INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

VASISHT DUDDU 2015137

SHUBHAM KHANNA 2015179

ANUBHAV JAIN 2015129

MOTIVATION

- Increasing malware complexity and sophistication
- Polymorphic and metamorphic malware change form dynamically and cannot be detected by traditional anti-virus
- Network traffic anomaly detection crucial for preventing certain attacks
- Hard Coded and Rule Based IDS not applicable
- Require to adapt defences to detect attacks based on past data

MALWARE DETECTION

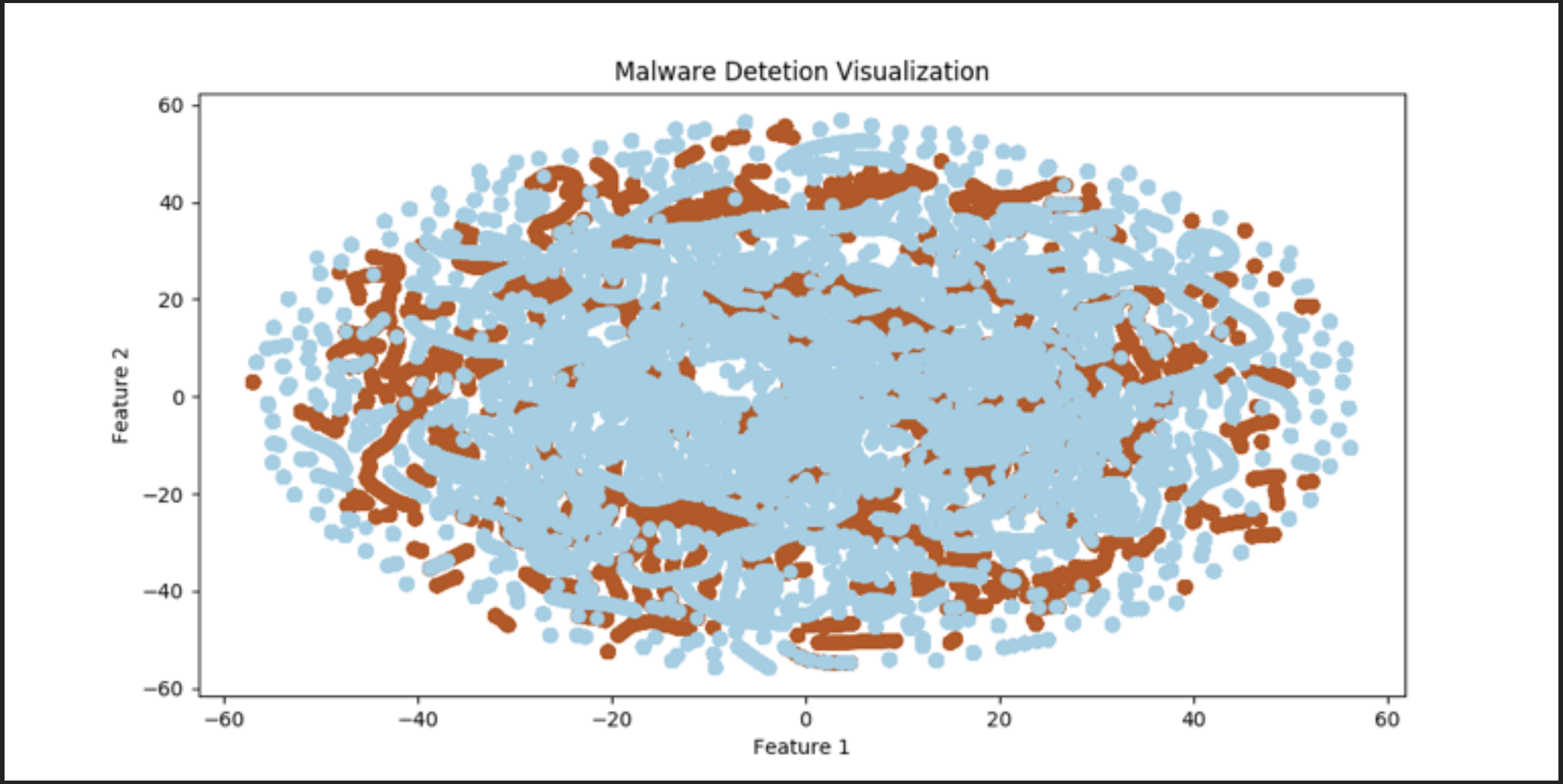
RELATED WORK

Source	Model	Accuracy	False Positives
Tek	Random Forest	99.35	0.56
Adobe	Random Forest	98.21	6.7

Data set used by Adobe is more extensive and contains more features for complex malware

- Malware classification of PE32 executables
- Siddiqui et al. Accuracy : 94%
- Schultz et al. Accuracy : 97.76%
- Shafiq et al. Accuracy : 99%
- Ye et al. Accuracy : 92%
- Ye et al. Accuracy : 93.8%

DATA SET



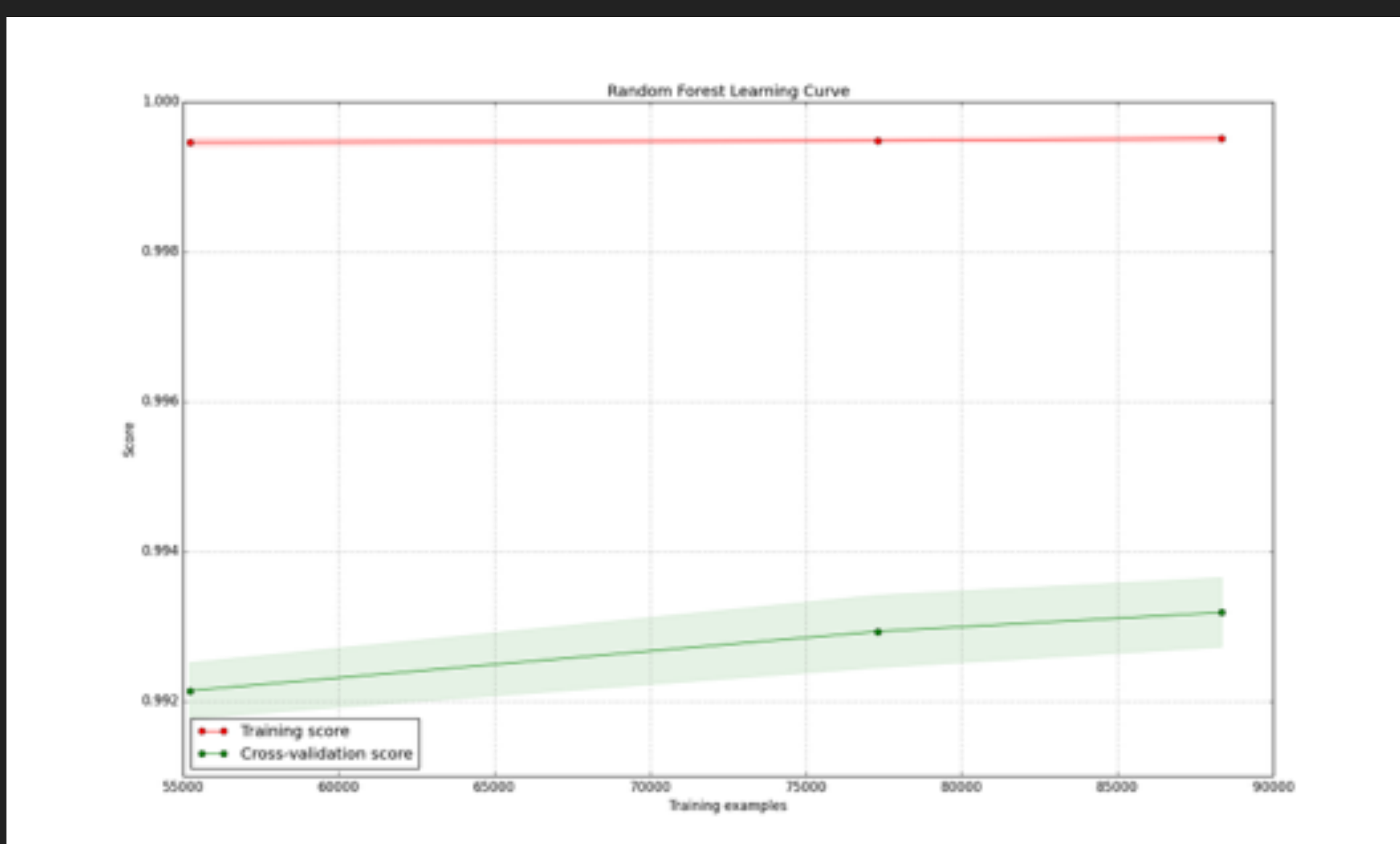
- Total Size : 138047 (Data is Non-Separable)
- Training Size : 96632 Testing Size : 41415
- Features: 54 After Feature Extraction : 14

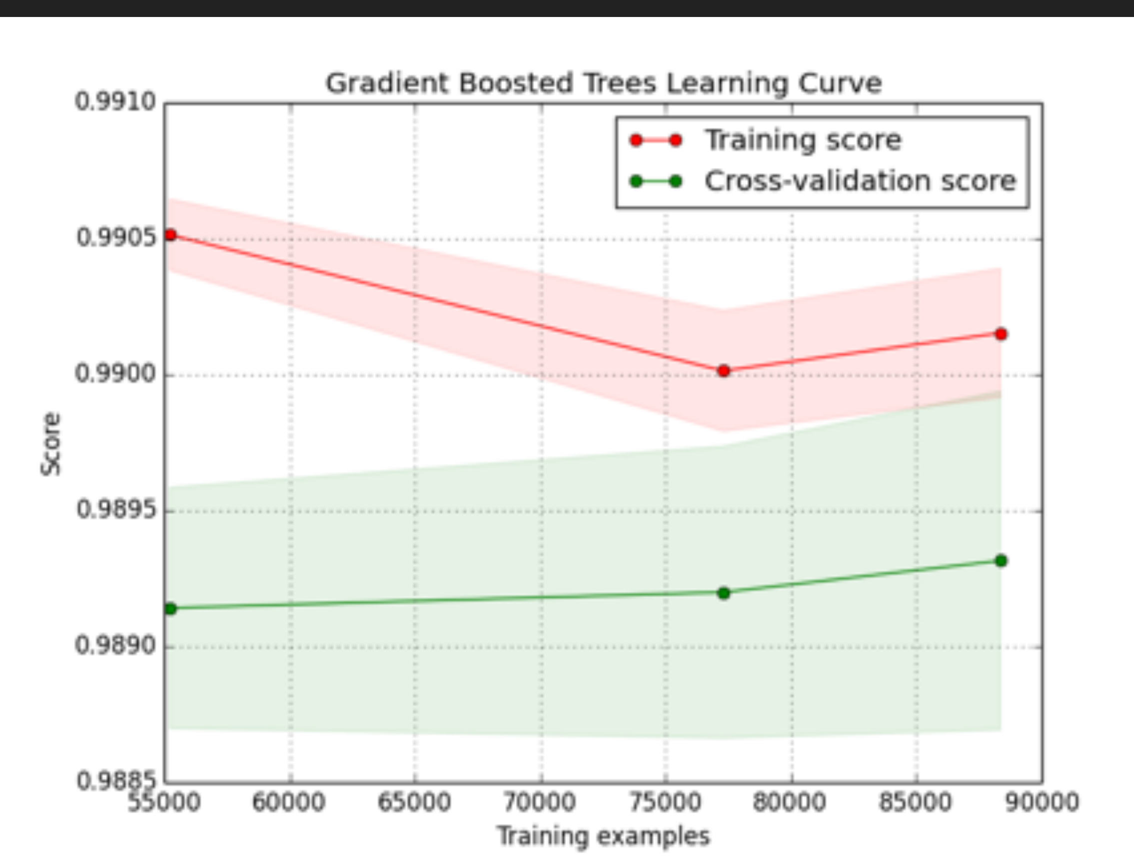
EVALUATION METRICS

- Classification Accuracy
- False Positive Percent (Evaluated using Confusion Matrix)
- AUC Score from ROC Curve
- Threshold tuning for better specificity vs sensitivity

ANALYSIS

- Objective: Reduce False positives maintaining a good accuracy
- Model Selection: Tried SVM, Adaboost, Logistic Regression, Random Forest and Gradient Boosted Trees
- Feature Selection: Using Tree Classifier to reduce to 14 features
- Tuned each classifier for better results





RESULTS

Model	Accuracy	False Positives	AUC Score
Logistic Regression	30.06	1	0.5
Random Forest	98.22	0.91	0.987
Gradient Boosted Trees	98.45	0.71	0.986

- Gradient Boosted Trees better than Random Forest due to lower false positives
- Logistic Regression and other linear models not suitable for the data
- Ensemble approaches are give better results than other classifiers
- Very close to state of the art (Tek)

FUTURE WORK

- Use UNSW-NB15 data set for IDS
- Training: 175,341 samples; Testing set : 82,332 samples ; 49 features with multiple output classes
- Train deep neural networks(NN) to predict network attacks
- Use different architectures and parameters for NN
- Tuning and Analysis: Grid search, cross validation, randomisation of data set
- Evaluation metrics: Classification Accuracy, False Negative, AUC Score, ROC curve for further adjustments

REFERENCES

- Machine learning for malware detection: https://www.randhome.io/blog/2016/07/16/machine-learning-for-malware-detection/
- Towards Classification of Polymorphic Malware: https://www.blackhat.com/docs/webcastTowardsClassificationofPolymorphicMalware-Final.pdf
- M. Siddiqui, M. C. Wang, and J. Lee. Detecting trojans using data mining techniques. In D. M. A. Hussain, A. Q. K. Rajput, B. S. Chowdhry, and Q. Gee, editors, IMTIC, volume 20 of Communications in Computer and Information Science, pages 400-411
- M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo. Data mining methods for detection of new malicious executables. In Proceedings of the 2001 IEEE Symposium on Security and Privacy
- M. Z. Shafiq, S. M. Tabish, F. Mirza, and M. Farooq. Pe-miner: Mining structural information to detect malicious executables in realtime. In Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection
- Y. Ye, L. Chen, D. Wang, T. Li, Q. Jiang, and M. Zhao. Sbmds: an interpretable string based malware detection system using svm ensemble with bagging. Journal in Computer Virology
- Y. Ye, D. Wang, T. Li, and Ye. Imds: Intelligent malware detection system. In Proccedings of ACM International Conference on Knowlege Discovery and Data Mining (SIGKDD 2007

NETWORK TRAFFIC DETECTION

PREVIOUS WORK

Source	Model	Accuracy	False Positives
Tek	Random Forest	99.35	0.56
Adobe	Random Forest	98.21	6.7

Model	Accuracy	False Positives	AUC Score
Logistic Regression	30.06	1	0.5
Random Forest	98.22	0.91	0.987
Gradient Boosted Trees	98.45	0.71	0.986

- Worked on PE32 Malware Dataset
- Gradient Boosted Trees better than Random Forest due to lower false positives
- Ensemble approaches give better results than other classifiers
- Very close to state of the art (Tek)
- Now, worked on UNSW Dataset for network traffic classification

DATASET

Category	Training Set	Testing Set
Normal	56,000	37,000
Analysis	2,000	677
Backdoor	1,746	583
DoS	12,264	4089

Category	Training Set	Testing Set
Exploits	33,393	11,132
Fuzzers	18,184	6,062
Generic	40,000	18,871
Reconnaissance	10,491	3,496
Shellcode	1,133	378
Worms	130	44
Total	175,341	82,332

- Data set has a hybrid of the real modern normal and the contemporary synthesised attack activities of the network traffic
- Pcap files from Argus and Bro-IDS over 3 networks with 9 attack families
- Port information of source and destination, service, packet count and connection information

RELATED WORK

Source	Model	Accuracy	False Positives
Chowdhury *et al.*[1]	SVM	88.03	4.2
Chowdhury *et al.*[1]	SVM(with processing)	98.76	0.09
Moustafa *et al.*[4]	Expectation-Maximisation clustering	77.2	13.1
Moustafa *et al.*[4]	Logistic Regression	83.0	14.5
Moustafa *et al.*[4]	Naive Bayes	79.5	23.5
Mogal *et al.*[6]	Naive Bayes	99.96	-
Mogal *et al.*[6]	Logistic Regression	99.89	-

EVALUATION

- Objective: Reduce False positives maintaining a good accuracy
- Classification Accuracy
- False Positive Percent (Evaluated using Confusion Matrix)
- AUC Score from ROC Curve
- Used tree based feature selection to extract 6 most important features from 49 features

RESULTS

ID	Architecture	Activation Functions	Accuracy	False Positives	AUC Score	Other
NN1	200,150,50	Sigmoid	88.29	32	83.72	learnrate=0.001
NN2	300,200,150,100,50,10	ReLU	88.21	32.58	83.70	learnrate=0.001
NN3	300,200,150,100,50,10	Tanh	75.55	7.47	79.25	learnrate=0.001
NN4	200,150,150,50,10,2	Sigmoid, ReLU, Tanh, Softmax	85.69	39.57	80.15	learnrate=0.01, Dropout=0.45
NN5	150, 300, 450, 50	Sigmoid, ReLU, Softmax	73.92	11.16	77.19	learnrate=0.0001,learnmom=0.9,dropout=0.45,

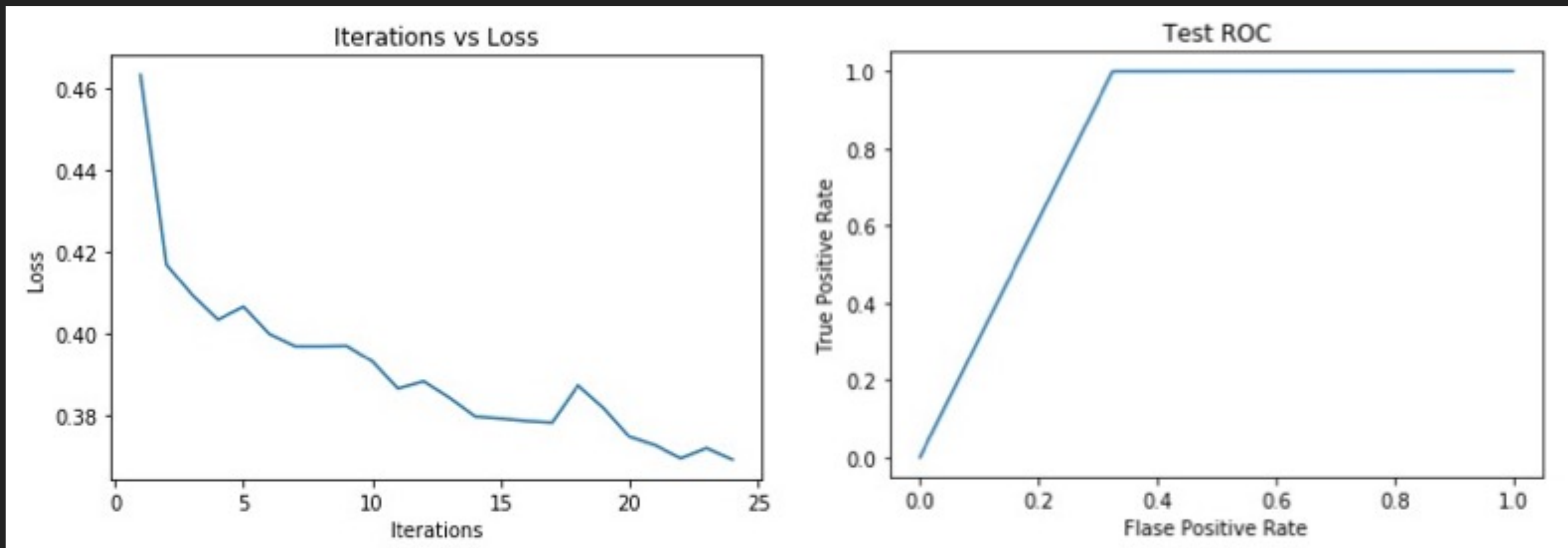
- Some architectures performed better than previous work
- Could not get high accuracy with low false positives
- Results could be improved by iterating for larger number
- Setting the threshold based on ROC can reduce the flase positives

ANALYSIS

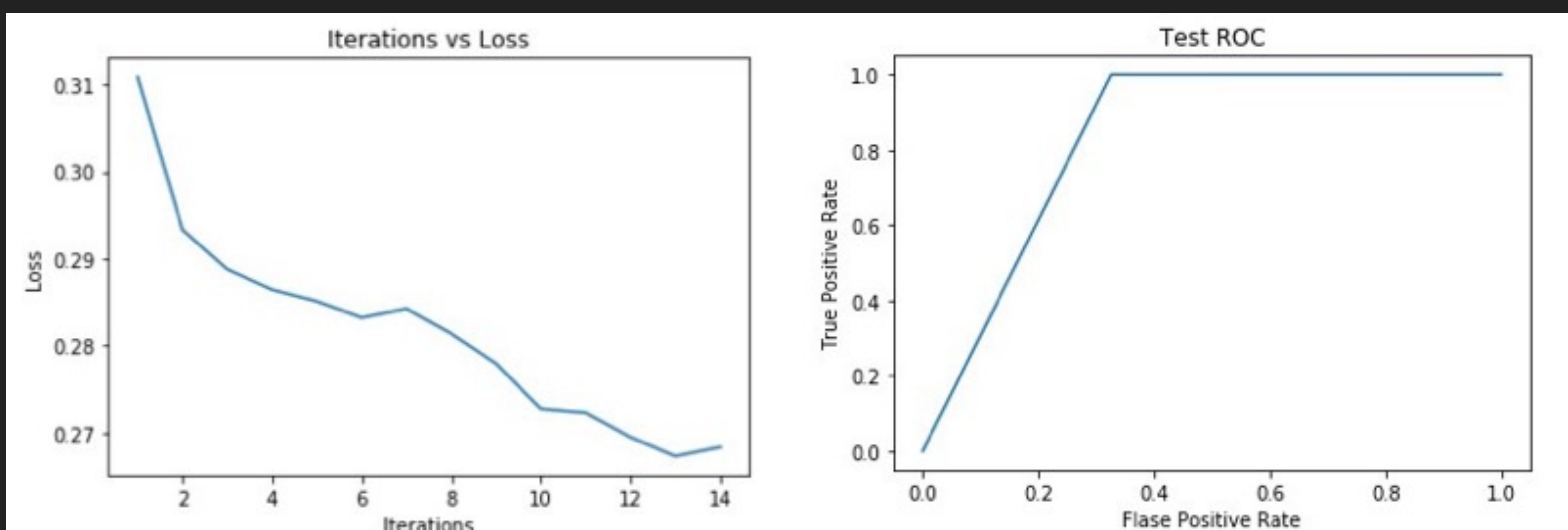
- Different Neural Network Architectures(Varying size and number of hidden layers)
- Pre-Processing: Standard Scalar
- Different Activation Function(Softmax, Sigmoid, ReLU, Tanh)
- Dropout value, learning rate, learning momentum

INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

NEURAL NETWORK 1

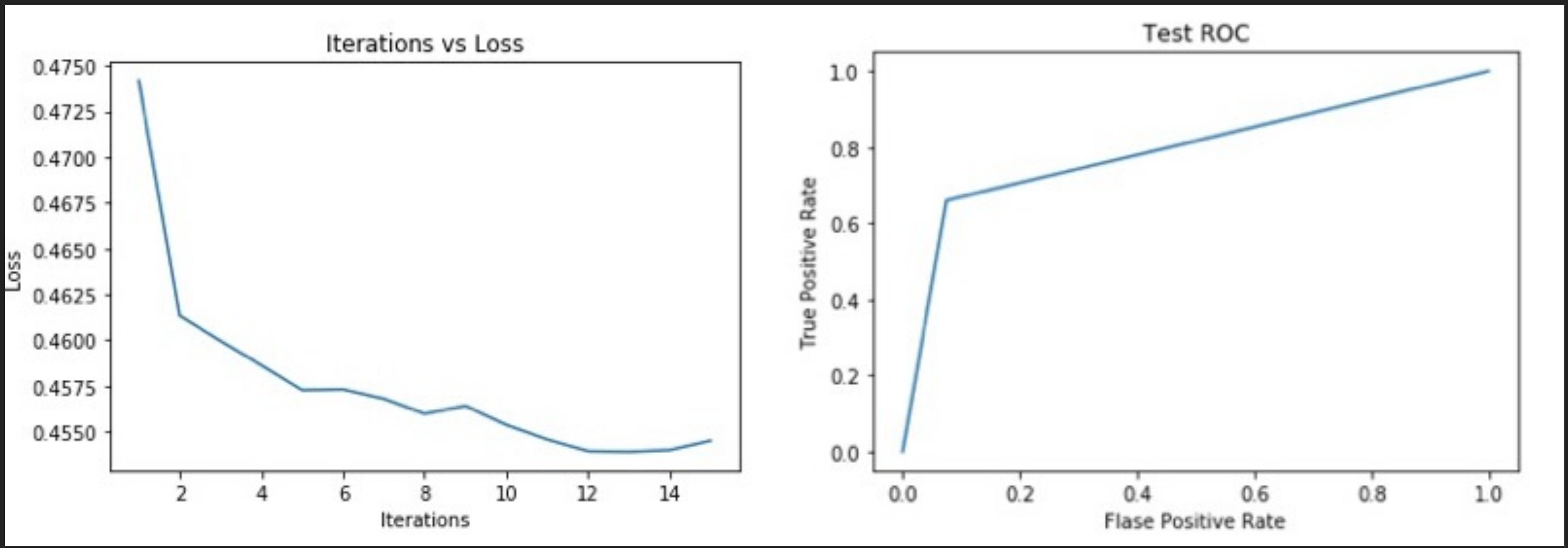


NEURAL NETWORK 2

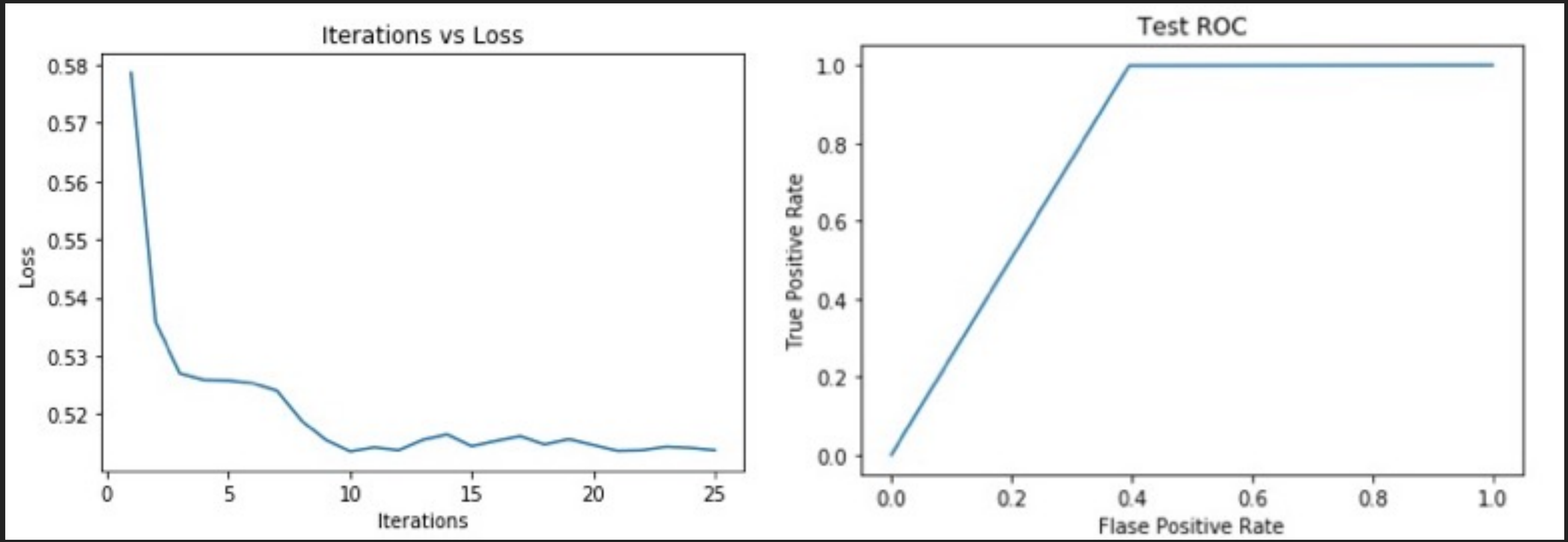


INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

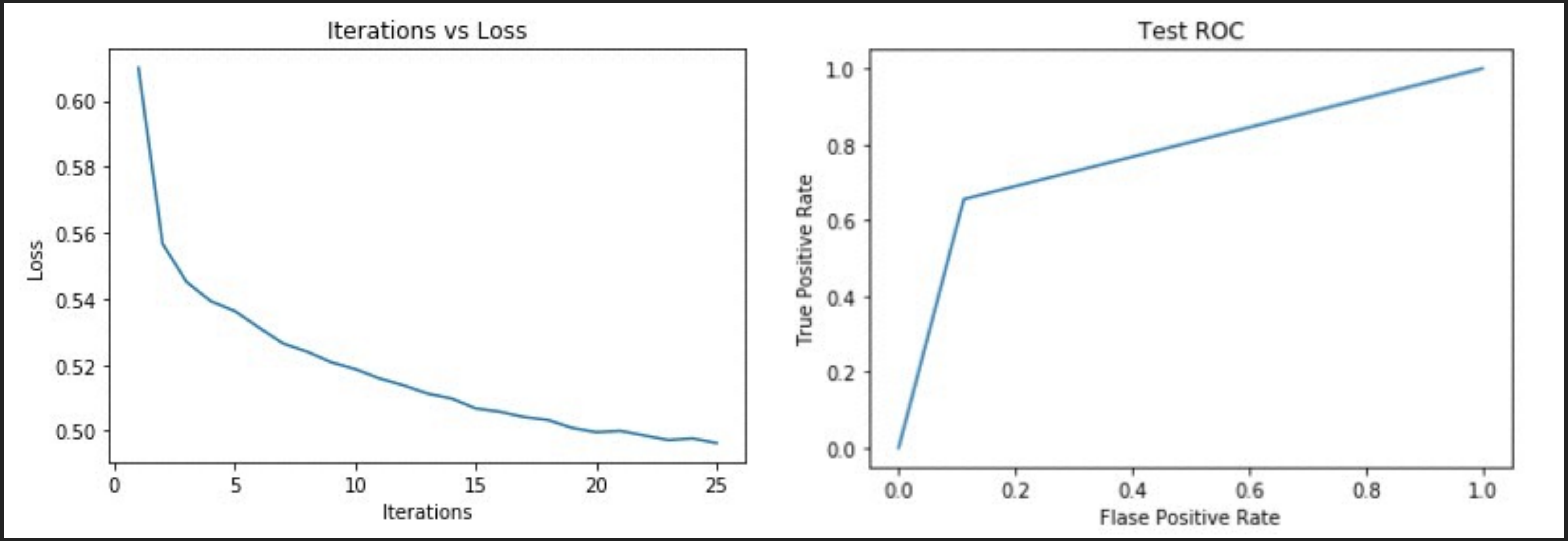
NEURAL NETWORK 3



NEURAL NETWORK 4



NEURAL NETWORK 5



CONTRIBUTIONS

- Vasisht Duddu: Model and parameter selection, feature Extraction, training and analysis for malware and network anomaly dataset
  - NN1.ipynb, NN2.ipynb, NN3.ipynb, NN4.ipynb, NN5.ipynb, gradient_boosted_trees.ipynb, random_forest.ipynb, logistic_regression.ipynb
- Shubham Khanna: Parameter tuning and learning curve analysis for malware dataset
  - learning_curve.py, gradient_boosted_trees_version2.ipynb, random_forest_version2.ipynb, logistic_regression_version2.ipynb
- Anubhav Jain: Data visualisation for malware dataset and data processing for network anomaly dataset
  - visualize.py, Reading data for UNSW Dataset

REFERENCES

- UNSW-NB15 data set: https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/cybersecurity/ADFA-NB15-data sets/
- Nour Moustafa, Jill Slay , “UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems ”
- Nour Moustafa, Jill Slay , “The significant features of the UNSW-NB15 and the KDD99 Data sets for Network Intrusion Detection Systems”
- Nour Moustafa, Jill Slay , “A Hybrid Feature Selection For Network Intrusion Detection Systems: Central Points And Association Rules”
- M.N Chowdhury, K. Ferens, M. Ferens, "Network Intrusion Detection Using Machine Learning"
- D.G. Mogal, S.R Ghungrad, B.B Bhusare, “NIDS using Machine Learning Classifiers on UNSW-NB15 and KDDCUP99 Datasets”