Detecting Phishing Websites: A Deep Learning Approach with MLP

Jonathan Garza
Department of Computer Science
California State University, Fresno
Fresno, California
jonathangarza559@mail.fresnostate.edu

Ulysses Ochoa
Department of Computer Science
California State University, Fresno
Fresno, California
ulysses_ochoa@mail.fresnostate.edu

Ahmad Kirkland
Department of Computer Science
California State University, Fresno
Fresno, California
ahmadkirkland566@mail.fresnostate.edu

ABSTRACT

Phishing is a prevalent cybersecurity threat that exploits deceptive websites to steal sensitive information. This project aims to explore the application of a Multilayer Perceptron (MLP) to detect phishing websites using a database from the UCI Machine Learning Repository. The dataset consists of over 11,000 URLs with 30 features being extracted from website characteristics.

Preprocessing involved converting the feature range from [-1, 1] to [0,1] to ensure compatibility with the model. The MLP was chosen for its suitability in handling tabular data, its efficient hyperparameter tuning for medium-sized datasets, and its ability to effectively perform binary classification. In order to optimize the model's performance, we conducted hyperparameter exploration, which included adjusting the learning rates, optimizers, and epochs. This report details the dataset analysis, model developments, and hyperparameter tuning process, resulting in a robust solution for identifying phishing websites.

1 INTRODUCTION

Phishing websites pose a significant threat to cybersecurity, obtaining sensitive information under false pretenses. Despite advancements in detection techniques, the rapid evolution and obfuscation strategies of malicious URLs outpace traditional blacklisting methods. This model aims to address this challenge by utilizing a supervised learning approach to classify URLs as phishing or legitimate based on a dataset containing 30 features extracted from website attributes. Using a Multilayer Perceptron, the project aims to achieve high detection accuracy with minimal preprocessing. This approach highlights the advantages of using tabular data and explores hyperparameter optimization to enhance the model's efficiency in identifying phishing attempts.

2 RELATED WORKS

2.1 Research Papers

Numerous studies have explored methods to detect phishing websites. In Mohammad et al. " An Assessment of Features Related to Website Using An Automated Technique",[2] various essential features were identified like the presence of IP address, URL length, and the use of URL shortening services, which are all indicators of phishing websites. Their work emphasized the challenges of inconsistent feature definitions across studies and proposed novel feature rules to improve the classification accuracy

"Detecting Malicious URLs Using Lexical Analysis" [1] by Mamun et al. demonstrates the effectiveness of lexical analysis in detecting malicious URLs They highlight that machine learning models using lexical features could surpass traditional blacklisting by identifying zero-day phishing attempts. Their research also explores obfuscation techniques such as encoding and redirection, which phishing websites often use to evade detection. The integration of lightweight lexical features enables rapid classification while maintaining high accuracy, proving effective for proactive phishing detection.

2.2 Datasets

Phishing Websites Dataset: This dataset from the UCI Machine Learning Repository consists of over 11,000 website samples described by 30 features extracted from URL characteristics and website specifications. These features include indicators such as the presence of an IP address, URL length, amongst other specific characteristics in the URL. The dataset is preprocessed with integer values in the range [-1,1], where -1 denotes a negative feature, 0 represents an absent or neutral features, and 1 indicates a positive feature. Modifications were made to replace -1 values with 0 for compatibility with the chosen model. This dataset provided a structured foundation for developing the machine learning model to detect phishing websites efficiently.

3 METHODS

In this project, we used a Multilayer Perceptron (MLP) to classify phishing websites from safe ones. The choice of MLP was based on the nature of the dataset, which consists of 30 numerical features derived from website attributes (such as URL length, IP address presence, and the use of URL shortening services). MLPs are ideal for tabular datasets like ours, as they can model complex, non-linear relationships between features. The model architecture included an input layer with 30 neurons (one for each feature), followed by three hidden layers with 64, 32, and 16 neurons, respectively. The output layer consisted of a single neuron with a sigmoid activation function to output a probability score between 0 and 1, indicating whether a website was phishing (0) or safe (1). To introduce non-linearity and improve learning, ReLU activation functions were used in the hidden layers. The model was trained using the Adam optimizer with a learning rate of 0.001, and the binary cross-entropy loss function, which is appropriate for binary classification tasks. This architecture allowed for flexibility in tuning various hyperparameters such as the number of hidden layers, neurons, and dropout rate, ensuring we could optimize the model's performance through experimentation.

4 HYPERPARAMETER EXPLORATION

During the development and training of the phishing detection model, various hyperparameters were explored to optimize

performance. One notable observation was the behavior of model accuracy over increasing epochs and a change in the learning rate:

Epochs: When the number of training epochs was increased to the range of 25-30, the model's test accuracy plateaued at 95-96%. This plateau suggests that the model has reached its optimal performance under the current architecture and hyperparameters. Training beyond this point yielded no further improvement in accuracy.

Learning Rate Adjustment: One significant improvement was achieved by increasing the learning rate from 0.001 to 0.01, allowing the model to converge more effectively during training. This resulted in better generalization, with accuracy consistently reaching the upper 96% range across different dataset splits. The increase did not introduce instability or divergence during training, validating the appropriateness of the higher learning rate.

Learning Rate Scheduler: To improve optimization during training, a StepLR learning rate scheduler was implemented. This scheduler dynamically reduced the learning rate by half every 5 epochs (step_size=5, gamma=0.5), allowing the optimizer to take larger steps early in training and smaller, more precise updates later. The scheduler enhanced the model's convergence, helping it achieve consistent performance across multiple data splits while avoiding issues like overshooting or stagnation during optimization.

5 RESULTS

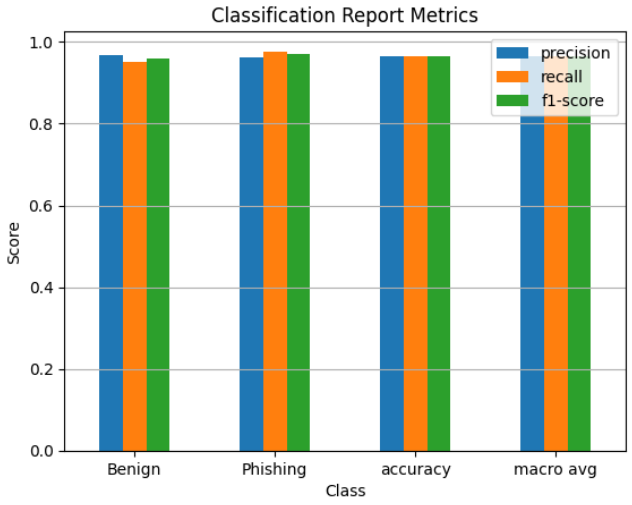

Figure 1: Comparison of classification metrics (precision, recall, and F1-score) for detecting benign and phishing websites, along with overall accuracy and macro-average scores.

The phishing detection model demonstrated exceptional performance in classifying phishing and benign websites. The evaluation metrics indicate strong and balanced results across both classes, highlighting the model's effectiveness in identifying threats while minimizing false positives.

Precision: The model achieved a precision of 96% for both benign and phishing websites. This indicates a low false-positive rate, ensuring benign websites are rarely misclassified as phishing.

Recall: A recall of 97% for both classes demonstrates the model's ability to detect nearly all phishing websites, with minimal false negatives.

F1-Score: The F1-score, which balances precision and recall, was consistently high at 97%, reflecting robust overall classification performance.

Accuracy: The overall accuracy of the model was between 95% and 97%, correctly predicting the class of websites in the vast majority of cases.

Macro Average: Precision, recall, and F1-score remained balanced across classes, confirming that the model performs equally well for both benign and phishing classifications.

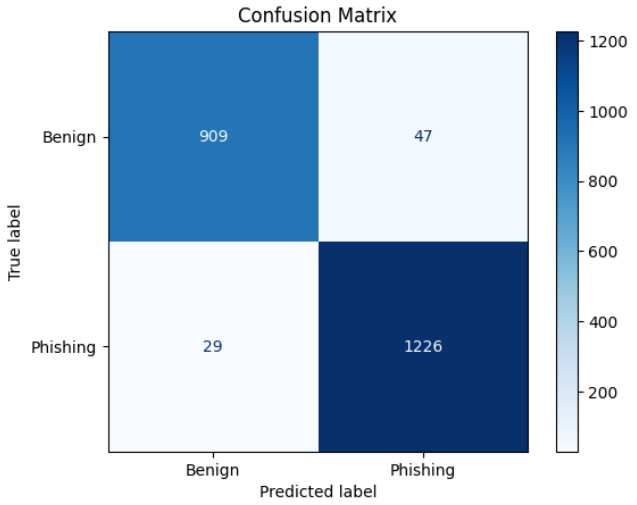

Figure 2: Confusion matrix illustrating the performance of the phishing website detection model, showing the counts of true positives, false positives, true negatives, and false negatives for benign and phishing labels.

The confusion matrix demonstrates the effectiveness of the phishing detection model. It was able to correctly identify 909 benign websites (true negatives) and 1,226 phishing websites (true positives), indicating its strong ability in distinguishing the two classes. However, the model misclassified 47 benign websites as phishing (false positives) and 29 phishing websites as benign(false negatives). These results highlight the model's high accuracy, as the majority of predictions align with the true labels. The relatively low number of false negatives is especially important in the context of phishing detection, as failing to identify a phishing website could expose users to significant security risks. Meanwhile, the slight number of false positives, while less critical, may still warrant further optimization to improve user experience and reduce unnecessary alerts. Overall, the matrix underscores the robustness and reliability of the model in detecting phishing attempts.

These results suggest that the model is well-suited for real-world deployment in phishing detection scenarios, where both high precision (minimizing false alarms) and high recall (capturing threats) are critical.

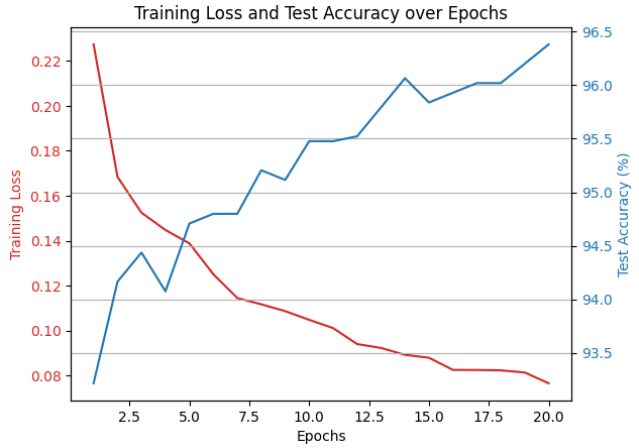

Figure 3: Training Loss and Test Accuracy Over Epochs: The graph illustrates the steady decrease in training loss alongside the consistent increase in test accuracy, highlighting the model's effective convergence and improving performance during training.

6 CONCLUSION

In this project, we developed a Multilayer Perceptron model to effectively detect phishing websites using a dataset containing 30 features derived from website attributes. The MLP architecture was well-suited for the task, as it handled the tabular data and captured complex, non-linear relationships between features. Through a series of hyperparameter exploration, including adjustments to the number of epochs and the learning rate, we optimize the model's performance. The learning rate adjustment to 0.01 and the introduction of a StepLR learning rate scheduler significantly improved the convergence and generalization, contributing to a high level of accuracy. The results of the model were outstanding, with precision and recall both reaching 96% and 97%, respectively for both benign and malignant websites. This demonstrates the model's ability to effectively identify phishing websites while minimizing false positives and negatives. The consistent F1-score of 97% across all classes further validated the robustness of the model. The overall accuracy, ranging from 95% to 97%, indicates reliable performance in real-world phishing detection scenarios. These findings suggest that the MLP-based approach is a strong candidate for practical deployment in phishing website detection systems, where balancing precision and recall is crucial to minimizing security risks. Future work may explore optimization techniques or the use of other deep learning models to enhance performance even further. However based on the results presented, the model shows significant promise in safeguarding users against phishing threats.

ACKNOWLEDGMENTS

We wish to acknowledge the comprehensive and insightful content of CSCI 167: Deep Learning, which greatly enhanced our understanding and application of the subject matter. Additionally, we thank the Department of Computer Science at California State University, Fresno, for fostering an environment conducive to our academic and practical pursuits in this field.

REFERENCES

[1] R. M. Mohammad, F. A. Thabtah, and L. McCluskey. 2012. An assessment of features related to phishing websites using an automated technique. In *2012 International Conference for Internet Technology and Secured Transactions*. 492–497.

[2] Arash Habibi Lashkari Natalia Stakhanova Mohammad Saiful Islam Mamun, Mohammad Ahmad Rathore and Ali A. Ghorbani. 2016. Detecting Malicious URLs Using Lexical Analysis. *Network and System Security* (2016), 467–482. Springer International Publishing.