

Algerian Democratic and Popular Republic
Ministry of Higher Education and Scientific Research
Higher School of Computer Science
- 08 May 1945 - Sidi Bel Abbès



THESIS

To obtain the Master's Degree

Field: **Computer Science**

Specialty: **Artificial Intelligence and Data Science**

Theme:

**Real-Time Anomaly Detection in Streaming Event Data: A
Literature Review and Comparative Study**

Presented by:

BENGHENIMA Hafsa

In front of the jury composed of:

Dr. ELOUALI Nadia

President

Dr. DIF Nassima

Examiner

Pr. BENSLIMANE Sidi Mohammed

Supervisor

Dr. BENABDERRAHMANE Sid Ahmed

Supervisor

Année universitaire **2025/2026**

Abstract

The exponential growth of event data generated across cybersecurity infrastructures, enterprise event logs, financial systems, and medical environments has made real-time anomaly detection in high-velocity streams a critical operational challenge. Most traditional methods were designed for static datasets, and they struggle to operate under streaming constraints such as single-pass processing, limited memory, and the absence of ground-truth labels during inference.

This thesis presents a structured literature review of existing approaches to anomaly detection in continuous event streams, covering unsupervised machine learning, deep learning, and frequency-based methods. The first part lays the groundwork by covering the main concepts in machine learning and deep learning, streaming data processing, and approximate frequency estimation, as well as the NLP and large language model techniques that support modern explanation-driven anomaly detection.

The second part reviews the state of the art across four main directions: sketch-based and frequency-driven approaches, deep learning methods for streaming anomaly detection, pattern-based rarity mining, and explainable anomaly detection using retrieval-augmented generation. A comparative analysis identifies persistent gaps in the literature, particularly the lack of unified frameworks that jointly address streaming compatibility, multi-signal score integration, and interpretable output generation across heterogeneous domains.

Keywords: Anomaly Detection, High-Velocity Event Streams, Unsupervised Learning, Retrieval-Augmented Generation, Explainability.

RÉSUMÉ

La croissance exponentielle des données d'événements générées au sein des infrastructures de cybersécurité, des journaux d'événements d'entreprise, des systèmes financiers et des environnements médicale a fait de la détection d'anomalies en temps réel dans les flux à haute vitesse un défi opérationnel majeur. Les méthodes traditionnelles de détection d'anomalies, conçues pour des ensembles de données statiques, ne répondent pas aux contraintes des environnements en streaming, notamment le traitement en un seul passage, la mémoire limitée et l'absence de vérités terrain au moment de l'inférence.

Cette mémoire présente une revue de littérature structurée des approches existantes de détection d'anomalies dans les flux d'événements continus, couvrant l'apprentissage automatique non supervisé, l'apprentissage profond et les méthodes basées sur la fréquence. La première partie pose les bases théoriques en présentant les concepts fondamentaux de l'apprentissage automatique et profond, le traitement des données en streaming, ainsi que les méthodes d'estimation de fréquence approximative, en plus des techniques de traitement du langage naturel et des grands modèles de langage qui soutiennent les systèmes modernes d'explication d'anomalies.

La deuxième partie passe en revue l'état de l'art selon quatre axes principaux : les méthodes basées sur les sketches et la fréquence, les approches d'apprentissage profond pour la détection en flux, les techniques de mining de motifs rares, ainsi que la détection explicable des anomalies via la génération augmentée par récupération (RAG). Une analyse comparative met ensuite en évidence les principales limites des travaux existants, en particulier l'absence de cadres unifiés capables de gérer simultanément les contraintes du streaming, l'intégration de plusieurs signaux d'anomalie et la génération d'explications interprétables dans des domaines hétérogènes.

Mots-clés : Détection d'anomalies, flux d'événements à haute vitesse, apprentissage non supervisé, génération augmentée par récupération (RAG), explicabilité.

ملخص

أدى النمو المتسارع لبيانات الأحداث المُولدة عبر البنى التحتية للأمن السيبراني، وسجلات الأحداث المؤسسية، والأنظمة المالية، والبيئات الطبية، إلى جعل كشف الشذوذ في الوقت الحقيقي ضمن التدفقات عالية السرعة تحدياً تشغيلياً بالغ الأهمية. تم تصميم معظم الطرق التقليدية للتعامل مع بيانات ثابتة، ولذلك فهي تواجه صعوبة عند العمل ضمن قيود بيئات التدفق، مثل المعالجة ذات المرور الواحد، والذاكرة المحدودة، وغياب القيم الحقيقية أثناء الاستدلال.

تقدم هذه الأطروحة مراجعة منهجية للأدبيات المتعلقة بالأساليب الحالية للكشف عن الشذوذ في تدفقات الأحداث المستمرة، وتغطي التعلم الآلي غير الخاضع للإشراف، والتعلم العميق، والأساليب المعتمدة على التكرار أو التردد. يضع الجزء الأول الأسس النظرية من خلال تغطية المفاهيم الأساسية في التعلم الآلي والتعلم العميق، ومعالجة البيانات المتدفقة، وتقنيات تقدير التردد التقريبي، بالإضافة إلى تقنيات معالجة اللغة الطبيعية ونماذج اللغة الكبيرة التي تدعم أنظمة تفسير الشذوذ الحديثة.

أما الجزء الثاني فيستعرض أحدث ما توصل إليه البحث في أربعة محاور رئيسية: الأساليب المعتمدة على الهياكل المدججة والتردد، أساليب التعلم العميق لكشف الشذوذ في التدفقات، استخراج الأنماط النادرة، وكشف الشذوذ القابل للتفسير باستخدام التوليد المعزز بالاسترجاع. ثم يتم تقديم تحليل مقارنة يسلط الضوء على أهم القيود في الأعمال الحالية، خصوصاً غياب أطر موحدة قادرة على التعامل مع قيود التدفق، ودمج إشارات شذوذ متعددة، وتوليد تفسيرات قابلة للفهم عبر مجالات مختلفة.

الكلمات المفتاحية: كشف الشذوذ، تدفقات الأحداث عالية السرعة، التعلم غير المراقب، التوليد المعزز بالاسترجاع، (RAG) القابلية للتفسير.