

الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 08 ماي 1945 بسيدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès



THESIS

to obtain the diploma of **Master**.

Field: **Computer Science**

Specialty: **Artificial Intelligence and Data Science (IASD)**

Thesis Title

**A Survey on Log-Based Anomaly Detection in
Distributed Systems: Methods, Benchmarks, and
Explainability**

Presented by:
Toumi Adem

Submission Date: **June 2026**
Defended before a jury composed of:

Dr. KHALDI Miloud
Dr. Sidi Mohammed Benslimane
Dr. Mohammed Bekkouche
Dr. ANANI Djihed

Président
Supervisor
Co-Supervisor
Examineur

Academic Year : 2025/2026

Abstract

Modern distributed systems generate massive volumes of execution logs that serve as the primary observability interface for detecting operational anomalies. As the scale and heterogeneity of these systems grow, manual inspection becomes infeasible, making automated, data-driven approaches essential for system reliability.

This thesis presents a systematic literature survey and comparative synthesis of log-based anomaly detection in distributed systems. Unlike implementation-oriented studies that propose and evaluate novel detection architectures, this work focuses on analyzing, categorizing, and cross-comparing existing representative methodologies across three paradigms—supervised machine learning classifiers, deep learning sequential models, and unsupervised reconstruction-based methods. Using the standard benchmark datasets HDFS, BGL, and Spirit as a unified comparative foundation, we systematically compile and contrast reported performance metrics, experimental protocols, and architectural design choices from the literature. Our synthesis identifies critical structural gaps: the absence of unified cross-dataset evaluation frameworks, the insufficient treatment of class imbalance in experimental reporting, and the near-total lack of explainability mechanisms that would allow system operators to interpret and trust anomaly alerts. We formulate four research questions that guide the comparative goals of this survey and outline structured directions for future research aimed at building systems that are simultaneously accurate, generalizable, imbalance-robust, and interpretable.

The findings of this work provide a structured, evidence-based foundation for advancing the field and are directly applicable to improving the reliability of large-scale infrastructures, cloud environments, and supercomputing clusters.

Keywords— Anomaly detection, execution logs, distributed systems, machine learning, deep learning, state of the art, explainable AI.

الملخص

تُنتج الأنظمة الموزعة الحديثة كميات هائلة من سجلات التنفيذ التي تُعد الواجهة الأساسية لمراقبة واكتشاف الحالات الشاذة التشغيلية. ومع تزايد حجم وتعقيد هذه الأنظمة، يصبح الفحص اليدوي غير عملي، مما يجعل الأساليب الآلية المعتمدة على البيانات ضرورية لضمان موثوقية النظام.

تقدم هذه الأطروحة دراسة منهجية لطرق اكتشاف الحالات الشاذة المعتمدة على السجلات في الأنظمة الموزعة. نقوم بمراجعة ومقارنة الأعمال الممثلة عبر ثلاثة نماذج: المصنفات الخاضعة للإشراف في التعلم الآلي، والنماذج التسلسلية للتعلم العميق، والأساليب غير الخاضعة للإشراف المعتمدة على إعادة البناء. يتم تقييم هذه المناهج باستخدام مجموعات بيانات مرجعية قياسية، بما في ذلك HDFS و BGL و Spirit.

تحدد تحليلاتنا فجوات حاسمة في الأدبيات الحالية، وتحديدًا غياب أطر التقييم الموحدة عبر مجموعات البيانات المختلفة، والمعالجة غير الكافية لعدم توازن الفئات، والغياب شبه التام لآليات التفسير التي تسمح لمشغلي الأنظمة بتفسير تنبيهات الحالات الشاذة والثقة بها.

توفر نتائج هذا العمل أساساً منظماً للبحوث المستقبلية ويمكن تطبيقها لتحسين موثوقية البنى التحتية واسعة النطاق، والبيئات السحابية، ومجموعات الحوسبة الفائقة.

الكلمات المفتاحية--- اكتشاف الحالات الشاذة، سجلات التنفيذ، الأنظمة الموزعة، التعلم الآلي، التعلم العميق، أحدث ما توصل إليه العلم، الذكاء الاصطناعي القابل للتفسير.

Résumé

Les systèmes distribués modernes génèrent des volumes massifs de journaux d'exécution (logs) qui constituent l'interface d'observabilité principale pour la détection des anomalies opérationnelles. Face à l'échelle et à l'hétérogénéité croissantes de ces systèmes, l'inspection manuelle devient impraticable, rendant les approches automatisées basées sur les données essentielles pour la fiabilité des systèmes.

Cette mémoire présente une étude systématique des méthodes de détection d'anomalies basées sur les logs pour les systèmes distribués. Nous passons en revue et comparons des travaux représentatifs à travers trois paradigmes : les classificateurs supervisés, les modèles séquentiels d'apprentissage profond, et les méthodes non supervisées basées sur la reconstruction. Ces approches sont évaluées sur des ensembles de données de référence tels que HDFS, BGL et Spirit.

Notre analyse identifie des lacunes critiques dans la littérature existante, notamment l'absence de cadres d'évaluation croisée unifiés, le traitement insuffisant du déséquilibre de classes, et le manque quasi total de mécanismes d'explicabilité permettant aux opérateurs d'interpréter et de faire confiance aux alertes.

Les résultats de ce travail fournissent une base structurée pour les recherches futures et s'appliquent à l'amélioration de la fiabilité des infrastructures à grande échelle, des environnements cloud et des supercalculateurs.

Mots clés— Détection d'anomalies, journaux d'exécution, systèmes distribués, apprentissage automatique, apprentissage profond, état de l'art, IA explicable.