

A decorative background featuring a large, dark green monstera leaf with characteristic splits on the right side. In the center, there is a white, shallow bowl or dish. The overall aesthetic is clean and modern, with a mix of natural and geometric elements.

# STAT 574: Phishing URL Detection via an L1-Regularized Artificial Neural Network Pipeline

A Project Conducted By:  
Cory Suzuki

# Table of Contents

Introduction & EDA (1 min)

Binary Classifiers (2 mins)

Theory: Artificial Neural Network: L1

Regularization & Dropout Layers (2 mins)

Main Results (2 mins)

Concluding Remarks & Future Work (1 min)

# Introduction & Exploratory Data Analysis (EDA)



# Introduction

- Phishing URL's are a common threat to personal cybersecurity in the modern digital age
- On a weekly basis, 56% of companies have their databases and data warehouses breached through phishing scams [Keepnet Security, 2024]
- PhiUSILL Dataset Information: 45 features that are mixed with continuous, binary, and multinomial features. **12 features** survived after Spearman Correlation-based feature selection and Gradient Boosting feature importance
- PhiUSILL dataset extracted from UCI Machine Learning Repository
- Goal: Compare the accuracies between all binary classifiers introduced in the STAT 574 course and compare them with the performance of an L1-regularized Artificial Neural Network.



# Distributional EDA & Feature Overview

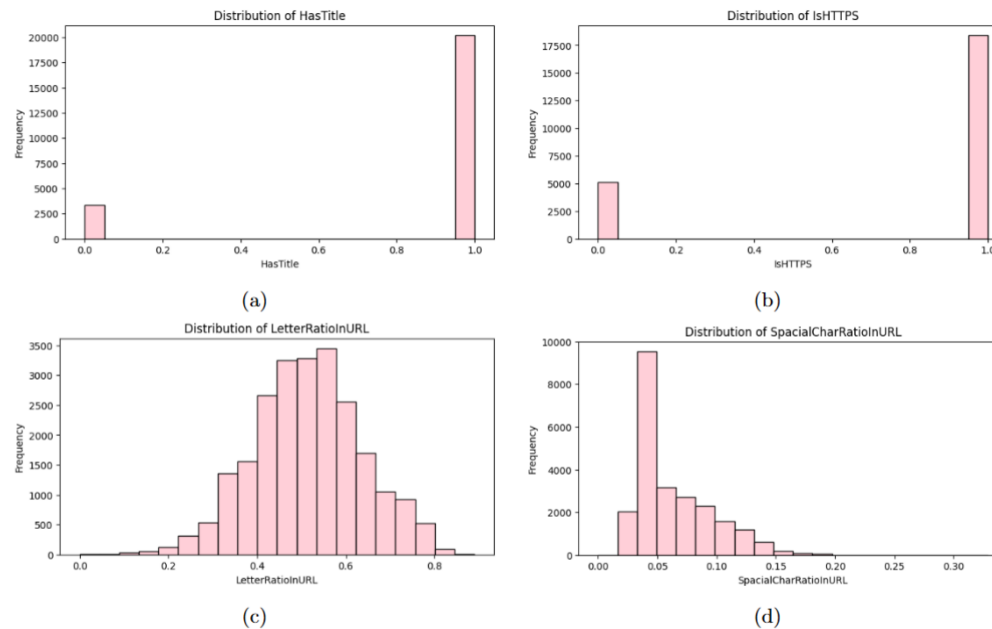


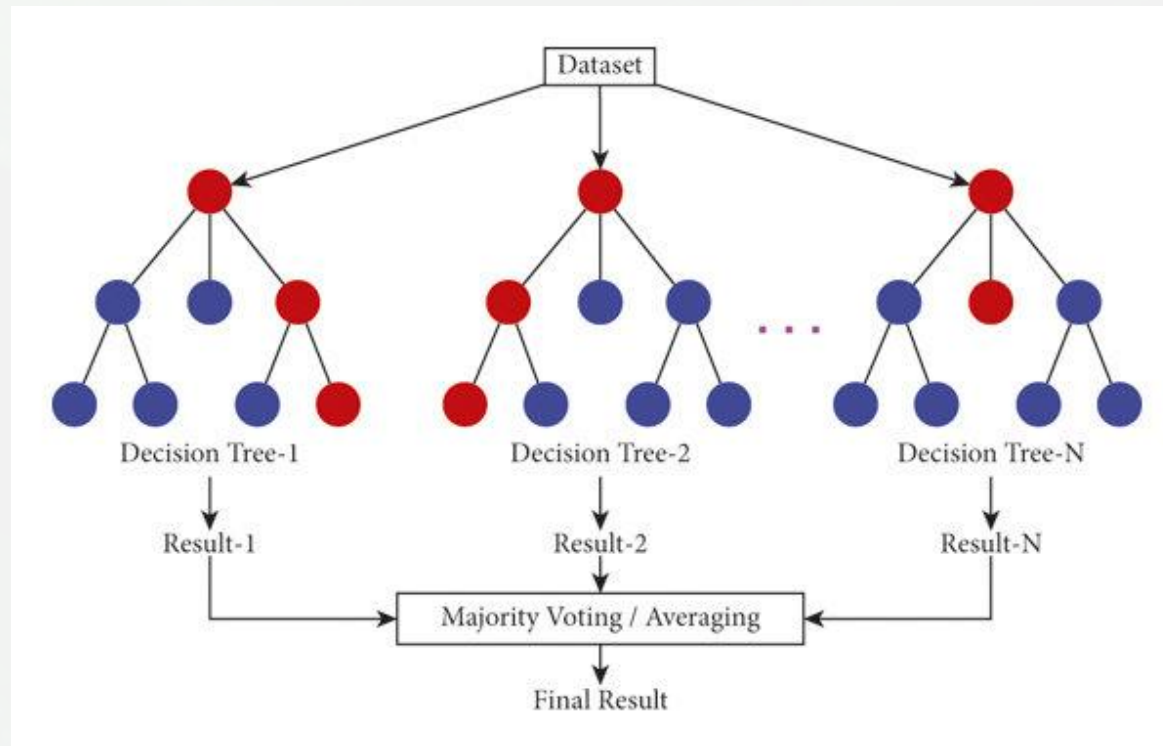
Figure 1: Distributional Histograms EDA

- Letter Ratio in URL is approximately normal
- Spatial Character Ratio in URL is right-skewed
- Binary classes tend to indicate a higher proportion of Phishing URL's

Variable	Type	Description	Responses
Label (Target)	Binary	URL is spam or not	- 1 (yes) - 0 (no)
LargestLineLength	Continuous	Largest Line length of URL	Numeric Value
URLSimilarityIndex	Continuous	How similar a URL is to another	Numeric Value index
LetterRatioInURL	Continuous	Letter Ratio in URL	Numeric Value
URLCharProb	Continuous	Probability of character in URL	Numeric Value
NoOfSubDomain	Multinomial	Number of Subdomains in URL	Finite Integer Values
IsHTTPS	Binary	Does the URL have an secure HTTP?	- 1 (yes) - 0 (no)
NoOfExternalRef	Multinomial	Number of external reference encryptions in URL	Finite Integer Values
CharContinuationRate	Continuous	Rate of Character continuations in URL	Numeric Value
HasTitle	Binary	Does the URL have a title in it?	- 1 (yes) - 0 (no)
NoOfSelfRedirect	Multinomial	Number of Redirects URL deploys to itself	Finite Integer Values
URLLength	Continuous	Length of the URL	Numeric Values
SpacialCharRatioInURL	Continuous	Number of spaces are in a URL	Numeric Values
NoOfDegitsInURL	Multinomial	Number of Digits in URL	Finite Integer Values
TLDLegitimateProb	Continuous	Probability that a Top Level Domain (TLD) or informally the suffix domain of a URL is legitimate	Numeric Value

Table 1: Table Summary of Variables

# Binary Classifiers Implementation



# Binary Classifier Performance

Binary Classifier	Accuracy	Misclassification rate
Decision Tree (Entropy Criterion)	1.0	0.0
Decision Tree (Gini Criterion)	1.0	0.0
Random Forest (Entropy Criterion)	1.0	0.0
Random Forest (Gini Criterion)	1.0	0.0
K-Nearest Neighbors (KNN)	0.9853	0.0147

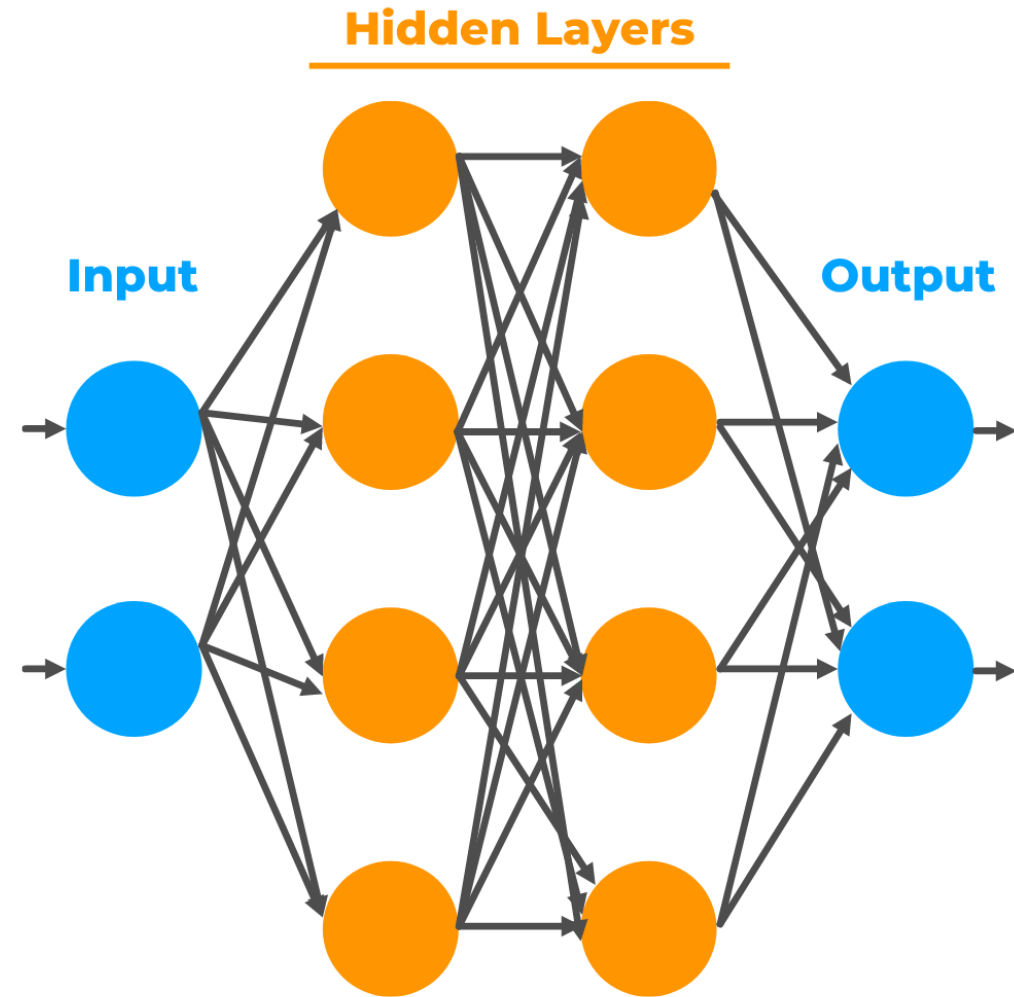


# Binary Classifier Performance (Continued)

Binary Classifier	Accuracy	Misclassification rate
Gradient Boosting	0.9997	0.0003
Naïve Bayes	0.9881	0.0119
Support Vector Machine (Linear)	1.0	0.0
Support Vector Machine (Polynomial)	0.5915	0.4085
Support Vector Machine (Radial)	0.7651	0.2349
Support Vector Machine (Sigmoid)	0.3051	0.6949

Can we do better? The answer is YES!

# Theory: Artificial Neural Networks and L1 Regularization



# ANN Architecture

- Extension/special case of the Artificial Neural Network (ANN) which implements additional parameters and layers such as L2 Regularization and Dropout
- In the proposed modified ANN, we also perform hyperparameter tuning via the Keras Random Tuner library.
- Idea: Create a skeleton ANN with min and max values of potential hyperparameters, and measure the accuracy of every set of hyperparameters for a small number of iterations.
- After tuning, use best set of hyperparameters and compute the prediction accuracy on the testing set with corresponding learning curves.

# L1 Regularization Parameter & Dropout Layer

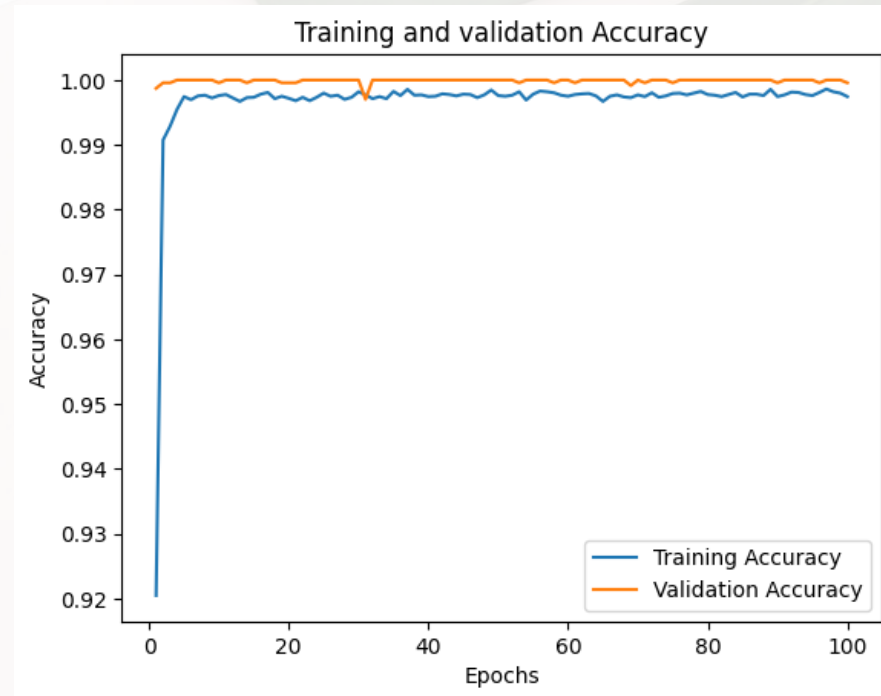
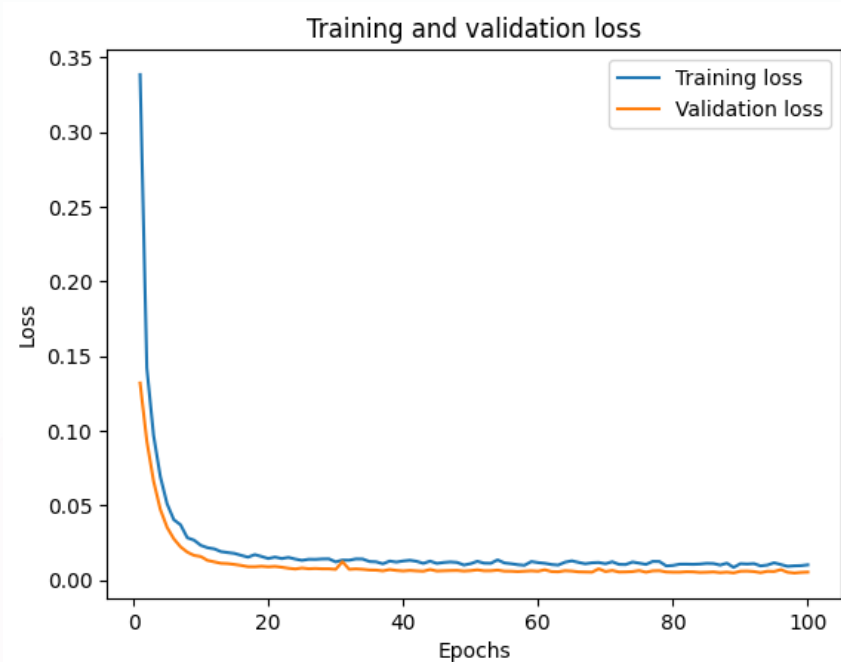
- L1 Regularization penalty parameter ensures that irrelevant features during training are shrunk down to zero to combat overfitting [Moon, 2024].
- This idea is inspired from the Least Absolute Shrinkage and Selection Algorithm (LASSO) formally introduced by Robert Tibshirani in 1996.
- Dropout layer ensures that a proportion of neurons in the ANN are dropped and calculates accuracy metrics during training to detect potential overfitting [Aurelion, 2017].
- Example: A dropout rate of 0.5 will drop half of the neurons in the current layer being trained.



# ANN Results & Analysis

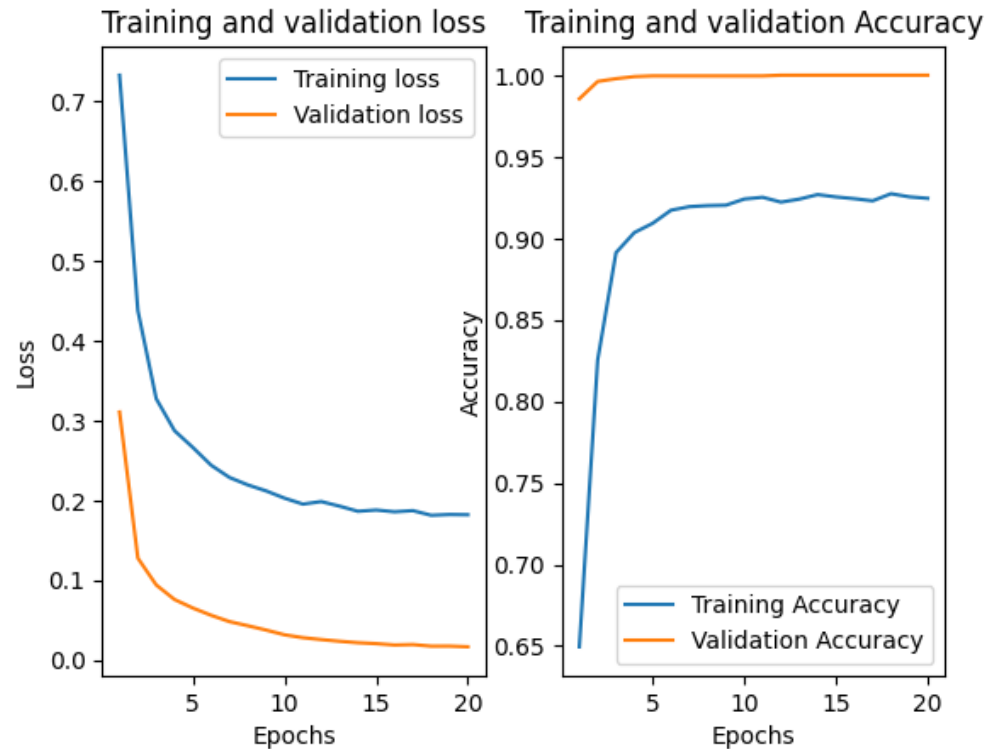


# Learning Curves for First Model (Model A)



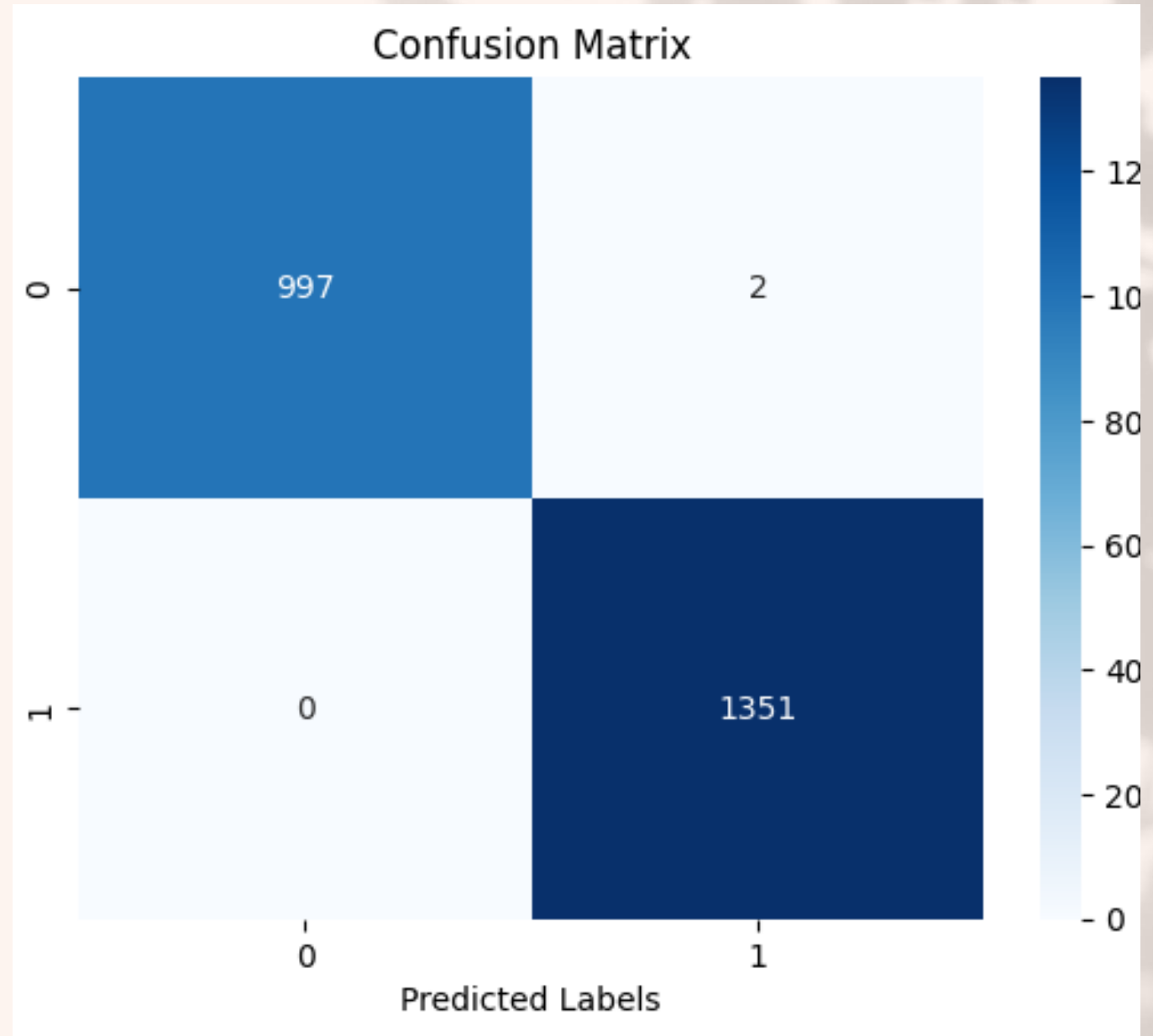


# Learning Curves for Final Model



- Shortened training period to around 20 epochs
- Tuned model shows little to no symptoms of severe overfitting
- High Accuracy plateaus during training on the validation set.
- Computed test set accuracy: 0.9999

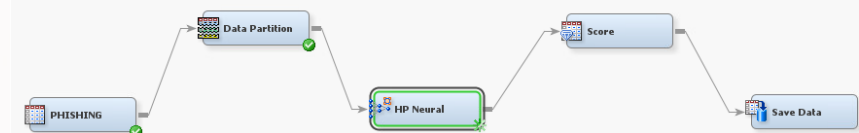
## Confusion Matrix for Best Model





# Verification of Results from R (Left) and SAS Enterprise Miner (Right)


```
[1] 0.9995
```



**The SAS System**

accuracy

0.977612



# Concluding Remarks & Future Work



[http://ww](http://www)

# Key Takeaways

## Fitted Binary Classifiers for Preliminary Analysis

- Some models showed signs of overfitting and had lower testing accuracies

## Refine delivery style

- L1 regularization and Dropout increased test accuracy and remedy overfitting in the ANN model

## Final Remark

- The ANN generalizes well over testing data and has the highest accuracy.

## Future Work:

- Implement Reinforcement Learning Algorithms
- Implement Transfer Learning Methodologies with the dataset
- Compare accuracies, can we do even better than an ANN?



# References

Aurelion, Geron. *Hands-On Machine Learning with SciKit-Learn and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Accessed April 8, 2025.

Keepnet Solutions. *2025 Phishing Statistics: Top Phishing Statistics and Trends You Must Know in 2025*. Accessed April 8, 2025.

Korosteleva, Olga. STAT 574: Data Mining Lecture Notes. *Artificial Neural Networks*. Accessed April 8, 2025.

Lee, Seungjoon. STAT 576: Data Informatics Lecture Notes. *Unsupervised and Semi-Supervised Learning*. Accessed April 20, 2025.

Prasad, A., & Chandra, S. (2023). *PhiUSIII: A Diverse Security Profile Empowered Phishing URL Detection Framework Based on Similarity Index and Incremental Learning*. *Computers & Security*, 103545. DOI: <https://doi.org/10.1016/j.cose.2023.103545>.



Thank you for  
your  
consideration!

