Bachelor Thesis Proposal

Analysis of Established Datasets and Increasing Complexity in Anomaly Detection

1 Introduction and Motivation

Anomaly detection—the identification of data points, patterns, or events that deviate significantly from expected normal behavior—is a central research area in machine learning with broad practical relevance. Its applications span safety-critical and industrial domains, including network intrusion detection, financial fraud identification, predictive maintenance, and medical diagnostics. In each of these settings, the timely and accurate identification of anomalous observations is essential for preventing financial losses, mitigating security threats, or averting physical harm.

Despite substantial advances in algorithmic approaches—from classical statistical methods to deep learning architectures—the empirical evaluation of anomaly-detection methods continues to depend heavily on a relatively small set of established benchmark datasets. A growing body of evidence suggests that many of these benchmarks exhibit limited complexity, artificial separability between normal and anomalous instances, or insufficient diversity in the types of anomalies they contain. This discrepancy between benchmark simplicity and real-world complexity raises fundamental questions about the generalizability and practical validity of reported experimental results.

This thesis proposal outlines a systematic investigation into the properties and limitations of established anomaly-detection datasets and proposes methods for deliberately increasing their complexity in order to enable more realistic and rigorous model evaluations.

2 Problem Statement

The central hypothesis underlying this thesis is that the restricted complexity of many commonly used benchmark datasets imposes a systematic limitation on the validity and transferability of experimental evaluations in anomaly detection. If the datasets used for benchmarking do not adequately reflect the characteristics of real-world anomaly scenarios—such as ambiguous decision boundaries, temporal non-stationarity, high dimensionality, and heterogeneous anomaly types—then the performance rankings and comparative conclusions derived from such evaluations may not generalize to practical deployment settings.

This thesis aims to address this gap by (a) systematically analyzing the properties and limitations of established datasets, (b) formalizing the key dimensions along which dataset complexity can be characterized, and (c) proposing and evaluating methods for controlled complexity augmentation.

3 Research Questions

The proposed thesis addresses the following research questions:

1. Which established datasets are commonly used for anomaly detection benchmarking, and what are their key structural and statistical characteristics?
2. Along which dimensions is the complexity of these datasets limited, and how do these limitations affect the validity of experimental evaluations?

3. Which methods can be employed to systematically increase dataset complexity while preserving comparability and enabling controlled experimentation?

4 Theoretical Background

4.1 Anomaly Detection Paradigms

Anomaly-detection approaches are commonly classified by the degree of supervision available during training. Supervised methods require labeled examples of both normal and anomalous instances, but are often impractical due to the scarcity of anomaly labels. Semi-supervised approaches train exclusively on normal data, detecting anomalies as deviations from the learned representation of normality. Unsupervised methods operate without labels entirely, relying on distributional properties such as density, distance, or reconstruction fidelity.

4.2 Prevalent Model Families

Relevant model families for this thesis include statistical methods (e.g., Gaussian mixture models, kernel density estimation), distance- and density-based approaches (e.g., k-nearest neighbors, Local Outlier Factor), reconstruction-based methods (e.g., autoencoders, variational autoencoders), and deep learning architectures (e.g., GANs, self-supervised contrastive learning, transformers). The effectiveness of each family is strongly dependent on dataset properties—a central concern of this work.

5 Planned Approach and Methodology

5.1 Dataset Survey and Analysis

A systematic literature review will be conducted to identify the most frequently used benchmark datasets for anomaly detection. Datasets will be categorized by domain (image-based industrial inspection, network security, time series) and analyzed with respect to their structural properties, anomaly characteristics, and documented limitations. The output of this phase will be a comparative overview highlighting the key strengths and deficiencies of existing benchmarks.

5.2 Complexity Framework

Building on the dataset analysis, a multi-dimensional framework for characterizing dataset complexity will be developed. The proposed framework encompasses four principal dimensions:

- Structural complexity: dimensionality, inter-feature correlations, and non-linear dependencies.
- Semantic complexity: ambiguity in the boundary between normal and anomalous behavior.
- Temporal complexity: concept drift, seasonal effects, and delayed or gradual-onset anomalies.
- Anomaly complexity: rarity, distributional overlap with normal data, and subtlety of deviations.

5.3 Complexity Augmentation Methods

Four complementary methods for systematically increasing dataset complexity will be developed and applied:

1. Domain-aware anomaly injection: embedding anomalies in a context-sensitive manner rather than through random perturbation, to improve ecological validity.
2. Controlled dimensionality augmentation: appending correlated, partially informative, or irrelevant features to simulate real-world noise and redundancy.
3. Temporal extension and drift simulation: transforming static datasets into time-series representations with realistic non-stationarity.
4. Multimodal data fusion: combining heterogeneous data modalities (e.g., image and sensor data) to reflect modern monitoring environments.

5.4 Evaluation Framework

A standardized evaluation protocol will be designed to assess the impact of complexity augmentation on anomaly-detection performance. Multiple detection models will be evaluated on both original and augmented

datasets using established metrics, including AUC-ROC, precision–recall curves, and robustness to noise. The analysis will focus not only on absolute performance but also on the stability and interpretability of model rankings as complexity increases

6 Expected Contributions

This thesis is expected to deliver the following contributions:

- A systematic comparative analysis of established anomaly-detection benchmark datasets, identifying their key properties and limitations.
- A multi-dimensional complexity framework that formalizes the axes along which dataset complexity can be characterized and extended.
- A set of practical augmentation methods for increasing dataset complexity in a controlled and reproducible manner.
- An empirical evaluation demonstrating the impact of complexity augmentation on model performance and ranking stability.

Together, these contributions aim to provide researchers and practitioners with both a diagnostic lens for assessing existing benchmarks and a practical toolkit for constructing more representative evaluation scenarios.