# Use of AI for Log Analysis in CI/CD Pipelines

Bachelor Thesis – Defence

Maid Ališić

July 23, 2025

FH Oberösterreich · Campus Hagenberg

Supervisor: FH-Prof. Dipl.-Ing. Dr. Stefan Wagner

# Road map

Research questions

Problem context

Method

Architecture

Data & evaluation

Results

Impact

# Research questions

# Research questions

**RQ$_{\textbf{main}}$**:
How can AI improve accuracy & explainability of CI/CD log analysis?

- **RQ1**: Capability of a cloud LLM (ChatGPT) to interpret logs
- **RQ2**: How does a local IF+RF stack compare?
- **RQ3**: Integration hurdles inside existing pipelines

# Problem context

# Why do we care?

- CI/CD emits $\approx$ 10–20 GB of build, test & deploy logs per day. [2]

# Why do we care?

- CI/CD emits ≈ 10–20 GB of build, test & deploy logs per day. [2]
- Manual `grep` slows the merge queue; critical faults slip through.

# Why do we care?

- CI/CD emits $\approx$ 10–20 GB of build, test & deploy logs per day. [2]
- Manual `grep` slows the merge queue; critical faults slip through.
- Business **Service-Level Objective**: feedback within $\leq$ 200 ms per pipeline.

# Why do we care?

- CI/CD emits $\approx$ 10–20 GB of build, test & deploy logs per day. [2]
- Manual `grep` slows the merge queue; critical faults slip through.
- Business **Service-Level Objective**: feedback within $\leq$ 200 ms per pipeline.
- Logs may expose customer IDs, therefore they must remain on-premises (no cloud export).

# Operational pain points

1. **Context-sensitivity** – identical tokens can be harmless or fatal.

# Operational pain points

1. **Context-sensitivity** – identical tokens can be harmless or fatal.
2. **Concept drift** – each merge may rename tests or switches. [4]

# Operational pain points

1. **Context-sensitivity** – identical tokens can be harmless or fatal.
2. **Concept drift** – each merge may rename tests or switches. [4]
3. **Latency pressure** – analysis must finish before runner teardown.

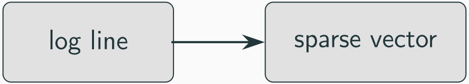
# Operational pain points

1. **Context-sensitivity** – identical tokens can be harmless or fatal.
2. **Concept drift** – each merge may rename tests or switches. [4]
3. **Latency pressure** – analysis must finish before runner teardown.
4. **Alert fatigue** – regex rule sets grow without bound. [3]

# Method

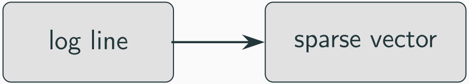
# Vectorisation pipeline ①

1. **Normalise** – strip timestamps, colours, IDs.



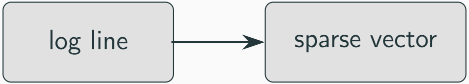
# Vectorisation pipeline ①

1. **Normalise** – strip timestamps, colours, IDs.
2. **Tokenise** into uni- and bi-grams.



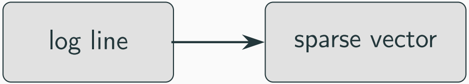
# Vectorisation pipeline ①

1. **Normalise** – strip timestamps, colours, IDs.
2. **Tokenise** into uni- and bi-grams.
3. Weight with TF-IDF.



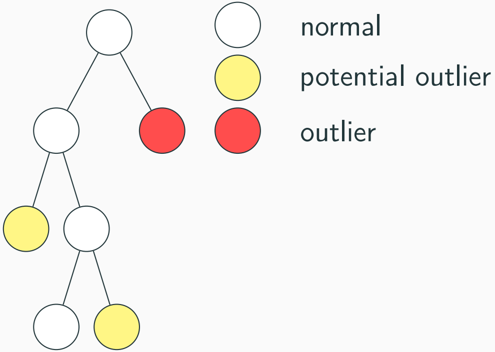
# Vectorisation pipeline ①

1. **Normalise** – strip timestamps, colours, IDs.
2. **Tokenise** into uni- and bi-grams.
3. Weight with TF-IDF.
4. Produce sparse vector.



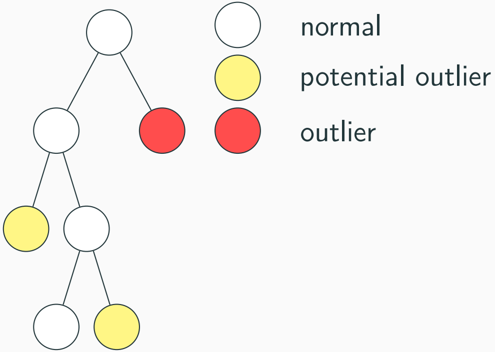
# Isolation Forest ② – intuition

- Random binary partitioning isolates unusual lines in fewer splits. [1]



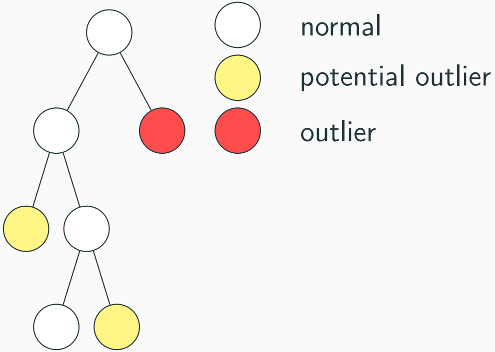
# Isolation Forest ② – intuition

- Random binary partitioning isolates unusual lines in fewer splits. [1]
- Score $s(x) = 2^{-h(x)/c(n)} \in [0, 1]$; high $\Rightarrow$ outlier.



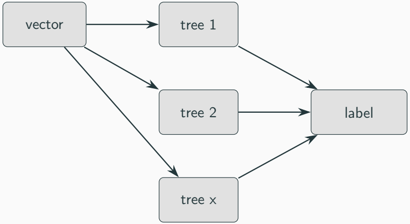
# Isolation Forest ② – intuition

- Random binary partitioning isolates unusual lines in fewer splits. [1]
- Score $s(x) = 2^{-h(x)/c(n)} \in [0, 1]$; high $\Rightarrow$ outlier.
- CPU-only: $\approx 30\,\mu s$ per line (prototype measurement).



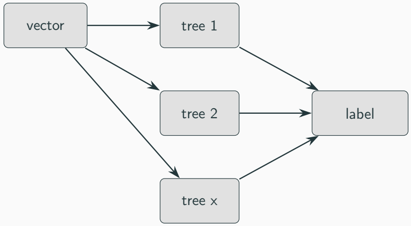
# Random Forest ③ – error labelling

- Maps each flagged line to a domain-specific error category. [5]



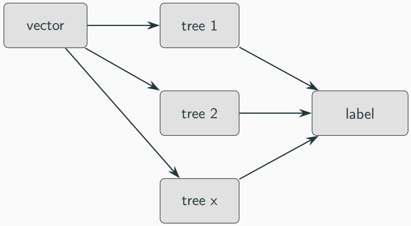
# Random Forest ③ – error labelling

- Maps each flagged line to a domain-specific error category. [5]
- Majority vote = deterministic, auditable output.



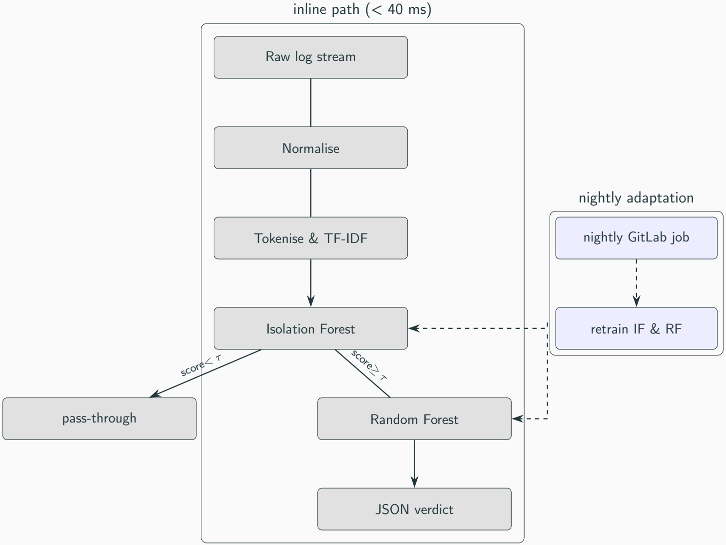
# Random Forest ③ – error labelling

- Maps each flagged line to a domain-specific error category. [5]
- Majority vote = deterministic, auditable output.
- Nightly retrain $< 90$ s; warm-start handles drift.



# Architecture

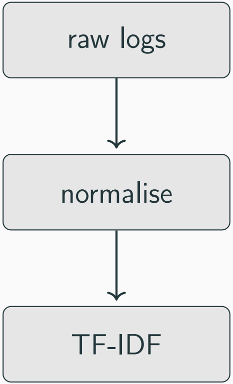
# End-to-end pipeline (< 40 ms inline)



# Data & evaluation

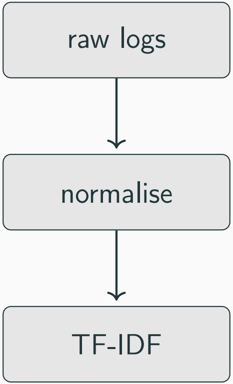
# Datasets & metrics

- 117 k *macOS* logs + 655 k *OpenSSH* logs. [6]



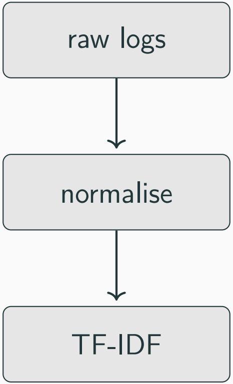
# Datasets & metrics

- 117 k *macOS* logs + 655 k *OpenSSH* logs. [6]
- 504 labelled anomalies (class imbalance $\approx 1 : 200$)



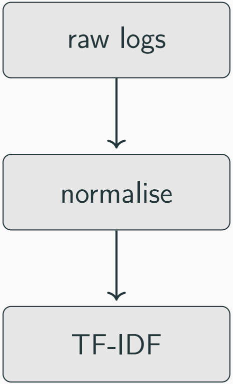
# Datasets & metrics

- 117 k *macOS* logs + 655 k *OpenSSH* logs. [6]
- 504 labelled anomalies (class imbalance ≈ 1 : 200)
- Split 70 / 15 / 15 % (training / validation / testing)



# Datasets & metrics

- 117 k *macOS* logs + 655 k *OpenSSH* logs. [6]
- 504 labelled anomalies (class imbalance $\approx 1:200$)
- Split 70 / 15 / 15 % (training / validation / testing)
- Metrics: Precision, Recall and $F_1$



# Results

# Headline numbers

| | Precision | Recall | $F_1$ |
|---|---|---|---|
| Detection (Isolation Forest) | 0.91 | 0.88 | 0.89 |
| Classification (Random Forest) | 0.99 | 0.99 | 0.99 |
| Regex Baseline | 0.286 | 0.286 | 0.286 |

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

# Impact

# Operational impact

- **Latency**: minutes $\rightarrow$ **milliseconds** (inline verdict).

# Operational impact

- **Latency**: minutes $\rightarrow$ **milliseconds** (inline verdict).
- **Cost-free**: 2.3 k lines of code, CPU-only, no token fees.

## Operational impact

- **Latency**: minutes $\rightarrow$ **milliseconds** (inline verdict).
- **Cost-free**: 2.3 k lines of code, CPU-only, no token fees.
- **GDPR compliant**: logs never leave the VPN.

# References


[1] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. **"Isolation Forest".** In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 413–422.

[2] Splunk Inc. ***Observability at Scale – Field Report 2023*.** 2023. URL: https://www.splunk.com/en_us/blog/devops/the-state-of-observability-2023-realizing-roi-and-increasing-digital-resilience.html.

[3] Shilin He et al. **"A Survey on Automated Log Analysis for Reliability Engineering".** In: *ACM Comput. Surv.* 54.6 (2021).

[4] João Gama et al. **"A survey on concept drift adaptation".** In: *ACM Comput. Surv.* 46.4 (2014).

[5] Leo Breiman. **"Random Forests".** In: *Machine Learning* 45.1 (2001), pp. 5–32.

[6] Jieming Zhu et al. ***Loghub: A Large Collection of System Log Datasets for AI-driven Log Analytics*.** 2023.

# Wrap-up

## Take-away

**Light-weight on-prem ML matches AIOps SaaS**

without latency, cost or privacy pain.

Questions welcome – thank you!