**Project Proposal**

*Hazel Griffith 100997489*

*Puxin Shi 101284355*

**Introduction**

For this project, we will be tackling the GAN-Based Training for Binary Classifier challenge that was a part of the IEEE BigData 2022 conference. This challenge involves training a binary classifier on an unbalanced data set and calculating its RMSE. Then balancing the dataset with synthetic data from a GAN-based solution and retraining the classifier on the now balanced dataset. The competition features two tracks for generating malignant samples: Pictorial-based GAN and Raw Data-based GAN.

**Motivation**

In the field of cyber-security, machine learning techniques have become the predominant method for identifying traces of malware. This is because malware developers have been using more advanced techniques to hide their malware from conventional identification techniques. For example, behavioral polymorphism allows malware to appear different yet retain it’s malicious functions. The machine learning models being trained to identify these illusive programs require real traces of malware to train them to detect future instances of malware. Unfortunately, it can be quite challenging to collect a sufficient number of malware traces to train a binary classifier and the result is that the training datasets can be unbalanced with more instances of benign system calls than those created by malware. If the training dataset is unbalanced, it can bias the classifier model towards classifying new data as the majority class.

**Related Works**

Pioroński and Górecki [1] utilized a Wasserstein GAN with gradient penalty to generate synthetic malware data from the tabular dataset. For classification, they employed the LightGBM model. When trained on the balanced dataset, the classifier achieved a 47.29% reduction in RMSE compared to the baseline model on the test dataset. In another study involving tabular data, Tonmoy and Zaman [2] used Optuna as a hyperparameter tuning method for optimizing the GAN model. As a relatively novel approach in this field, their results demonstrated the utility and effectiveness of Optuna. Similarly, Shin *et al.* [3] generated new samples from tabular data using three GAN models: CTGAN, CTAB-GAN, and complementary GAN. A key distinction from the previous two studies was that they not only generated malign samples from the malign class, but also produced both benign and malign samples from the benign class. Wang *et al.* [4] transformed the tubular data into images and employed a cGAN to generate malign samples. The study revealed that, after the transformation, the patterns for each class became distinct enough that even visual inspection alone could lead to correct classification. Alsheraifi *et al.* [5] explored six GAN models, both tabular-based and pictorial-based, alongside ten classifiers during the classification phrase. Their results indicated that the Random Forest classifier exhibited the most significant improvement when using artificial samples generated from a Vanilla-based GAN (LeakyReLU) with tabular data. Collectively, these studies demonstrate that both image-based and tabular-based GANs can effectively generate new synthetic samples, while still highlighting opportunities for testing additional classifiers to enhance classification performance.

**Objectives**

To solve this problem, generative adversarial networks (GANs) have been used to create synthetic data resembling malware traces. GANs have been successful at recreating any distribution of data to create new data points within the distribution without it appearing out of place. This would solve the problem of having an unbalanced training dataset as a GAN model could be trained to create new synthetic malware traces. The synthetic malware traces would be combined with the real ones within the training dataset to balance the positive and negative classes. Our objective then is to train a binary classifier on the real unbalanced dataset and measure its root mean squared error (RMSE) on the testing data. Then we would train a GAN model to generate synthetic data to balance the training dataset, and then train the same classifier on the balanced dataset with synthetic data included. We will determine its success by measuring the RMSE of the new classifier and comparing it with that of the first classifier to note any changes. This should improve the RMSE, but if the GAN isn’t trained correctly, the synthetic data could be too different from the real data and would instead mislead the classifier; lowering its accuracy.

**Dataset**

The unbalanced training set obtained from the Androzoo database contains 4465 samples, including 3,000 benign and 1,465 malign instances, each with 128 attributes. These attributes are monogram representations of system calls that lead to malicious or benign behavior. The test set consists of 950 samples, with approximately 52% designated as the public set and 48% as the private set. Since the competition did not provide the test set labels, the RMSE can only be evaluated by uploading the classification results to Kaggle.

**Proposed Method**

To add to the research that has already been conducted on this problem, we will be generating pictorial and tabular based data with GAN models that have been underused. For example the deep convolutional GAN model is excellent for creating images[6], and the wasserstein GAN with gradient penalty is a sophisticated model for generating tabular data[7]. We will also be testing different classifier models on the real and synthetic data, as according to the No Free Lunch Theorem there is no single classifier that is better than all others, instead some classifiers are better suited to the problem[8]. This is relevant because we are judging the success of our GAN model on how much the classifier improves, but if the classifier is poorly suited to the problem then the RMSE may not change or even become worse. So to get an accurate understanding of how well our GAN model’s synthetic data is improving classifier training we will train multiple classifiers and compare their results. A part of this project will involve investigating what types of classifiers have both been underused in related literature and would be well suited to this problem. Many papers used boosting algorithms as they help to eliminate high bias, and as the dataset is unbalanced we are more likely to see a bias towards the benign class. But this could provide better results on the unbalanced training set, that undersells how beneficial balancing the dataset with synthetic data could be. Therefore, it may be wise to test classifiers that are suited to problems with an expected bias, such as boosting algorithms, and those that are not well suited, like the naive Bayes algorithm.

**References**

[1] S. Pioroński, T. Górecki, “Using GAN to generate malicious samples suitable for binary classifier training,” in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 6522-6527, Dec. 2022.

[2] S. T. I. Tonmoy, S. M. Zaman, “OOG-Optuna Optimized GAN Sampling Technique for Tabular Imbalanced Malware Data,” in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 6534-6539, Dec. 2022.

[3] D. Shin, D. Han, S. Kyeong, “Performance enhancement of malware classifiers using generative adversarial networks,” in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 6528-6533, Dec. 2022.

[4] F. Wang, H. Al Hamadi, E. Damiani, “A visualized malware detection framework with CNN and conditional GAN,” in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 6540-6546, Dec. 2022.

[5] ​​H. H. Alsheraifi, H. M. Sajwani, S. M. Aljaberi, A. A. Alblooshi, A. H. Alhashmi, S. A. Sharif, E. Damiani, “Using Generative Adversarial Networks to Simulate System Calls of Malicious Android Processes,” in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 6516-6521, Dec. 2022.

[6] J. Luo, J. Huang, H. Li, “a case study of conditional deep convolutional generative adversarial networks in machine fault diagnosis,” *Journal of Intelligent Manufacturing*, vol. 32, pp. 407 - 425, May. 2020. Accessed: Sep. 27, 2024. [Online] Available: [C-DCGAN.article](https://link.springer.com/article/10.1007/s10845-020-01579-w)

[7] M. Zheng, T. Li, R. Zhu, et al. “conditional wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification,” *Information Sciences*, vol. 512, pp. 1009 - 1023, Feb. 2020. Accessed: Sep. 27, 2024. [Online] Available: [WGAN-GP.article](https://www.sciencedirect.com/science/article/abs/pii/S0020025519309715)

[8] E. Adebayo, *“a comprehensive empirical demonstration of the no free lunch theorem in statistical machine learning,”* M.S. Thesis, AIMS, Rwanda, Africa, 2019, [Online]. Available: [thesis.pdf](https://www.researchgate.net/profile/Ezekiel-Ogundepo-2/publication/339882872_A_Comprehensive_Empirical_Demonstration_of_the_No_Free_Lunch_Theorem_in_Statistical_Machine_Learning/links/5e6a46cc299bf135e60a1da4/A-Comprehensive-Empirical-Demonstration-of-the-No-Free-Lunch-Theorem-in-Statistical-Machine-Learning.pdf)