

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



Engineering Thesis

To obtain the diploma of **Engineer Degree**

Field of Study: **Computer Science**

Specialization: **Artificial Intelligence and data science (IASD)**

Theme

**Performance Evaluation of Network Traffic Analysis Platforms
and AI-Based Intrusion Detection Models Using Conditional
Variational Autoencoder (CVAE) for Data Augmentation and
real-time Deployment**

Presented by : **Abdelillah Serghine**

Defended on: **July 1, 2026**

In front of the jury composed of

Dr. Khaldi Miloud	ESI-SBA, Algeria	President of the Jury
Dr. Lahlou Laaziz	ÉTS, Canada	Thesis Supervisor
Pr. Benslimane Sidi Mohammed	ESI-SBA, Algeria	Co-Supervisor
Pr. Kara Nadjia	ÉTS, Canada	Co-Supervisor
Dr. Souyah Amina	ESI-SBA, Algeria	Examiner

Academic Year: 2025/2026

Abstract

Intrusion Detection Systems (IDS) play a critical role in protecting modern computer networks from increasingly sophisticated cyberattacks. With the rapid growth of cloud-native infrastructures, traditional rule-based detection methods are no longer sufficient. This thesis proposes a comprehensive data-driven framework for network intrusion detection that integrates network traffic analysis, data preprocessing, data augmentation, and artificial intelligence techniques.

The proposed work begins with a comparative study of three network traffic analysis platforms. These tools are evaluated based on their ability to process Kubernetes PCAP data and extract meaningful flow-based features. In addition, two benchmark datasets are used to validate the proposed approach.

To address the class imbalance problem commonly found in intrusion detection datasets, a Conditional Variational Autoencoder (CVAE) is employed to generate synthetic samples for underrepresented attack classes. This approach was selected because it can model complex data distributions and generate class-specific synthetic flows conditioned on attack labels. Compared with conventional oversampling methods, such as random oversampling or SMOTE. The balanced datasets are then used to train and evaluate multiple machine learning (ML) models as well as a Transformer-based model.

Experimental results demonstrate that CVAE-based data augmentation substantially improves classification performance for models sensitive to class imbalance, particularly DistilBERT, the F1-score increased From 75% to 99% on UNSW-NB15 and from 29% to 99% on NFStream processed dataset, while providing only marginal gains for already robust tree-based models such as XGBoost and Random Forest in terms of accuracy, precision, recall, and F1-score. Finally, the best-performing model is deployed in FlowVertex a real-time network analysis platform, demonstrating the practical applicability of the proposed framework in operational network monitoring environment, where network flows must be analyzed and classified with low latency.

Overall, this thesis highlights the importance of combining network traffic analysis tools, generative models, and advanced machine learning techniques to build robust and adaptive intrusion detection systems for modern cybersecurity environments.

Keywords : Intrusion Detection System (IDS), Network Traffic Analysis, Machine Learning, Deep Learning, Generative Artificial Intelligence, Cybersecurity, Data Augmentation.

Résumé

Les systèmes de détection d'intrusion (IDS) jouent un rôle essentiel dans la protection des réseaux informatiques modernes contre des cyberattaques de plus en plus sophistiquées. Avec la croissance rapide des infrastructures cloud-native, les méthodes traditionnelles basées sur des règles ne sont plus suffisantes. Cette mémoire propose une approche globale basée sur les données pour la détection d'intrusion réseau, intégrant l'analyse du trafic réseau, le prétraitement des données, l'augmentation des données et des techniques d'intelligence artificielle.

Le travail proposé commence par une étude comparative de trois plateformes d'analyse de trafic réseau. Ces outils sont évalués selon leur capacité à traiter des fichiers PCAP Kubernetes et à extraire des caractéristiques pertinentes basées sur les flux réseau. En outre, deux jeux de données de référence sont utilisés afin de valider l'approche proposée.

Pour résoudre le problème de déséquilibre des classes présent dans les jeux de données IDS, un autoencodeur variationnel conditionnel (CVAE) est utilisé pour générer des échantillons synthétiques pour les classes d'attaques minoritaires. Cette approche a été choisie car elle permet de modéliser des distributions de données complexes et de générer des flux synthétiques spécifiques à chaque classe, conditionnés par les étiquettes d'attaque. Elle se distingue des méthodes de suréchantillonnage classiques, telles que le suréchantillonnage aléatoire ou SMOTE. Les ensembles de données équilibrés sont ensuite utilisés pour entraîner et évaluer plusieurs modèles d'apprentissage automatique (ML), ainsi qu'un modèle basé sur les Transformers.

Les résultats expérimentaux démontrent que l'augmentation de données basée sur un CVAE améliore sensiblement les performances de classification des modèles sensibles au déséquilibre des classes, notamment DistilBERT, le score F1 a progressé de 75% à 99% sur l'ensemble de données UNSW-NB15 et de 29% à 99% sur l'ensemble de données traité avec NFStream, tandis qu'elle n'apporte que des gains marginaux aux modèles arborescents déjà robustes tels que XGBoost et Random Forest en termes d'exactitude, de précision, de rappel et de score F1. Enfin, le modèle le plus performant est déployé sur FlowVertex, une plateforme d'analyse de réseau en temps réel, démontrant ainsi l'applicabilité pratique du cadre proposé dans un environnement de surveillance opérationnelle des réseaux, où les flux réseau doivent être analysés et classifiés avec une faible latence.

Dans l'ensemble, cette mémoire souligne l'importance de combiner les outils d'analyse de trafic réseau, les modèles génératifs et les techniques avancées d'apprentissage automatique pour construire des systèmes IDS robustes et adaptatifs pour les environnements de cybersécurité modernes.

Mots-clés : Système de Détection d'Intrusion (IDS), Analyse du Trafic Réseau, Apprentissage Automatique, Apprentissage Profond, Intelligence Artificielle Générative, Cyber-sécurité, Augmentation de Données.

الملخص

تلعب أنظمة كشف التسلل (IDS) دوراً محورياً في حماية شبكات الحاسوب الحديثة من الهجمات الإلكترونية المتطورة باستمرار. ومع النمو المتسارع للبنى التحتية السحابية، لم تعد الأساليب التقليدية القائمة على القواعد كافية. تقترح هذه الأطروحة منهجاً شاملاً قائماً على البيانات لكشف التسلل الشبكي، يدمج تحليل حركة مرور الشبكة، ومعالجة البيانات المسبقة، وتوسيع نطاق البيانات، وتقنيات الذكاء الاصطناعي.

يبدأ العمل بدراسة مقارنة لثلاث منصات لتحليل حركة مرور الشبكة. يتم تقييم هذه الأدوات بناءً على قدرتها على معالجة ملفات PCAP الخاصة بـ Kubernetes واستخراج السمات ذات الصلة من تدفقات الشبكة. علاوة على ذلك، تُستخدم مجموعتا بيانات مرجعيتان، للتحقق من قابلية المنهج المقترح.

لمعالجة مشكلة عدم توازن الفئات الشائعة في مجموعات بيانات كشف الاختراقات، يُستخدم مُشَقِّر تلقائي تبايني شرطي (CVAE) لتوليد عينات اصطناعية لفئات الهجمات الأقل تمثيلاً. تم اختيار هذا النهج لقدرته على نمذجة توزيعات البيانات المعقدة وتوليد تدفقات اصطناعية خاصة بكل فئة بناءً على تصنيفات الهجمات. بالمقارنة مع طرق أخذ العينات الزائدة التقليدية، مثل أخذ العينات الزائدة العشوائية أو SMOTE، تُستخدم مجموعات البيانات المتوازنة بعد ذلك لتدريب وتقييم نماذج متعددة للتعلم الآلي، بالإضافة إلى نموذج قائم على Transformer.

أظهرت النتائج التجريبية أن زيادة البيانات القائمة على CVAE تحسّن بشكل ملحوظ أداء التصنيف للنماذج الحساسة لعدم توازن الفئات، وخاصةً نموذج DistilBERT حيث ارتفع معامل F1 من 75% إلى 99% على مجموعة بيانات UNSW-NB15، ومن 29% إلى 99% على مجموعة بيانات NFStream المُعالجة، بينما لم تُحقق سوى مكاسب طفيفة للنماذج الشجرية القوية بالفعل، مثل XGBoost و Random و Forest من حيث الدقة والضبط والاستدعاء ومعامل F1. وأخيراً، تم نشر النموذج الأفضل أداءً في FlowVertex وهي منصة لتحليل الشبكات في الوقت الفعلي، مما يبرهن على التطبيق العملي للإطار المقترح في بيئة مراقبة الشبكات التشغيلية، حيث يجب تحليل وتصنيف تدفقات الشبكة بزمان استجابة منخفض. بشكل عام، تؤكد هذه الأطروحة على أهمية الجمع بين أدوات تحليل حركة مرور الشبكة، والنماذج التوليدية، وتقنيات التعلم الآلي المتقدمة لبناء أنظمة كشف تسلل قوية وقابلة للتكيف في بيئات الأمن السيبراني الحديثة.

الكلمات المفتاحية : نظام كشف التسلل (IDS) تحليل حركة مرور الشبكة، التعلم الآلي، التعلم العميق، الذكاء الاصطناعي التوليدي، الأمن السيبراني، زيادة البيانات.