**Review of Algorithms, Techniques Used in Spam Classification**

In the year 2018, Yuan [1] proposed a blend of features pertaining to URL and web page for the detection of phishing websites. Along with the basic features like the length of the URL, unusual characters, etc, statistical features like mean, median and variance and lexical features like title and the content of the web page were also considered. Several algorithms like KNN, Logistic Regression, Random Forest, Deep Forest, XGBoost were applied. It was found that Deep Forest followed by XGBoost manifested high accuracy and less training time.

Koray [2] detects the phishing websites by analyzing and extracting features from the URL. The Random Word Detection module is used which decomposes URL into small features which are then used to classify if the websites are legitimate or not. Seven di\_erent machine learning algorithms like Naive Bayes, KNN, SVM, Random Forest, Decision Tree, Adaboost, K-star were implemented on a humongous amount of data. It was seen that Random Forest produced the highest accuracy of around 97.98 percent among all the techniques applied.

In the world of online advertising, fraud clicks are one of the most momentous issues. The research done by (Minastireanu and Mesnita [3] tackles the problem of fraud clicks by using the latest machine learning technique viz. LightGBM on the dataset which contains millions of clicks. The K-Fold Cross-Validation technique is used as a feature engineering which helps in improving the performance. The accuracy achieved by the model was 98% and was found to be the fastest with respect to computational speed and low on memory consumption. Looking at the above research papers, it was found that LightGBM and XGBoost are suitable as it performs faster with a less computational speed, unlike the deep learning techniques. Also, Random Forest performed well giving high accuracy and hence in this research, these algorithms are chosen.

**Analysis Table :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author | Area of  Classification | Applied Classifiers | Features Extracted | Results |
| (Yuan et al.  (2018)) | Phishing  Classification | Logistic Regression,  Decision  Tree, GBDT,  Deep Forest,  Random Forest,  XGBoost, KNN | TF-IDF for feature  extracted  from URL and  links | Deep Forest  with 97.7%  and XGBoost  with 97.1% |
| (Koray et al.  (2019)) | Phishing  Classification | Naive Bayes,  KNN, SVM,  Random Forest,  Decision Tree,  Adaboost, K-star | NLP features,  word vectors | Random  Forest with  97.98% |
| (Minastireanu  and Mesnita  (2019)) | Fraud Advertisement  Classification | LightGBM, XGBoost.  Stochastic  Gradient Boosting | IP, OS, channel,  device, click time | LightGBM  with 98%  accuracy |

**References :**

[1] Yuan, H., Chen, X., Li, Y., Yang, Z. and Gv Liu, W. (2018). Detecting Phishing

Websites and Targets Based on URLs and Webpage Links, Proceedings - International

Conference on Pattern Recognition 2018-August: 3669{3674.

[2] Koray, O., Buber, E., Demir, O. and Diri, B. (2019). Machine learning based phishing

detection from URLs, Expert Systems With Applications 117: 345{357.

URL: https://doi.org/10.1016/j.eswa.2018.09.029

[3] Minastireanu, E.-A. and Mesnita, G. (2019). Light GBM Machine Learning Algorithm to

Online Click Fraud Detection Light GBM Machine Learning Algorithm to Online Click

Fraud Detection, Journal of Information Assurance & Cybersecurity 3(April): 1{15.