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 **Master's Degree**
Field of Study: **Computer Science**
Specialization: **ISI**

Theme

A Comparative Study of Malicious URL Detection Using Machine Learning, Deep Learning and Metaheuristic Optimization

---

Presented by

ALILAT Zohra

BENDOUBBA Hana

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 rise of malicious websites has made cybersecurity a critical concern in the age of growing internet usage. These threats often mimic legitimate platforms to steal sensitive information, resulting in severe financial and data losses. Traditional detection techniques such as manual feature extraction and rule-based systems are increasingly inadequate against the evolving sophistication of cyberattacks. This thesis presents a theoretical analysis of machine learning and deep learning approaches for malicious URL detection as discussed in the literature. It evaluates the strengths and limitations of ensemble methods, NLP-based models, and feature-driven techniques. Particular emphasis is placed on the role of hyperparameter optimization and feature selection in improving model performance. The study also identifies major challenges in detection systems, including scalability, real time analysis, and adaptability across diverse attack vectors. Ultimately, this research highlights the importance of developing flexible, efficient, and interpretable methods for future malicious URL detection systems. It aims to provide a solid theoretical foundation for advancing cybersecurity research and enhancing defenses against evolving online threats.

Keywords: Malicious URL Detection, Hyperparameter Optimization, Metaheuristics, Machine Learning, Deep Learning.

الملخص

أدى انتشار المواقع الإلكترونية الخبيثة إلى جعل الأمن السيبراني قضية حاسمة في ظل الاستخدام المتزايد للإنترنت. تحاكي هذه التهديدات في كثير من الأحيان منصات شرعية لسرقة معلومات حساسة، مما يؤدي إلى خسائر كبيرة مالية وبياناتية. أصبحت الطرق التقليدية للكشف، مثل استخراج الميزات اليدوي والأنظمة المعتمدة على القواعد، غير كافية أمام تطور تعقيد الهجمات السيبرانية. يقدم هذا البحث تحليلاً نظرياً لاستراتيجيات التعلم الآلي والتعلم العميق المستخدمة في كشف عناوين URL الضارة كما وردت في الأدبيات. يتم تقييم نقاط القوة والضعف في أساليب التعلم التجميعي، ونماذج معالجة اللغة الطبيعية (NLP)، والطرق المستندة إلى الميزات. يتم التركيز بشكل خاص على أهمية تحسين الضبط الفائق (Hyperparameter Optimization) واختيار الميزات لتحقيق أداء نموذجي فعال. كما تحدد الدراسة أبرز التحديات التي تواجه أنظمة الكشف، مثل قابلية التوسع، المعالجة في الوقت الحقيقي، والقدرة على التعميم على أنواع متعددة من الهجمات. في الختام، تؤكد هذه الدراسة على ضرورة تطوير أساليب مرنة وفعالة وقابلة للفهم للكشف المستقبلي عن روابط المواقع الخبيثة. وتمثل هذه الدراسة أساساً نظرياً متيناً لتطوير أبحاث الأمن السيبراني وتعزيز الدفاعات ضد التهديدات الإلكترونية المتطورة.

الكلمات المفتاحية: اكتشاف الروابط الضارة، التحسين عالي الأداء، الميتاهيرستيك، التعلم الآلي، التعلم العميق.

Résumé

La prolifération des sites web malveillants a fait de la cybersécurité un enjeu majeur à l'ère de l'utilisation croissante d'Internet. Ces menaces imitent souvent des plateformes légitimes pour dérober des informations sensibles, entraînant d'importantes pertes financières et de données. Les techniques de détection traditionnelles, telles que l'extraction manuelle de caractéristiques et les systèmes basés sur des règles, deviennent de moins en moins efficaces face à la sophistication croissante des cyberattaques. Ce mémoire présente une analyse théorique des approches d'apprentissage automatique et d'apprentissage profond appliquées à la détection des URL malveillantes dans la littérature. Il évalue les forces et les limites des méthodes par ensemble, des modèles basés sur le traitement du langage naturel (NLP) et des approches fondées sur les caractéristiques. Une attention particulière est portée à l'optimisation des hyperparamètres et à la sélection des caractéristiques, essentielles à la performance des modèles. L'étude identifie également les principaux défis liés aux systèmes de détection, notamment la montée en charge, le traitement en temps réel et la capacité de généralisation face à des vecteurs d'attaque variés. En définitive, cette recherche souligne l'importance de développer des méthodes flexibles, efficaces et interprétables pour les systèmes futurs de détection d'URL malveillantes. Elle constitue une base théorique solide pour faire progresser la recherche en cybersécurité et renforcer la protection face aux menaces en ligne émergentes.

Mots-Clés : Détection d'URL malveillantes, Optimisation des hyperparamètres, Métaheuristiques, Apprentissage automatique, Apprentissage profond.