# CONCLUSION

Phishing detection mechanism aims to improve current blacklist methods, protecting users from malicious login forms. Our work provides an updated dataset PILU-90K for researchers to train and test their approaches. This dataset includes legitimate login URLs which are the most representative scenario for real-world phishing detection.

We explored several URL-based detection models using deep learning and machine learning solutions trained with phishing and legitimate home URLs. The main advantage of our approach is the low false-positive rate when classifying this type of URL. Among the different evaluated models, TFIDF combined with N-gram and LR algorithm obtained the best results with a 96:50% accuracy. In comparison with the current state-of-the-art, reviewed in Section II, our approach present three main advantages:

**No dependence on external services**. A limitation of the description methods that use features such as WHOIS domain age, page ranking on Google or Alexa or online blacklists, is their dependence on those services. Network slowdowns and service shortages can negatively impact analysis time, making real-time execution infeasible. Since phishing websites have a short lifespan [12], low detection times are required to warn users before accessing phishing websites.

**Login website detection**. Unlike other methods, which are trained with homepage URLs as representatives of the legitimate class, our model was trained with legitimate login websites. This ensures the correct classification of those websites. Therefore, our approach can be applied to the real case scenario where users have to predict whether a login form page is legitimate or phishing.

**Updated and real-world dataset**. PLU-60K is focused on using updated legitimate login URLs. As demonstrated, models trained with old datasets were not able to endure their performance over time. We provide an updated phishing URL dataset for models to learn from nowadays phishing URLs and trends, which are crucial for real-world performance.

We demonstrated that phishing URL detection systems trained with legitimate land page URLs fail to classify legitimate login URLs correctly. The best-tested models could only classify 69:50% of these URLs correctly, which implies a high false-positive rate. For this reason, we recommend that a phishing detector, which intends to be used in a real situation, should be trained using *legitimate login* websites (such as PLU-60K) instead of homepages. The main drawback of using login websites for training is that, due to the similarity between phishing and legitimate samples, overall accuracy is slightly reduced. The tradeoff against the state-of-the-art methods is still fair due to their high false-positive rate.

Different categories for current phishing attacks were identified by using a domain frequency analysis. While standalone and compromised domains were the most common approaches, free hosting services, cloud web servers and malware blog posts represent many current phishing attacks due to their cost and effectiveness for phishing campaigns.

Finally, we demonstrated that machine learning models using handcrafted URL features decreased their performance over time, up to 10:42% accuracy in the case of the Light GBM algorithm from the year 2016 to 2020. For this reason, machine learning methods should be trained with recent URLs to prevent substantial ageing from the date of its release. In the future, we will add more information about the samples into the analysis, such as the source code of the website and a screenshot of its content, which could be useful to increase the phishing detection performance. In addition, we will enlarge our dataset, including such information. Finally, observing that deep learning techniques and automatic feature extraction obtained promising results over traditional feature extraction, we intend to explore different URL codifications to improve detection performance.