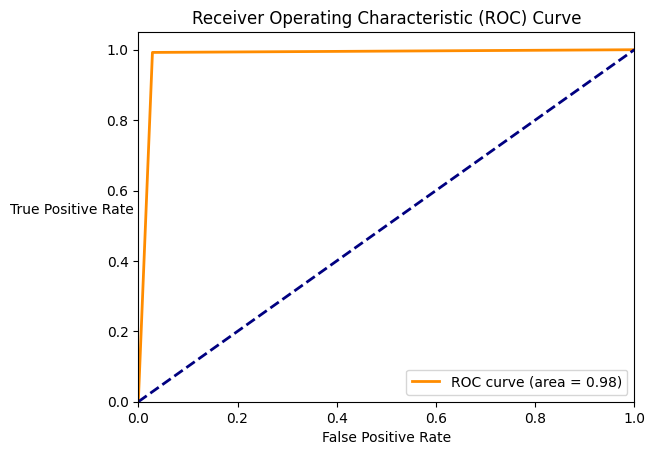
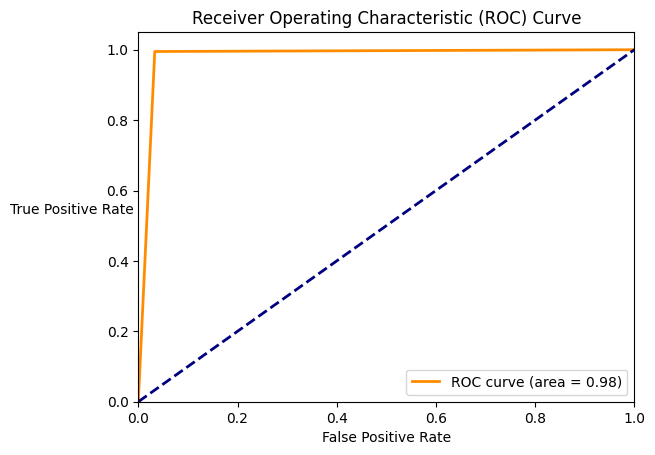
**Evaluation of results obtained using Artificial Bee Colony algorithm**

1. Binary classification using Standard Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced test dataset** | | **Imbalanced test dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 13 | 11 | 13 | 11 |
| **Confusion matrix** | [[20103, 589], [476, 61600]] | [[20006, 686], [314, 61762]] | [[1404119, 33045], [80723, 216214]] | [[1382983, 54181], [76129, 220808]] |
| **Accuracy** | 0.987 | 0.988 | 0.934 | 0.925 |
| **Precision** | 0.991 | 0.989 | 0.867 | 0.803 |
| **Recall** | 0.992 | 0.995 | 0.728 | 0.744 |
| **F1-Score** | 0.991 | 0.992 | 0.792 | 0.772 |
| **AUC Score** | 0.982 | 0.981 | 0.853 | 0.853 |
| **Balanced accuracy** | 0.982 | 0.981 | 0.853 | 0.853 |
| **MCC** | 0.966 | 0.968 | 0.757 | 0.728 |
| **NPV** | 0.977 | 0.985 | 0.946 | 0.948 |
| **FDR** | 0.009 | 0.011 | 0.133 | 0.197 |
| **Cohen Kappa** | 0.966 | 0.968 | 0.753 | 0.727 |

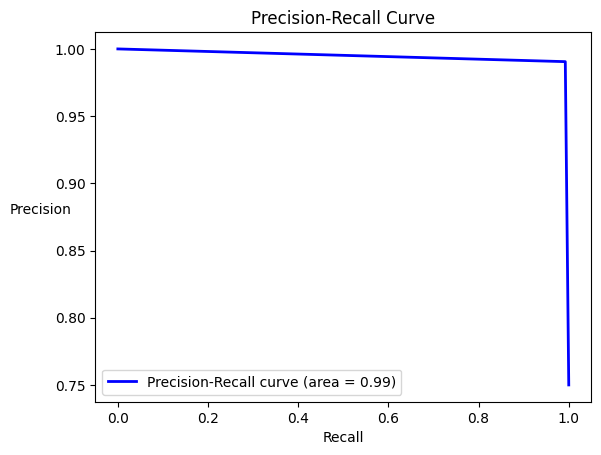
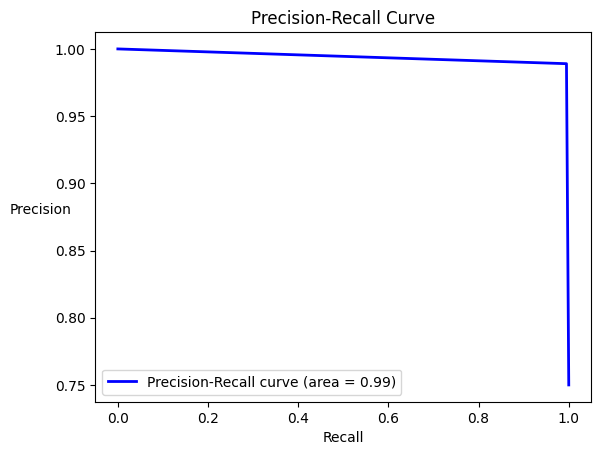
**Balanced test dataset: Comparison of ROC curves**

**Solution 1** **Solution 2**

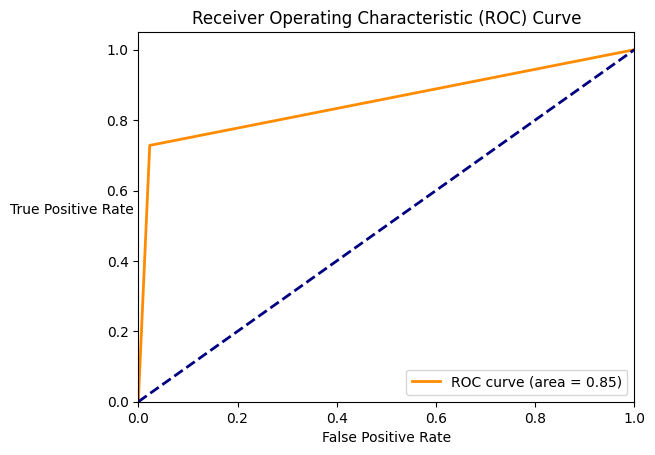
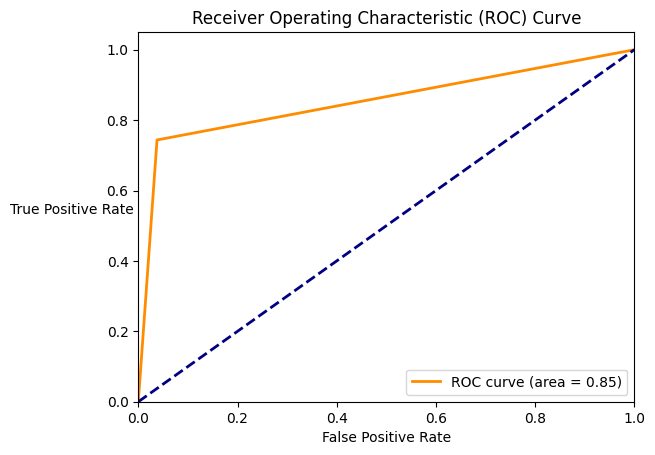
**Balanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**

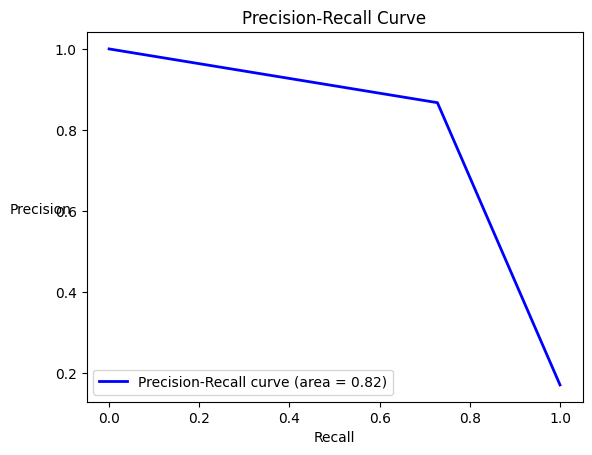
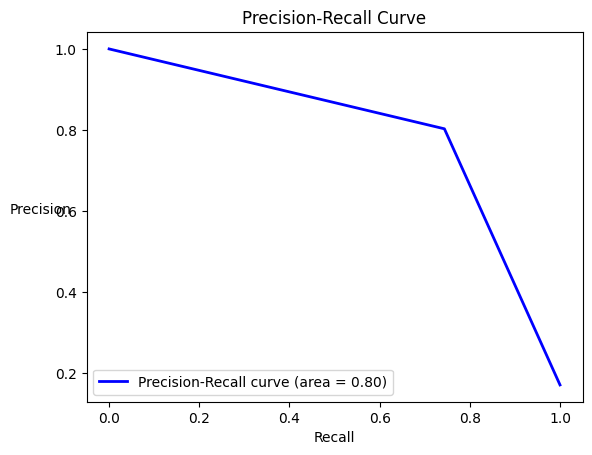
**Imbalanced test dataset: Comparison of ROC curves**

**Solution 1 Solution 2**

**Imbalanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**

**Balanced dataset: -**

1. Both solutions have almost **equal and very high accuracy**. Thus, both models perform correct prediction around 98% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less. Solution 1 has slightly higher precision than solution 2.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives). Solution 2 has slightly higher recall than solution 1.

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have almost **equal and** **very high AUC scores** which is closer to 1. Thus, both models have great ability to differentiate between benign and malicious events.

6. Both solutions have **almost equal and very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

7. Both solutions **have almost equal and very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

8. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 98% times it was correct and the event was actually benign (that is not malicious). Solution 2 has higher negative predictive value than solution 1.

9. Both solutions **have very low false discovery rate**. Thus, for both models when an event was classified as malicious, around 10% of the times it was incorrect and the event was actually benign. Thus, both the models have noise close to 10%.

10. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

11. Both solutions **have ROC curve closer to axes, and the elbow is closer to coordinate (0, 1)** which indicates both the models have larger value of area in their respective ROC curve. Thus, both the models are good classifiers.

12. Both solutions have **Precision – Recall curve have very high value for both precision and recall**. Thus, both the models are good classifiers.

**Imbalanced dataset: -**

1. Both the solutions **have a very high accuracy**. Thus, both models perform correct prediction around 93% of the times.

2. Both solutions **have high precision**, greater than 80%. Thus, the model incorrectly classifies some of the normal events as malicious, resulting in false positives.

3. Both solutions have **average recall**, around 70%. Thus, the model incorrectly classifies multiple malicious events as benign.

4. Both solutions have **average F1-score**, around 78%. This is because the model’s performance drops while correctly predicting both normal and malicious events.

5. Both solutions have **equal and high AUC score**: 85%. Thus, both models have consistent ability to differentiate between benign and malicious events.

6. Both solutions have **equal and high balanced accuracy**: 85%. Thus, both models have high precision and recall.

7. Both solutions have **average MCC**, around 70%. MCC score is greater than 0 and closer to 1. Thus, both models have some agreement with the actual labels.

8. Both solutions **have very high negative predictive value**: 94%. Thus, for both models when an event was classified as benign, around 94% times it was correct and the event was actually benign (that is not malicious). Solution 2 has higher negative predictive value than solution 1.

9. Both solutions have **low false discovery rate**. Thus, for solution 1 if an event is classified as malicious, 13% times it is misclassified and the event was actually normal (benign). For solution 2, if an event is classified as malicious, 19% of times it is misclassified and the event was actually normal (benign). As the result, model for solution 1 was better than the model for solution 2.

10. Both solutions have Cohen Kappa score in the range of 0.61 and 0.80. Thus, the models **have substantial agreement with the expected model**.

11. Both solutions **have ROC curve closer to axes, and the elbow is closer to coordinate (0, 0.8)** which indicates both the models have relatively large value of area in their respective ROC curve and are medium fit.

12. Both solutions have **Precision – Recall curve with high area**. The area of P-R curve for solution 1 is greater than the area of P-R curve for solution 2. Thus, model for solution 1 performs better than the model for solution 2.

2. Multi-class classification using Standard Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced dataset** | | **Imbalanced dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 27 | 30 | 27 | 30 |
| **Confusion matrix** | [[19998, 74, 32, 588], [70, 20616, 0, 6], [87, 2, 20603, 0], [104, 0, 1, 20587]] | [[19939, 60, 23, 630], [143, 20542, 1, 6], [82, 0, 20605, 5], [84, 0, 2, 20606]] | [[1392586, 10923, 4728, 28927], [155, 28994, 2, 12], [167, 3, 20521, 1], [53871, 94, 22, 193095]] | [[1399394, 9646, 7061, 21063], [319, 28827, 2, 15], [162, 0, 20526, 4], [54696, 1163, 20, 191203]] |
| **Accuracy** | 0.988 | 0.988 | 0.943 | 0.946 |
| **Precision** | 0.989 | 0.989 | 0.845 | 0.864 |
| **Recall** | 0.996 | 0.995 | 0.817 | 0.813 |
| **F1-Score** | 0.992 | 0.992 | 0.831 | 0.838 |
| **Balanced accuracy** | 0.981 | 0.98 | 0.893 | 0.894 |
| **MCC** | 0.969 | 0.967 | 0.797 | 0.807 |
| **NPV** | 0.987 | 0.985 | 0.963 | 0.962 |
| **FDR** | 0.011 | 0.011 | 0.155 | 0.136 |
| **Cohen Kappa** | 0.984 | 0.983 | 0.803 | 0.811 |

**Balanced dataset: -**

1. Both solutions have almost **equal and very high accuracy**. Thus, both models perform correct prediction around 98% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives).

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have **almost equal and very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

6. Both solutions **have almost equal and very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

7. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 98% times it was correct and the event was actually benign (that is not malicious). Solution 2 has higher negative predictive value than solution 1.

8. Both solutions **have very low false discovery rate**. Thus, for both models when an event was classified as malicious, around 1% of the times it was incorrect and the event was actually benign. Thus, both the models have noise close to 1%.

9. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

**Imbalanced dataset: -**

1. Both the solutions **have a very high accuracy**. Thus, both models perform correct prediction around 94% of the times.

2. Both solutions **have high precision**, greater than 84%. Thus, the model incorrectly classifies some of the normal events as malicious, resulting in false positives.

3. Both solutions have **high recall**, around 81%. Thus, the model incorrectly classifies some malicious events as benign.

4. Both solutions have **average F1-score**, around 83%. This is because the model’s performance drops while correctly predicting both normal and malicious events.

5. Both solutions have almost **equal and high balanced accuracy**: 89%. Thus, both models have high precision and recall.

6. Both solutions have **high MCC**, around 80%. MCC score is greater than 0 and closer to 1. Thus, both models have some agreement with the actual labels.

7. Both solutions **have almost equal and very high negative predictive value**: 96%. Thus, for both models when an event was classified as benign, around 96% times it was correct and the event was actually benign (that is not malicious). Solution 1 has slightly higher negative predictive value than solution 1.

8. Both solutions have **low false discovery rate**. Thus, for solution 1 if an event is classified as malicious, 15% times it is misclassified and the event was actually normal (benign). For solution 2, if an event is classified as malicious, 13% of times it is misclassified and the event was actually normal (benign). As the result, model for solution 2 was better than the model for solution 1.

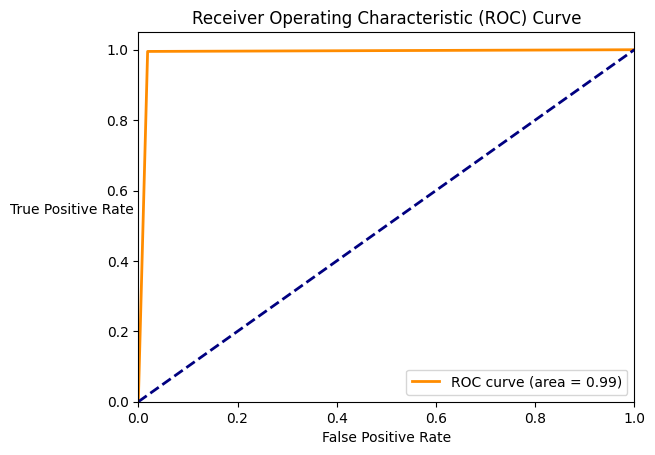
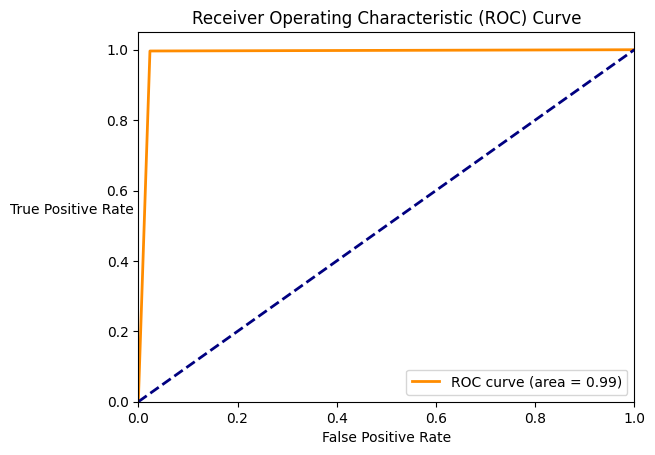
9. Both solutions have Cohen Kappa score greater than 0.80. Thus, the models **have near perfect agreement with the expected model**.

3. Binary classification using Robust Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced dataset** | | **Imbalanced dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 14 | 16 | 14 | 16 |
| **Confusion matrix** | [[20207, 485], [224, 61852]] | [[20304, 388], [301, 61775]] | [[824867, 612288], [120690, 176247]] | [[920408, 516756], [246720, 50217]] |
| **Accuracy** | 0.991 | 0.992 | 0.577 | 0.56 |
| **Precision** | 0.992 | 0.994 | 0.224 | 0.089 |
| **Recall** | 0.996 | 0.995 | 0.594 | 0.169 |
| **F1-Score** | 0.994 | 0.994 | 0.325 | 0.116 |
| **AUC Score** | 0.986 | 0.988 | 0.584 | 0.405 |
| **Balanced accuracy** | 0.986 | 0.988 | 0.584 | 0.405 |
| **MCC** | 0.977 | 0.978 | 0.127 | -0.153 |
| **NPV** | 0.989 | 0.985 | 0.872 | 0.789 |
| **FDR** | 0.008 | 0.006 | 0.776 | 0.911 |
| **Cohen Kappa** | 0.977 | 0.978 | 0.101 | -0.14 |

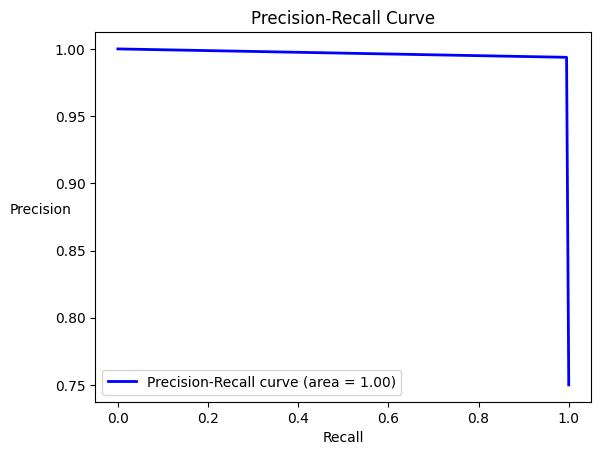
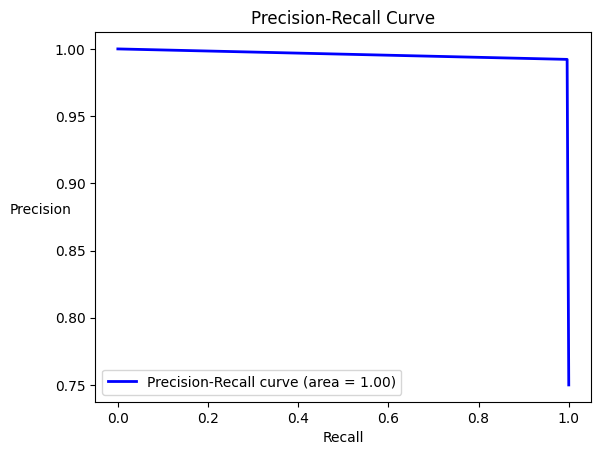
**Balanced test dataset: Comparison of ROC curves**

**Solution 1 Solution 2**

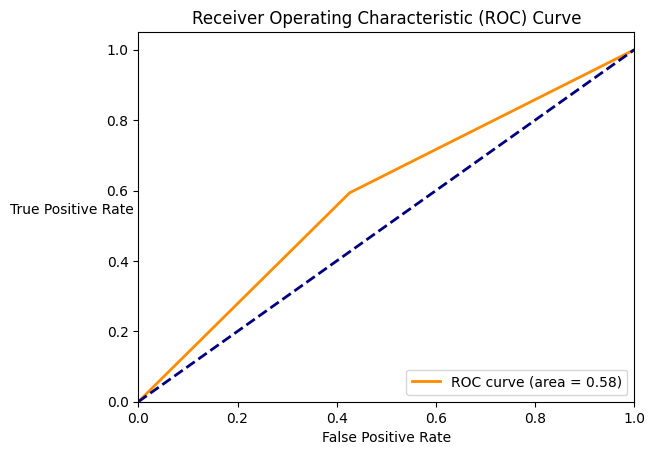
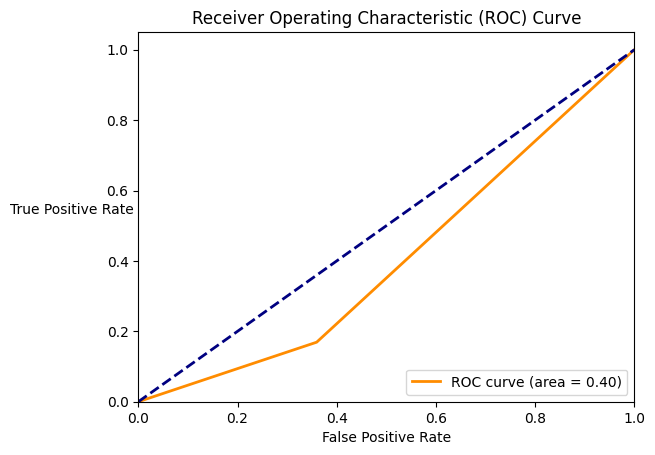
**Balanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**

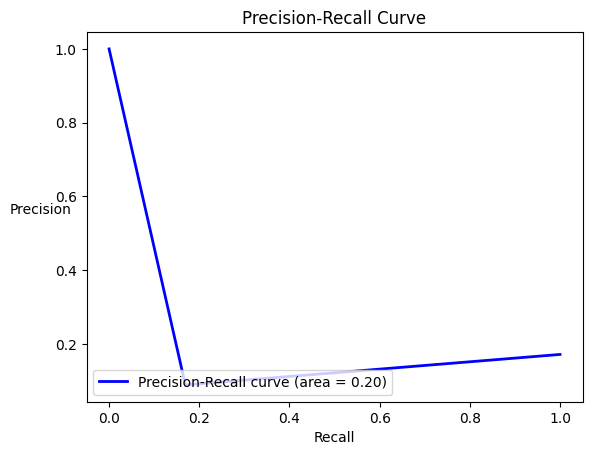
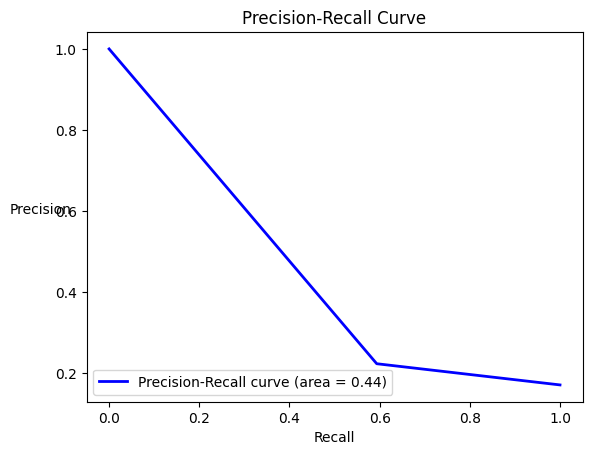
**Imbalanced test dataset: Comparison of ROC curves**

**Solution 1 Solution 2**



**Imbalanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**

**Balanced dataset: -**

1. Both solutions have **very high accuracy**. Thus, both models perform correct prediction around 99% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less. Solution 2 has slightly higher precision than solution 1.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives). Solution 1 has slightly higher recall than solution 1.

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have almost **equal and** **very high AUC scores** which is closer to 1. Thus, both models have great ability to differentiate between benign and malicious events.

6. Both solutions have **almost equal and very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

7. Both solutions **have almost equal and very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

8. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 98% times it was correct and the event was actually benign (that is not malicious). Solution 1 has higher negative predictive value than solution 2.

9. Both solutions **have very low false discovery rate**. Thus, for both models when an event was classified as malicious, around 8% of the times it was incorrect and the event was actually benign. Thus, both the models have noise less than 8%.

10. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

11. Both solutions **have ROC curve closer to axes, and the elbow is closer to coordinate (0, 1)** which indicates both the models have larger value of area in their respective ROC curve. Thus, both the models are good classifiers.

12. Both solutions have **Precision – Recall curve have very high value for both precision and recall**. Thus, both the models are good classifiers.

**Imbalanced dataset: -**

1. Both the solutions **have a very low accuracy**. Thus, both models perform correct prediction around 56% of the times.

2. Both solutions **have very low precision**. Thus, the model incorrectly classifies many normal events as malicious, resulting in false positives. Model for solution 1 has better precision than the model for solution 2.

3. Both solutions have **low recall**. Thus, the model incorrectly classifies multiple malicious events as benign. Model for solution 1 has better recall than the model for solution 2.

4. Both solutions have **low F1-score.** This is because the model’s performance is significantly low while correctly predicting both normal and malicious events.

5. Both solutions have **low AUC score**. Thus, both models are inconsistent to differentiate between benign and malicious events.

6. Both solutions have **low balanced accuracy**. Thus, both models have low precision and recall.

7. Both solutions have **low MCC**. MCC score for solution 1 is positive but closer to 0, thus its performance is very close to random guessing. MCC score of solution 2 is closer to -1, thus, it has total disagreement between the model’s predictions and the actual labels.

8. Both solutions **have high negative predictive value**. Model for solution 1 has NPV: 87% and model for solution 2 has NPV: 78%. Thus, for both models when an event was classified as benign, around 87%% times and 78% times respectively it was correct and the event was actually benign (that is not malicious). Solution 1 has higher negative predictive value than solution 2.

9. Both solutions have **high false discovery rate**. Thus, for solution 1 if an event is classified as malicious, 77% times it is misclassified and the event was actually normal (benign). For solution 2, if an event is classified as malicious, 91% of times it is misclassified and the event was actually normal (benign).

10. Both solutions have Cohen Kappa score in the range of 0 and 0.1. Thus, the models **have no agreement with the expected model**.

11. Both solutions have **very small area under the ROC curve**. For solution 1, the ROC curve is below random guessing line. For solution 2, the ROC curve is just above the random guessing line. Thus, both the classifiers are very bad fit.

12. Both solutions have **Precision – Recall curve with small area**. Thus, both models have very poor precision and recall.

4. Multi-class classification using Robust Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced dataset** | | **Imbalanced dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 32 | 25 | 32 | 25 |
| **Confusion matrix** | [[20260, 61, 37, 334], [54, 20631, 0, 7], [24, 0, 20668, 0], [183, 4, 2, 20503]] | [[20294, 51, 24, 323], [44, 20641, 0, 7], [17, 1, 20674, 0], [141, 0, 5, 20546]] | [[809752, 354838, 206919, 65655], [638, 27591, 757, 177], [20021, 0, 348, 323], [150550, 24460, 515, 71557]] | [[1264989, 68723, 24664, 78788], [916, 28178, 36, 33], [16201, 0, 346, 4145], [108326, 2932, 129, 135695]] |
| **Accuracy** | 0.992 | 0.993 | 0.532 | 0.828 |
| **Precision** | 0.993 | 0.994 | 0.137 | 0.488 |
| **Recall** | 0.996 | 0.997 | 0.368 | 0.567 |
| **F1-Score** | 0.994 | 0.995 | 0.199 | 0.525 |
| **Balanced accuracy** | 0.987 | 0.989 | 0.465 | 0.724 |
| **MCC** | 0.978 | 0.981 | -0.051 | 0.422 |
| **NPV** | 0.987 | 0.99 | 0.825 | 0.91 |
| **FDR** | 0.007 | 0.006 | 0.863 | 0.512 |
| **Cohen Kappa** | 0.989 | 0.99 | 0.075 | 0.444 |

**Balanced dataset: -**

1. Both solutions have almost **equal and very high accuracy**. Thus, both models perform correct prediction around 99% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives).

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have **almost equal and very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

6. Both solutions **have almost equal and very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

7. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 99% times it was correct and the event was actually benign (that is not malicious). Solution 2 has higher negative predictive value than solution 1.

8. Both solutions **have very low false discovery rate**. Thus, for both models when an event was classified as malicious, around 0.7% of the times it was incorrect and the event was actually benign. Thus, both the models have noise less than 1%.

9. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

**Imbalanced dataset: -**

1**. Solution 1 has low accuracy**: 0.532 and **solution 2 has relatively higher accuracy**: 0.828. Thus, overall correctness of model for solution 2 is better than the model for solution 1.

2. Both solutions **have low precision**. Thus, the model incorrectly classifies many normal events as malicious, resulting in false positives.

3. Both solutions have **low recall**. Thus, the model incorrectly classifies many malicious events as benign.

4. Both solutions have **low F1-score**. This is because the model’s performance drops while correctly predicting both normal and malicious events.

5. **Solution 1 has low balanced accuracy**: 0.465, and **solution 2 has relatively higher balanced accuracy**: 0.724. Although solution 2 has lower precision and recall and still has higher balanced accuracy. It may be due to imbalanced nature of the dataset and its higher accuracy also supports the results.

6. Both solutions have **low MCC**, closer to 0. MCC for solution 1 is negative and closer to -1, which indicates there is strong disagreement between the model’s predictions and the actual labels. MCC for solution 2 is positive and closer to 0, which indicates that the model performs similar to random guessing.

7. Both solutions **have high negative predictive value**. Thus, for both models when an event was classified as benign, for solution 1 it was correct 82.5% times and for solution 2 it was correct 91% times and the event was actually benign (that is not malicious). Solution 2 has higher negative predictive value than solution 1.

8. **Solution 1 has high false discovery rate**, and **solution 2 has low false discovery rate**. Thus, for solution 1 if an event is classified as malicious, 86.3% times it is misclassified and the event was actually normal (benign). For solution 2, if an event is classified as malicious, 51.2% of times it is misclassified and the event was actually normal (benign). As the result, model for solution 2 was better than the model for solution 1.

9. Solution 1 has Cohen Kappa score less than 0.1, thus, the **model for solution 1 has** **no agreement with the expected model**. Solution 2 has Cohen Kappa score in the range of 0.41 to 0.60, thus, the **model for solution 2 has** **moderate agreement with the expected model**.

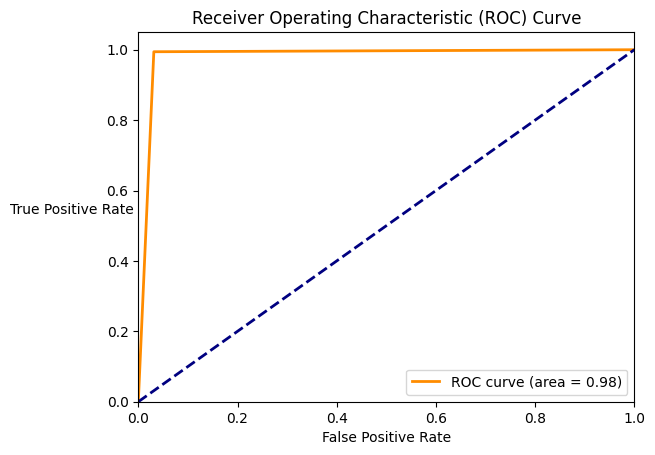
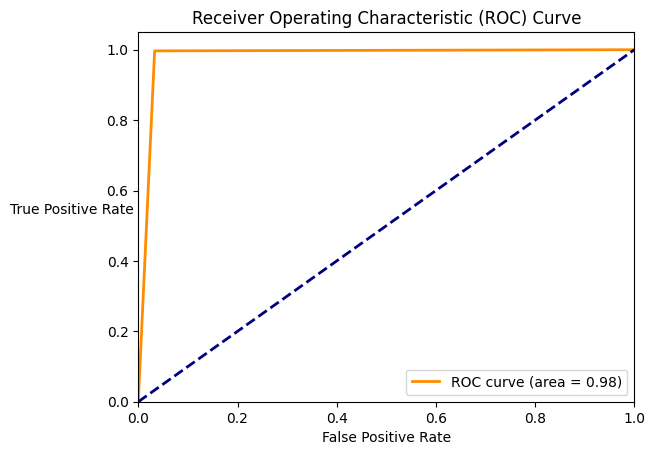
**Evaluation of results obtained using Flower Pollination Algorithm**

1. Binary classification using Standard Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced dataset** | | **Imbalanced dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 20 | 25 | 20 | 25 |
| **Confusion matrix** | [[20045, 647], [356, 61720]] | [[20012, 680], [206, 61870]] | [[1401121, 36043], [69228, 227649]] | [[1400635, 36529], [60777, 236160]] |
| **Accuracy** | 0.988 | 0.989 | 0.939 | 0.944 |
| **Precision** | 0.99 | 0.989 | 0.863 | 0.866 |
| **Recall** | 0.994 | 0.997 | 0.767 | 0.795 |
| **F1-Score** | 0.992 | 0.993 | 0.812 | 0.829 |
| **AUC Score** | 0.981 | 0.982 | 0.871 | 0.885 |
| **Balanced accuracy** | 0.981 | 0.982 | 0.871 | 0.885 |
| **MCC** | 0.968 | 0.971 | 0.778 | 0.797 |
| **NPV** | 0.983 | 0.99 | 0.953 | 0.958 |
| **FDR** | 0.01 | 0.011 | 0.137 | 0.134 |
| **Cohen Kappa** | 0.968 | 0.971 | 0.776 | 0.796 |

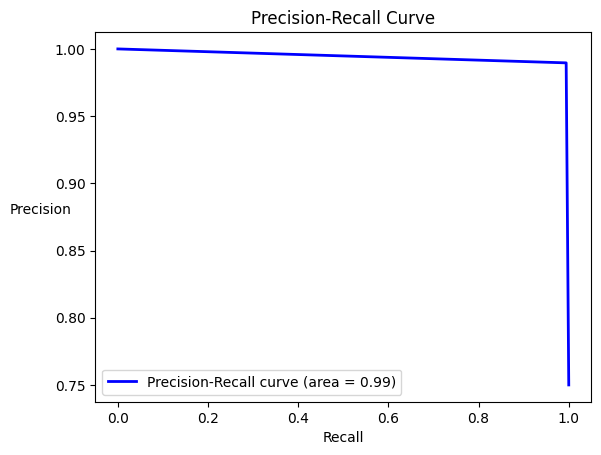
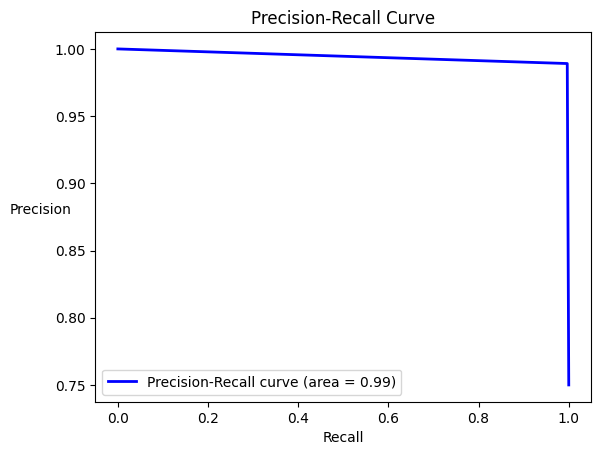
**Balanced test dataset: Comparison of ROC curves**

**Solution 1 Solution 2**

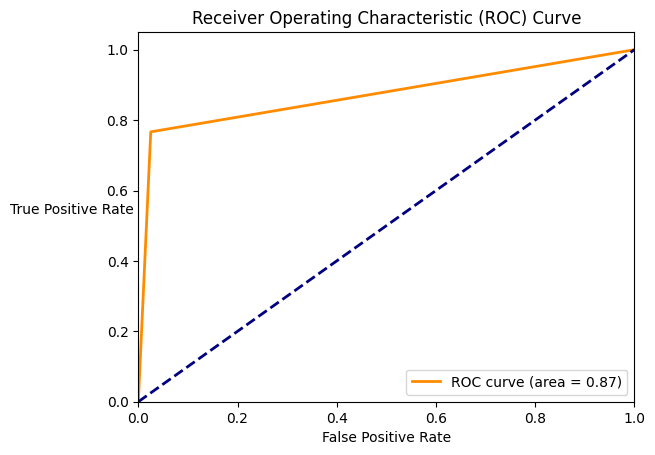
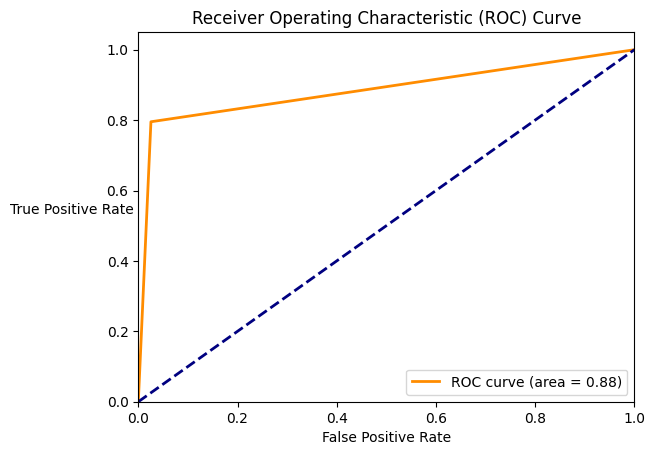
**Balanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**

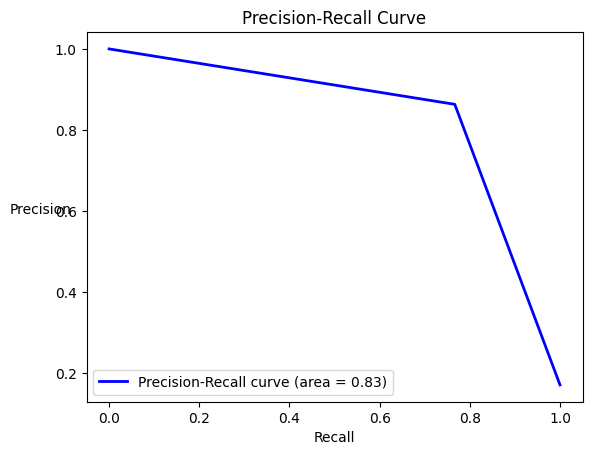
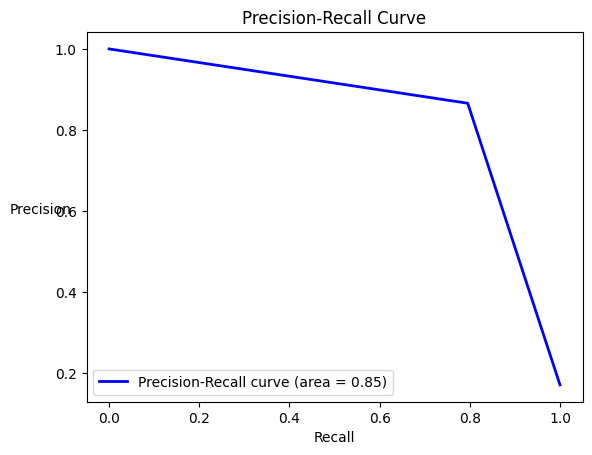
**Imbalanced test dataset: Comparison of ROC curves**

**Solution 1 Solution 2**

**Imbalanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**

**Balanced dataset: -**

1. Both solutions have almost **equal and very high accuracy**. Thus, both models perform correct prediction around 98% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less. Solution 1 has slightly higher precision than solution 2.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives). Solution 2 has slightly higher recall than solution 1.

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have almost **equal and** **very high AUC scores** which is closer to 1. Thus, both models have great ability to differentiate between benign and malicious events.

6. Both solutions have **almost equal and very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

7. Both solutions **have almost equal and very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

8. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 98% times it was correct and the event was actually benign (that is not malicious). Solution 2 has slightly higher negative predictive value than solution 1.

9. Both solutions **have very low false discovery rate**. Thus, for both models when an event was classified as malicious, around 1% of the times it was incorrect and the event was actually benign. Thus, both the models have noise close to 1%.

10. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

11. Both solutions **have ROC curve closer to axes, and the elbow is closer to coordinate (0, 1)** which indicates both the models have larger value of area in their respective ROC curve. Thus, both the models are good classifiers.

12. Both solutions have **Precision – Recall curve have very high value for both precision and recall**. Thus, both the models are good classifiers.

**Imbalanced dataset: -**

1. Both the solutions **have a very high accuracy**. Thus, both models perform correct prediction around 94% of the times.

2. Both solutions **have high precision**, greater than 86%. Thus, the model incorrectly classifies some of the normal events as malicious, resulting in false positives.

3. Both solutions have **average recall**, greater than 75%. Thus, the model incorrectly classifies multiple malicious events as benign. Solution 2 has better recall than solution 1.

4. Both solutions have **high F1-score**, around 82%. This is because the model’s performance drops while correctly predicting both normal and malicious events.

5. Both solutions have **high AUC score**, around 87%. Thus, both models have strong ability to differentiate between benign and malicious events.

6. Both solutions have **high balanced accuracy**: 87%. Thus, both models have high precision and recall.

7. Both solutions have **average MCC**, around 79%. MCC score is greater than 0 and closer to 1. Thus, both models have some agreement with the actual labels.

8. Both solutions **have very high negative predictive value** around 95%. Thus, for both models when an event was classified as benign, around 95% times it was correct and the event was actually benign (that is not malicious). Solution 2 has slightly higher negative predictive value than solution 1.

9. Both solutions have **low false discovery rate** around 13%. Thus, for solution 1 if an event is classified as malicious, 13% times it is misclassified and the event was actually normal (benign). For solution 2, if an event is classified as malicious, 13% of times it is misclassified and the event was actually normal (benign).

10. Both solutions have Cohen Kappa score in the range of 0.61 and 0.80. Thus, the models **have substantial agreement with the expected model**.

11. Both solutions **have ROC curve closer to axes, and the elbow is closer to coordinate (0, 0.8)** which indicates both the models have relatively large value of area in their respective ROC curve and are medium fit.

12. Both solutions have **Precision – Recall curve with high area**. The area of P-R curve for solution 2 is greater than the area of P-R curve for solution 1. Thus, model for solution 2 performs better than the model for solution 1.

2. Multi-class classification using Standard Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced dataset** | | **Imbalanced dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 14 | 19 | 14 | 19 |
| **Confusion matrix** | [[19980, 58, 17, 637], [145, 20539, 0, 8], [33, 0, 20659, 0], [106, 1, 1, 20584]] | [[19954, 205, 25, 508], [133, 20555, 0, 4], [71, 0, 20621, 0], [307, 1, 2, 20382]] | [[1398768, 10113, 2113, 26170], [589, 28570, 0, 4], [527, 0, 20165, 0], [[96845, 563, 20, 149654]] | [[1416627, 10478, 2477, 7582], [329, 28828, 0, 6], [496, 1, 20195, 0], [64697, 471, 24, 181890]] |
| **Accuracy** | 0.988 | 0.985 | 0.921 | 0.95 |
| **Precision** | 0.989 | 0.988 | 0.838 | 0.918 |
| **Recall** | 0.995 | 0.992 | 0.669 | 0.779 |
| **F1-Score** | 0.992 | 0.99 | 0.744 | 0.843 |
| **Balanced accuracy** | 0.981 | 0.978 | 0.821 | 0.882 |
| **MCC** | 0.968 | 0.96 | 0.705 | 0.818 |
| **NPV** | 0.986 | 0.975 | 0.935 | 0.956 |
| **FDR** | 0.011 | 0.012 | 0.162 | 0.082 |
| **Cohen Kappa** | 0.984 | 0.98 | 0.707 | 0.819 |

**Balanced dataset: -**

1. Both solutions have almost **equal and very high accuracy**. Thus, both models perform correct prediction around 98% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives).

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have **almost equal and very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

6. Both solutions **have almost equal and very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

7. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 98% times it was correct and the event was actually benign (that is not malicious). Solution 1 has slightly higher negative predictive value than solution 2.

8. Both solutions **have very low false discovery rate**. Thus, for both models when an event was classified as malicious, around 1% of the times it was incorrect and the event was actually benign. Thus, both the models have noise close to 1%.

9. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

**Imbalanced dataset: -**

1. Both the solutions **have a very high accuracy**. Thus, both models perform correct prediction around 93% of the times.

2. Both solutions **have high precision**, greater than 84%. Thus, the model incorrectly classifies some of the normal events as malicious, resulting in false positives. Model for solution 2 has higher precision than the model for solution 1.

3. Both solutions have **low recall**, model for solution 1 has recall around 67% and model for solution 2 has recall around 78%. Thus, the model incorrectly classifies many malicious events as benign.

4. Both solutions have **low F1-score**, model for solution 1 has F1-score around 74% and model for solution 2 has F1-score around 84%. This is because the model’s performance drops while correctly predicting both normal and malicious events.

5. Both solutions have almost **equal and high balanced accuracy** around 82%. Thus, both models have high precision and recall. However, this may be observed due to imbalanced nature of the dataset and high accuracy of both models.

6. Both solutions have **low MCC**, model for solution 1 has MCC around 70% and model for solution 2 has MCC around 82%. MCC score is greater than 0 and closer to 1. Thus, both models have some agreement with the actual labels.

7. Both solutions **have very high negative predictive value** around 94%. Thus, for both models when an event was classified as benign, around 94% times it was correct and the event was actually benign (that is not malicious). Solution 2 has slightly higher negative predictive value than solution 1.

8. Both solutions have **low false discovery rate**. Thus, for solution 1 if an event is classified as malicious, 16% times it is misclassified and the event was actually normal (benign). For solution 2, if an event is classified as malicious, 8% of times it is misclassified and the event was actually normal (benign). As the result, model for solution 2 was better than the model for solution 1.

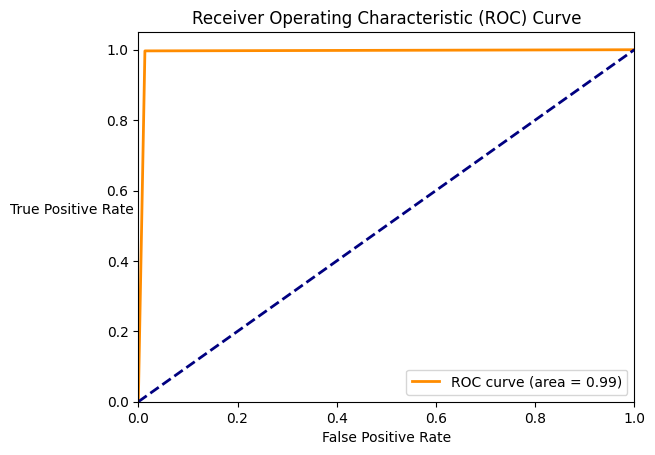
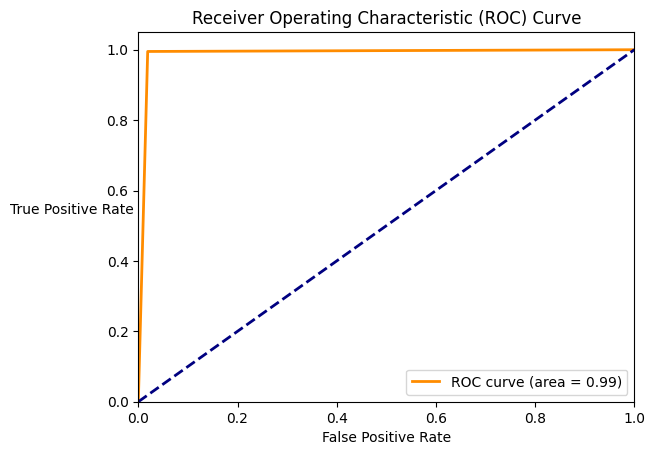
9. Solution 1 has Cohen Kappa score in the range of 0.61 to 0.80, thus, **for solution 1 model there is substantial agreement with the expected model**. Solution 2 has Cohen Kappa score in the range of 0.81 to 0.99, thus, **for solution 2 model there is near perfect agreement with the expected model**.

3. Binary classification using Robust Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced dataset** | | **Imbalanced dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 18 | 18 | 18 | 18 |
| **Confusion matrix** | [[20301, 391], [307, 61769]] | [[20417, 275], [207, 61869]] | [[1245124, 192040], [242030, 54907]] | [[975308, 461856], [177082, 119855]] |
| **Accuracy** | 0.992 | 0.994 | 0.75 | 0.632 |
| **Precision** | 0.994 | 0.996 | 0.222 | 0.206 |
| **Recall** | 0.995 | 0.997 | 0.185 | 0.404 |
| **F1-Score** | 0.994 | 0.996 | 0.202 | 0.273 |
| **AUC Score** | 0.988 | 0.992 | 0.526 | 0.541 |
| **Balanced accuracy** | 0.988 | 0.992 | 0.526 | 0.541 |
| **MCC** | 0.977 | 0.984 | 0.055 | 0.066 |
| **NPV** | 0.985 | 0.99 | 0.837 | 0.846 |
| **FDR** | 0.006 | 0.004 | 0.778 | 0.794 |
| **Cohen Kappa** | 0.977 | 0.984 | 0.055 | 0.06 |

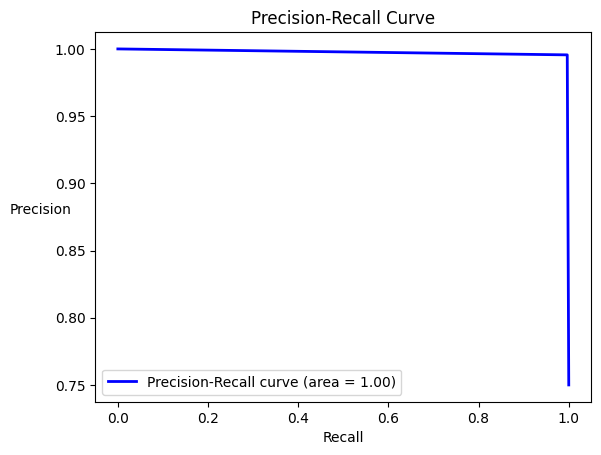
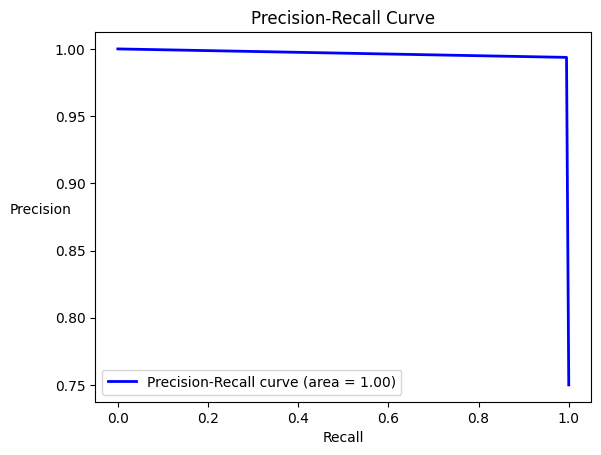
**Balanced test dataset: Comparison of ROC curves**

**Solution 1 Solution 2**



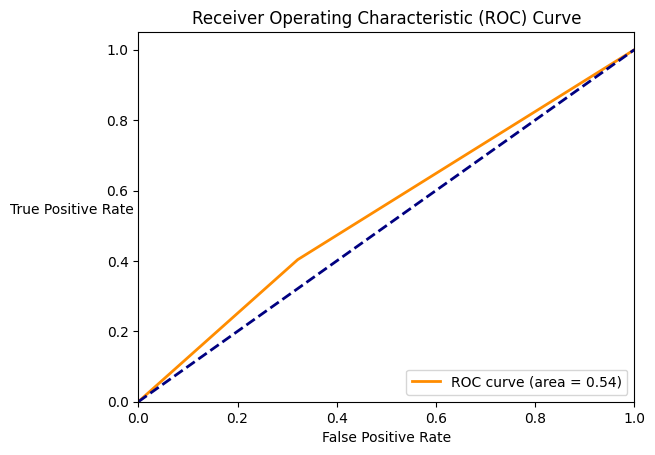
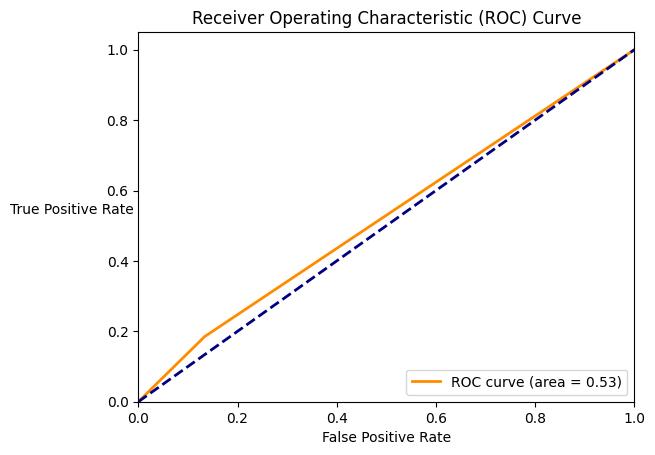
**Balanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**



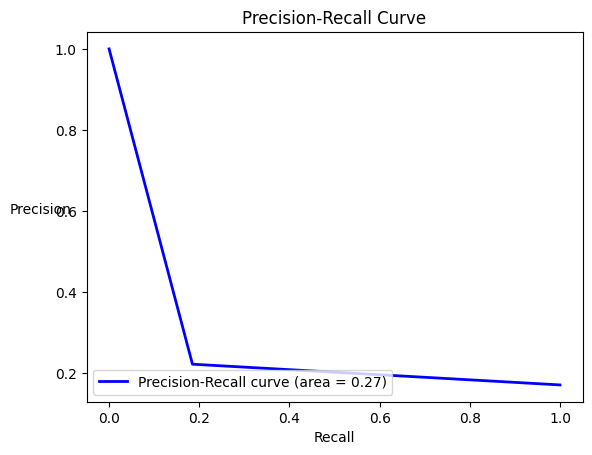
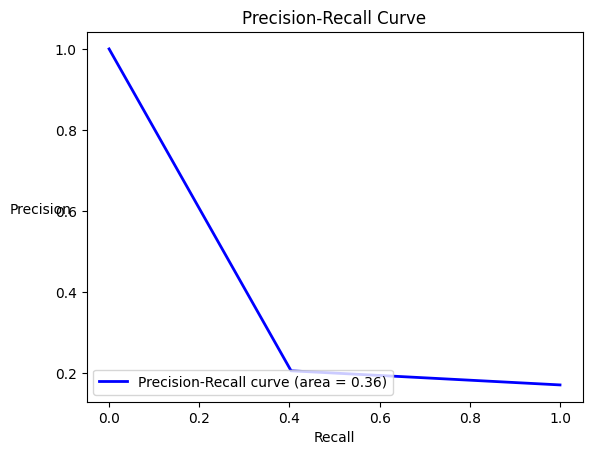
**Imbalanced test dataset: Comparison of ROC curves**

**Solution 1 Solution 2**



**Imbalanced test dataset: Comparison of Precision – Recall curves**

**Solution 1 Solution 2**

**Balanced dataset: -**

1. Both solutions have **very high accuracy**. Thus, both models perform correct prediction around 99% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less. Solution 2 has slightly higher precision than solution 1.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives). Solution 2 has slightly higher recall than solution 2.

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have almost **very high AUC scores** which is closer to 1. Thus, both models have great ability to differentiate between benign and malicious events.

6. Both solutions have **very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

7. Both solutions **have very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

8. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 98% times it was correct and the event was actually benign (that is not malicious). Solution 2 has slightly higher negative predictive value than solution 1.

9. Both solutions **have very low false discovery rate**. Thus, for both models when an event was classified as malicious, less than 0.6% of the times it was incorrect and the event was actually benign. Thus, both the models have noise less than 0.6%.

10. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

11. Both solutions **have ROC curve closer to axes, and the elbow is closer to coordinate (0, 1)** which indicates both the models have larger value of area in their respective ROC curve. Thus, both the models are good classifiers.

12. Both solutions have **Precision – Recall curve have very high value for both precision and recall**. Thus, both the models are good classifiers.

**Imbalanced dataset: -**

1. Both the solutions **have a very low accuracy**. Thus, both models perform correct prediction in the range of 60 to 75%.

2. Both solutions **have very low precision**. Thus, the model incorrectly classifies many normal events as malicious, resulting in false positives. Model for solution 1 has better precision than the model for solution 2.

3. Both solutions **have very low recall**. Thus, the model incorrectly classifies multiple malicious events as benign. Model for solution 2 has better recall than the model for solution 1.

4. Both solutions have **low F1-score.** This is because the model’s performance is significantly low while correctly predicting both normal and malicious events.

5. Both solutions have **low AUC score**. Thus, both models are inconsistent to differentiate between benign and malicious events.

6. Both solutions have **low balanced accuracy**. Thus, both models have low precision and recall.

7. Both solutions have **low MCC**. MCC score for both the solutions is positive and closer to 0, thus their performance is very close to random guessing.

8. Both solutions **have high negative predictive value**. Model for both solutions have NPV around 84% Thus, for both models when an event was classified as benign, around 84% times it was correct and the event was actually benign (that is not malicious). Solution 2 has slightly higher negative predictive value than solution 1.

9. Both solutions have **average false discovery rate**. Thus, for solution 1 if an event is classified as malicious, around 78% times it is misclassified and the event was actually normal (benign). For solution 2, if an event is classified as malicious, 79% of times it is misclassified and the event was actually normal (benign).

10. Both solutions have Cohen Kappa score in the range of 0 and 0.1. Thus, the models **have no agreement with the expected model**.

11. Both solutions have **very small area under the ROC curve**. For solution 1, the ROC curve is very close to random guessing line. For solution 2, the ROC curve has relatively more distance from the random guessing line. Thus, both the classifiers are very bad fit.

12. Both solutions have **Precision – Recall curve with very small area**. Thus, both models have very poor precision and recall.

4. Multi-class classification using Robust Scaler

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Balanced dataset** | | **Imbalanced dataset** | |
| **Solution 1** | **Solution 2** | **Solution 1** | **Solution 2** |
| **Number of features** | 28 | 25 | 28 | 25 |
| **Confusion matrix** | [[20339, 41, 25, 287], [117, 20572, 0, 3], [48, 0, 20644, 0], [163, 2, 6, 20521]] | [[20349, 51, 20, 272], [77, 20611, 0, 4], [14, 0, 20678, 0], [138, 5, 2, 20547]] | [[833874, 161274, 35569, 406447], [461, 26706, 5, 1991], [16542, 0, 0, 4150], [145300, 3734, 60, 97988]] | [[712300, 134029, 265956, 324879], [578, 27464, 68, 1053], [19610, 0, 349, 733], [91661, 2172, 859, 152390]] |
| **Accuracy** | 0.992 | 0.993 | 0.556 | 0.516 |
| **Precision** | 0.994 | 0.994 | 0.171 | 0.199 |
| **Recall** | 0.995 | 0.996 | 0.434 | 0.617 |
| **F1-Score** | 0.995 | 0.995 | 0.246 | 0.301 |
| **Balanced accuracy** | 0.989 | 0.99 | 0.507 | 0.556 |
| **MCC** | 0.978 | 0.982 | 0.011 | 0.085 |
| **NPV** | 0.984 | 0.989 | 0.837 | 0.864 |
| **FDR** | 0.006 | 0.006 | 0.829 | 0.801 |
| **Cohen Kappa** | 0.989 | 0.991 | 0.068 | 0.138 |

**Balanced dataset: -**

1. Both solutions have almost **equal and very high accuracy**. Thus, both models perform correct prediction around 99% of the times.

2. Both solutions have **very high precision**, thus, the false positives for both models were less.

3. Both solutions have **very high recall**, thus, both models correctly classify most of the malicious events (True Positives).

4. Both solutions have F1-score close to 1, thus, both models **have good performance**.

5. Both solutions have **almost equal and very high balanced accuracy** which is closer to 1. Thus, both models have high precision and recall.

6. Both solutions **have very high MCC** which is closer to 1. Thus, both models have very close agreements with actual labels of each event.

7. Both solutions **have very high negative predictive value**. Thus, for both models when an event was classified as benign, around 98% times it was correct and the event was actually benign (that is not malicious). Solution 2 has slightly higher negative predictive value than solution 1.

8. Both solutions **have equal and very low false discovery rate**. Thus, for both models when an event was classified as malicious, around 0.6% of the times it was incorrect and the event was actually benign. Thus, both the models have noise less than 1%.

9. Both solutions have Cohen Kappa in the range of 0.81 to 0.99. Thus, the models **have near perfect agreement** and are closer to the expected model.

**Imbalanced dataset: -**

1**.** Both solution 1 and solution 2 **have very low accuracy.** Solution 1 has accuracy around 55% and solution 2 has accuracy around 51%. Thus, overall correctness of model for solution 1 is better than the model for solution 2.

2. Both solutions **have low precision**. Thus, the model incorrectly classifies many normal events as malicious, resulting in false positives.

3. Both solutions have **low recall**. Thus, the model incorrectly classifies many malicious events as benign.

4. Both solutions have **low F1-score**. This is because the model’s performance drops while correctly predicting both normal and malicious events.

5. **Solution 1 has low balanced accuracy**: 0.507, and **solution 2 has relatively higher balanced accuracy**: 0.556. Although solution 2 has lower precision and recall and still has higher balanced accuracy. It may be due to imbalanced nature of the dataset and its higher accuracy also supports the results.

6. Both solutions have **low MCC**, closer to 0. MCC values for both solutions are positive and closer to 0, which indicates that both models perform similar to random guessing.

7. Both solutions **have high negative predictive value**. Thus, for both models when an event was classified as benign, for solution 1 it was correct around 83% times and for solution 2 it was correct around 86% times and the event was actually benign (that is not malicious). Solution 2 has higher negative predictive value than solution 1.

8. Both solutions **have high false discovery rate**. Thus, for both models when if an event is classified as malicious, greater than 80% times it is misclassified and the event was actually normal (benign). As the result both models generate lots of noise.

9. Solution 1 has Cohen Kappa score less than 0.1, thus, the **model for solution 1 has** **no agreement with the expected model**. Solution 2 has Cohen Kappa score in the range of 0.10 to 0.20, thus, the **model for solution 2 has** **slight** **agreement with the expected model**.

**Comparison of results between Standard Scaler and Robust Scaler**

**Comparison of results between Balanced test datasets and Imbalanced test datasets**

**Comparison of results between Artificial Bee Colony algorithm and Flower Pollination Algorithm**

**Comparison of results obtained with the models in literature survey.**