Literature Survey - Web Phishing Detection

Team Members:

DEVALLA CHARAN SRI SAI (111519104022)
GALLA VISHNU SAI SAKETH (111519104029)
DASARARAJU GNANENDRA (111519104020)
CHEJARLA VINAY KARTHIK (111519104017)

S.No	Title of The Paper	Methodology Used	Observations	Results + Conclusion	Limitations
1	An Optimized Stacking Ensemble Model for Phishing Websites Detection (2021)	The optimisation was carried out using a genetic algorithm (GA) to tune the parameters of several ensemble machine learning methods, including random forests, AdaBoost, XGBoost, Bagging, GradientBoost, and LightGBM	Higher accuracies than currently proposed models	The detection accuracy reached 97.16%, 98.58%, and 97.39% for Dataset 1, Dataset 2, and Dataset 3	Does not take into account the weightage of features
2	Phishing Detection using Machine Learning based URL Analysis: A Survey (2021)	Conducted a literature survey of all the top publications, along with a summary of the different features that are extracted	The top features used are Address bar features, abnormal based features, HTML based features and Domain based features	97.36% is the highest reported accuracy on the UCI dataset, where Random Forest is the most robust	Doesn't go into detail with neural network approaches
3	Detection of phishing websites by using machine learning-based URL analysis (2020)	Utilizing 48 characteristics across 3 distinct datasets, 8 different methods are used.	On each of the three datasets, Random Forest has the highest accuracy. Also desired is Artificial Neural Network.	First Dataset - 94.59% accuracy Second Dataset - 90.5% accuracy Third Dataset - 91% accuracy	The accuracies are very less for the other two datasets

S.No	Title of The Paper	Methodology Used	Observations	Results + Conclusion	Limitations
4	Phishing websites detection using a novel multipurpose dataset and web technologies features (2022)	Uses a public dataset - PILWD-134K and classification using LightGBM	54 Features can be extracted	97.95% accuracy observed from the LightBGM Classifier	Doesn’t use brand and logo recognition methods for more efficient classification
5	Phishing attacks detection using machine learning approach (2020)	Uses PCA-style feature selection techniques before applying the RF and DT classifiers	Since RF has less variance, it might manage the overfitting issue.	Random Forest delivers an accuracy of 97%	Doesn’t explore a wide variety of features that aren’t restricted to the domain