الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي - 08 ماي 1945 - بسيدي بلعباس
Higher School of Computer Science
-8 May 1945- Sidi Bel Abbes

Thesis

To obtain the diploma of **Engineering Degree**
Field of Study: **Computer Science**
Specialization: **ISI**

Theme

A Metaheuristic-Optimized Framework for Robust Malicious URL Detection Using ML and DL with Real-Time Application

---

Presented by

BENDOUBBA Hana

ALILAT Zohra

Defending on **09 July 2025**
In front of the jury composed of

Dr. BENDAOUD Fayssal	President
Dr. KHALDI Miloud	Supervisor
Dr. CHIKH Asma	Examinator

Academic Year: 2024 - 2025

Abstract

The rapid expansion of the Internet has made it an integral part of modern life, facilitating communication and data exchange across numerous domains. However, this growth has also led to a surge in cyber threats, particularly those leveraging malicious URLs to deceive users and compromise sensitive data. Traditional detection methods often rely on manual feature extraction, which limits scalability and effectiveness in dynamic threat environments. This study proposes an enhanced approach to malicious URL detection by applying hyperparameter optimization (HPO) to machine learning and deep learning models, using metaheuristic algorithms such as Optuna and Randomized Search. Two benchmark datasets were used for evaluation: the Kaggle "Malicious URLs" dataset, based on raw URL strings, and the ISCX-URL-2016 dataset, which includes engineered features. The results show that optimized models significantly outperform baseline configurations. For instance, the BERT-Optuna model achieved an accuracy of 98.84% and an F1-score of 99.02%, while LightGBM with TF-IDF and Random Search reached 98.46% accuracy. On the ISCX dataset, the CatBoost-RF-Optuna model achieved 98.31% accuracy with a low False Acceptance Rate (FAR) of 0.0069 and an AUC-ROC of 99.93%. By integrating advanced feature representations and automated hyperparameter tuning, the proposed models not only enhance classification performance but also maintain computational efficiency. Additionally, a practical browser extension was developed to enable real-time malicious URL detection, bridging the gap between research and application.

Keywords: Malicious URL Detection, Hyperparameter Optimization, Metaheuristics, Machine Learning, Deep Learning, BERT, Optuna, LightGBM, Random Search, CatBoost.

الملخص

أدى التوسع السريع للإنترنت إلى جعله جزءً لا يتجزأ من الحياة الحديثة، حيث يسهل التواصل وتبادل البيانات في مختلف المجالات. ومع ذلك، فقد أدى هذا النمو أيضاً إلى زيادة التهديدات السيبرانية، خاصة تلك التي تستغل الروابط الخبيثة URLs لخداع المستخدمين وسرقة المعلومات الحساسة. وتعتمد طرق الكشف التقليدية غالباً على الاستخراج اليدوي للخصائص، مما يحد من قابليتها للتوسع وفعاليتها في بيئات التهديد الديناميكية. تقترح هذه الدراسة نهجاً محسناً للكشف عن روابط URL الخبيثة من خلال استخدام تقنيات تحسين المعاملات الفائقة (HPO) في نماذج التعلم الآلي والتعلم العميق، باستخدام خوارزميات ميتاهورسيتية مثل Optuna و Randomized Search. تم تقييم المنهجية باستخدام مجموعتي بيانات مرجعية: مجموعة بيانات Kaggle المعتمدة على سلاسل URL الخام، ومجموعة بيانات ISCX-URL-2016 التي تتضمن خصائص هندسية. أظهرت النتائج أن النماذج المحسنة تفوقت بشكل كبير على النماذج الأساسية. على سبيل المثال، حقق نموذج BERT-Optuna دقة قدرها98.84% ودرجة F1 بنسبة99.02% ، بينما وصل نموذج LightGBM مع TF-IDF و Random Search إلى دقة قدرها98.46% . أما على مجموعة بيانات ISCX، فقد حقق نموذج CatBoost-RF-Optuna دقة بنسبة98.31% مع معدل قبول خاطئ (FAR) منخفض يبلغ 0.0069 وقيمة AUC-ROCبلغت 99.93% . من خلال دمج تمثيلات متقدمة للخصائص وتقنيات آلية لضبط المعاملات، تمكنت النماذج المقترحة من تعزيز أداء التصنيف مع الحفاظ على الكفاءة الحاسوبية. كما تم تطوير إضافة (Extension)لمتصفح الويب لتفعيل الكشف الفوري عن الروابط الخبيثة، مما يربط بين الجانب النظري والتطبيق العملي.

الكلمات المفتاحية: اكتشاف الروابط الخبيثة ، تحسين المعاملات الفائقة ، الخوارزميات الميتاهورسيتية ، التعلم الآلي، التعلم العميق ، BERT، Optuna، LightGBM، CatBoost، Random Search.

Résumé

L'expansion rapide d'Internet en a fait un élément essentiel de la vie moderne, facilitant la communication et l'échange de données dans de nombreux domaines. Cependant, cette croissance s'est accompagnée d'une augmentation des cybermenaces, notamment celles qui exploitent les URL malveillantes pour tromper les utilisateurs et compromettre des informations sensibles. Les méthodes de détection traditionnelles reposent souvent sur une extraction manuelle des caractéristiques, ce qui limite leur évolutivité et leur efficacité face à des environnements de menaces dynamiques. Cette étude propose une approche améliorée pour la détection des URL malveillantes en appliquant l'optimisation des hyperparamètres (HPO) à des modèles d'apprentissage automatique et profond, à l'aide d'algorithmes métaheuristiques tels que Optuna et Randomized Search. Deux jeux de données de référence ont été utilisés pour l'évaluation : le dataset Kaggle « Malicious URLs », basé sur des chaînes d'URL brutes, et le dataset ISCX-URL-2016, contenant des caractéristiques extraites. Les résultats montrent que les modèles optimisés surpassent significativement les configurations de base. Par exemple, le modèle BERT-Optuna a atteint une précision de 98,84% et un score F1 de 99,02%, tandis que LightGBM avec TF-IDF et Random Search a atteint une précision de 98,46%. Sur le jeu de données ISCX, le modèle CatBoost-RF-Optuna a obtenu une précision de 98,31%, un taux de fausses acceptations (FAR) réduit à 0,0069 et un AUC-ROC de 99,93%. En combinant des représentations avancées des caractéristiques avec un réglage automatisé des hyperparamètres, les modèles proposés améliorent non seulement les performances de classification, mais conservent également une efficacité computationnelle. De plus, une extension de navigateur pratique a été développée pour permettre une détection en temps réel des URL malveillantes, comblant ainsi le fossé entre la recherche et l'application réelle.

Mots-clés : Détection d'URL malveillantes, optimisation des hyperparamètres, métaheuristiques, apprentissage automatique, apprentissage profond, BERT, Optuna, LightGBM, Random Search, CatBoost.