RepoShield – Technical Reference Document

1. Overview

RepoShield is a predictive repository risk analysis system designed to evaluate and rank software repositories based on security, vulnerability likelihood, and maintenance risk. It aggregates signals from multiple analyzers such as static analysis tools, machine learning models, dependency vulnerability scanners, and repository activity metrics to compute a single, normalized risk score.

The primary objective of RepoShield is to support early risk detection, prioritization, and informed decision-making for developers and security teams.

---

2. Risk Scoring Formula

The core scoring mechanism of RepoShield is defined as:

$$\text{Risk Score} = \frac{\sum_{i=1}^{n}(\text{risk}_i \times \text{confidence}_i^2)}{\sum_{i=1}^{n}\text{confidence}_i}$$

Where: - n is the total number of analyzers - $\text{risk}_i \in [0,1]$ is the predicted risk from analyzer i - $\text{confidence}_i \in [0,1]$ is the confidence or reliability score of analyzer i

---

3. Mathematical Justification

The formula represents a confidence-weighted average of multiple risk predictions.

Key Rationale:

- Confidence is treated as a weighting factor to reflect trust in each analyzer.
- Squaring the confidence value increases separation between high-confidence and low-confidence analyzers, ensuring reliable analyzers dominate the final score.
- Normalization by the total confidence prevents score inflation and allows fair comparison across repositories with varying analyzer coverage.

This approach is inspired by ensemble learning, Bayesian model averaging, and weighted decision fusion commonly used in predictive systems.

---

4. Theoretical Foundations

RepoShield's aggregation strategy is grounded in: - Weighted averaging techniques in ensemble machine learning - Bayesian confidence-based model weighting - Multi-signal risk aggregation used in modern vulnerability prioritization systems

---

5. Supported Analyzer Types

Each analyzer outputs a tuple of (risk, confidence):

- Static code vulnerability analysis
- Dependency vulnerability analysis (CVE-based)
- Machine learning vulnerability prediction models
- Commit behavior and anomaly detection
- Repository health metrics (activity trends, issue resolution, contributor count)

The system is extensible and supports adding new analyzers without changing the aggregation logic.

---

6. Complexity Analysis

For a single repository evaluated by n analyzers:

Time Complexity

- Risk aggregation: $O(n)$
- Total runtime: $\sum O(f_i) + O(n)$, where f_i is the execution cost of analyzer i

Space Complexity

- Aggregation memory: $O(n)$

The scoring layer is computationally lightweight; the dominant cost lies in analyzer execution.

---

7. Research References

The following research works support the design and methodology of RepoShield:

- Open and Adaptable Approach to Vulnerability Risk Scoring
https://doi.org/10.32604/jcs.2025.064958

- Data-Driven Vulnerability Prioritization and Exploit Prediction
https://arxiv.org/abs/2302.14172

- Software Vulnerability Prediction: A Systematic Review
  https://doi.org/10.1016/j.infsof.2023.107303

- Hybrid Vulnerability Scoring Systems Beyond CVSS
  https://doi.org/10.1016/j.cose.2023.103256

- Anomalicious: Detecting Malicious and Risky Commits
  https://arxiv.org/abs/2103.03846

8. Summary

RepoShield provides a scalable, explainable, and confidence-aware framework for repository risk assessment. The proposed scoring formula ensures that high-confidence analyzers have a stronger influence while maintaining normalization and interpretability. Backed by established research in vulnerability scoring and ensemble learning, RepoShield is suitable for real-world predictive security analysis and prioritization workflows.