# Report:

**Genetic Algorithm for Network Intrusion Classification:**

# Evaluation Summary

The table below shows the evaluation results of the genetic algorithm over five generations with various configurations. The aim was to select a subset of features from the UNSW-NB15 dataset that provides the highest classification accuracy using a logistic regression classifier.

|  |  |  |
| --- | --- | --- |
| **Generation:1** | **Best Fitness Score** | **Selected Features Count** |
| 0 | 0.8779263737203798 | proto', 'service', 'state |
| 1 | 0.8911289172773674 | proto', 'service', 'state |
| 2 | 0.8917562519604209 | proto', 'service', 'state |
| 3 | 0.900253785394508 | proto', 'service', 'state |
| 4 | 0.904302945621489 | proto', 'service', 'state |

|  |  |  |
| --- | --- | --- |
| **Generation:2** | **Best Fitness Score** | **Selected Features Count** |
| 0 | 0.8719952094442385 | proto', 'service', 'state |
| 1 | 0.8845704183181727 | proto', 'service', 'state |
| 2 | 0.8977729618751604 | proto', 'service', 'state |
| 3 | 0.9016510308249451 | proto', 'service', 'state |
| 4 | 0.902449456785195 | proto', 'service', 'state |

|  |  |  |
| --- | --- | --- |
| **Generation:3** | **Best Fitness Score** | **Selected Features Count** |
| 0 | 0.8680030796429895 | proto', 'service', 'state |
| 1 | 0.8838860532093872 | proto', 'service', 'state |
| 2 | 0.8897031566340643 | proto', 'service', 'state |
| 3 | 0.8926117083464028 | proto', 'service', 'state |
| 4 | 0.8970885967663749 | proto', 'service', 'state |

# Methodology

# Fitness Function:

The fitness function used for this genetic algorithm was classification accuracy. It was chosen because the primary objective of the model is to accurately classify network intrusions, and accuracy is a direct measure of this**.**

# Selection:

A roulette wheel selection method was used to select individuals for reproduction based on their fitness proportionate to the population**.**

# Crossover:

A single-point crossover method was employed to combine pairs of parents and create offspring for the next generation**.**

Mutation:

Random bit-flip mutation was applied to the offspring to introduce genetic diversity and explore more of the feature space**.**

# Classifier:

Logistic regression was utilized for its simplicity and effectiveness in binary classification tasks**.**

# Insights and Discussion

The genetic algorithm successfully identified a subset of features from the dataset that provided high classification accuracy. As the generations progressed, there was an improvement in the fitness score, indicating that the algorithm was effectively evolving the population toward better solutions. It was observed that certain features consistently appeared in the selected subset, suggesting their importance in classifying network intrusions.

The implementation choices, including the type of genetic operators and the fitness function, were made to balance the exploration and exploitation in the genetic search process. The logistic regression model was chosen for its computational efficiency, allowing for rapid evaluation of individuals within the population. Future work could include testing different classifiers, fine-tuning the genetic algorithm's parameters, and incorporating feature engineering techniques to potentially improve performance

# Conclusion

The genetic algorithm proved to be a potent method for feature selection in the intrusion classification task. The evolution of the population over successive generations led to the discovery of feature subsets that enhanced the model's accuracy. This approach can be extended to other feature selection and optimization problems within the field of cybersecurity and beyond.