로그인 실패 횟수 임계값이 이상 접속 탐지 성능에 미치는 영향

부제: 탐지 누락률과 오탐률의 상충관계에 관한 재현 가능한 합성 데이터 연구
영문 제목: Effects of Login Failure Count Thresholds on Authentication Anomaly Detection: A Reproducible Synthetic-Data Study
작성자: 김봉경
작성일: 2026 년 11 월 10 일
자료 범위: 실제 인증 기록이 아닌 합성 데이터. 과제 T10 의 기존 정규화 논문과 별개의 C 안 신규 연구.

초록

본 연구는 로그인 실패 횟수 기반 이상 접속 탐지에서 임계값 선택이 누락률과 오탐률에 미치는 영향을 검증했다. 기존 정규화 실험의 6 회 기준 결과를 예비 관찰로 구분하고, 새로운 난수 시드 10 개로 생성한 합성 인증 로그 6,551 건을 5 분 단위 1,000 구간(공격 300·정상 700)으로 구성했다. 4~10 회 임계값을 모두 비교하되, 확인적 가설은 6 회에서 5 회로 낮추면 누락률이 감소하면서 오탐 증가폭은 5%p 이내라는 조건으로 미리 정했다. 5 회 기준 누락률은 9.67%로 6 회(22.33%)보다 12.67%p 낮았지만, 오탐률은 14.57%에서 29.00%로 14.43%p 증가하여 가설은 지지되지 않았다. 사후 탐색적 대체 분포에서도 7 회 기준 오탐이 발생했고, 임계값 성능은 데이터 분포와 비용 가정에 의존함을 보였다. 이 결과는 지정한 합성 규칙에 한정되며 실제 기업 보안 정책의 최적 임계값을 뜻하지 않는다.

주제어: 인증 로그, 무차별 대입 공격, 임계값, 오탐률, 탐지 누락률, 재현 실험

1. 서론

1.1 연구 배경과 문제 정의

여러 서비스의 인증 실패 기록을 이용해 반복적인 로그인 시도를 감시하는 것은 보안 운영의 기본적인 탐지 유형 중 하나다. MITRE ATT&CK T1110 은 무차별 대입을 자격증명 공격 유형으로 설명하며, 탐지 전략 DET0463 에서 시간 창과 실패 횟수를 조정 가능한 탐지 요소로 제시한다(MITRE ATT&CK, 2026). 그러나 로그인 실패 횟수만으로 공격인지 정상 사용자 실수인지 완벽하게 판정할 수는 없다. 낮은 임계값은 경보를 많이 만들어 오탐 부담을 늘릴 수 있고, 높은 임계값은 일부 공격을 놓칠 수 있다.

기존 T10 연구에서는 원문 직접 탐지와 정규화 후 탐지에 동일한 6 회 기준을 적용했고 두 방식의 결과가 일치했다. 해당 연구는 로그 처리 구조를 비교했으나 탐지 기준 자체의 영향은 검증하지 못했다. 본 연구는 그 한계를 바탕으로 **탐지 임계값 자체를 독립변수로 바꾸는 새로운 연구**로 설계했다. 기존 데이터는 새 가설을 검증하는 데 재사용하지 않았다.

1.2 연구 질문과 가설

연구 질문은 "동일한 합성 인증 로그에서 5 분 이내 로그인 실패 횟수 임계값을 변경할 때 누락률과 오탐률의 균형은 어떻게 달라지는가?"다. 신규 데이터를 생성하기 전에 다음 가설을 확정했다.

> **H1:** 동일한 5 분 단위 인증 로그에서 실패 횟수 임계값을 6 회에서 5 회로 낮추면, 공격 탐지 누락률이 감소하면서 정상 구간 오탐률 증가폭은 5%p 이내일 것이다.

H1 의 지지 조건은 (a) FNR(5) < FNR(6), (b) FPR(5) − FPR(6) ≤ 5.00%p 를 모두 만족하는 것이다. 하나라도 충족하지 못하면 지지하지 않는다고 판단한다. 4, 7, 8, 9, 10 회 결과는 정책별 특성을 확인하기 위한 탐색적 비교이며, 가장 좋은 값이 나오도록 가설을 사후 변경하지 않는다.

1.3 주제 전환과 가설 구체화의 기록

이번 연구의 출발점은 「보안 로그 정규화가 이상 탐지 누락률을 낮추는가」였다. 초기 합성 실험에서는 원문 탐지와 정규화 후 탐지가 필요한 정보를 모두 추출했고, 두 방식의 탐지 결과 차이가 0%p 였다. 사용자는 이 비교가 연구로서 충분히 의미 있는지 질문했다. AI 는 기존 연구 유지, 형식 변화 대응력 추가 검증, 임계값 비교 연구의 세 방향을 제안했다. 사용자는 C 안 계획을 요청했고 이어 실행을 선택했다. 이 경과는 C 안의 연구 질문을 정한 배경이며, 이전 실험 결과는 신규 H1 의 검증 자료에 포함되지 않았다.

AI 는 임계값을 6 회에서 5 회로 낮추는 비교와 함께 **오탐률 증가 허용폭 5%p** 를 연구용 판정 조건으로 제안했다. 5%p 는 외부 보안 표준이나 실증 논문의 권고값이 아니라, 누락 감소 효과와 오탐 부담을 동시에 평가하기 위한 **임의적 허용폭**이다. 사용자가 이 수치를 독립적으로 선정하거나 학술적으로 검증한 기록은 없다. 신규 데이터 생성 전에 작성된 C 안 계획에 이 조건을 고정했고, 결과를 확인한 뒤에도 변경하지 않았다. 구체적인 대화 경과와 역할 구분은 docs/05_AI_가설검토_수정이력.md 에 정리했다.

2. 선행 연구

Axelsson(2000)은 침입 탐지에서 오탐과 실제 공격의 발생 빈도(기저율)를 함께 고려해야 한다고 분석했다. Javed(2016)는 실제 기업 인증 기록을 분석해 경보 빈도와 탐지의 균형을 다루는 기법을 연구했다. Singh 등(2024)은 생산 환경의 500 대 이상 서버에서 수집한 SSH 로그를 분석하여 빈도 기반 차단 방식과 다른 방어 접근을 비교했다. Khandait 등(2021)은 운영 SSH 인증 로그를 분석하고 로그인

시도와 공격 탐지에 관한 실증 사례를 보고했다. 이들 실증 연구는 인증 로그가 관측 가능한 연구 대상임을 보여주지만, 본 연구의 합성 데이터 생성 범위와 5 회 임계값을 정당화하는 근거는 아니다.

MITRE ATT&CK(2026)의 탐지 전략은 실패 횟수·시간 창이 탐지 구성요소임을 제시한다. NIST SP 800-63B-4(2025)는 온라인 추측 공격에 대한 인증 시도 제한 등 대응 배경을 제공한다. 그러나 **인증 수단/계정별 시도 제한과 본 논문의 IP 별 5 분 경보 임계값은 다른 개념**이므로 NIST 가 5 회 또는 6 회 기준을 권고한 것처럼 해석하지 않는다. 출처·초록의 직접 검증 과정과 제외 문헌은 docs/02_선행연구_인용검증.md 에 별도로 기록했다.

3. 연구 방법

3.1 설계와 조작적 정의

본 연구는 조건별 로그를 다르게 생성하지 않고, **동일한 1,000 구간**에 7 개의 임계값을 반복 적용하는 짝지은 비교 설계다. 관측 단위는 특정 IP 의 UTC 기준 5 분 구간이며, 실패 이벤트 수가 임계값 이상이면 해당 구간을 이상으로 판정한다. 공격·정상 정답은 별도 라벨 파일로 저장하여 탐지 계산에 사용하지 않고 평가 시에만 결합한다.

설계 요소	정의 및 고정 조건
독립변수	실패 횟수 임계값 4, 5, 6, 7, 8, 9, 10 회
확인적 비교	5 회와 6 회; FNR 감소와 FPR 증가폭 5%p 이내를 동시 충족하는지
고정 조건	동일 로그, IP 별 5 분 집계, 동일 파서, 동일 정답표, 동일 실행 코드
반복 횟수	독립 난수 시드 10 개, 회차당 100 구간
신규 표본	공격 300 구간·정상 700 구간, 총 1,000 구간
주요 종속변수	FNR = FN/(TP+FN)×100, FPR = FP/(FP+TN)×100
보조 종속변수	정밀도 TP/(TP+FP), 재현율 TP/(TP+FN), F1 조화평균
가설 판단	FNR(5) < FNR(6)와 FPR 상승 ≤ 5.00%p 동시 충족

3.2 데이터 생성과 독립성

Python 표준 라이브러리 난수 생성기를 사용해 새로운 시드 202611101~202611110 으로 10 회 생성했다. 각 회차는 공격 30 개·정상 70 개의 5 분 구간으로 구성했다. **정답 라벨을 먼저 배치한 뒤**, 공격 구간은 실패 횟수 4~12 회(균등 정수 추출), 정상 구간은 0~6 회(균등 정수 추출)로 만들었다. 여기에 성공 로그인 1~3 회를 더하고 사건을 SSH 텍스트, 웹 인증 텍스트, JSON 형식으로 무작위 배분했다. 정답과 관측 실패 횟수는 일치하지 않을 수 있으며, 분포가 겹치는 4~6 회 영역이 의도적으로 존재한다. 사용자명과 IP 는 모두 가상이며, IP 는 RFC 5737 문서용 예약 대역을 이용했다.

새 난수 시드와 파일은 이전 T10 프로젝트와 다르다. 생성 규칙을 임계값 6 회 결과가 나오는 방향으로 사후 수정하지 않았다. 다만 **정답을 이용해 실패 횟수 분포를 생성하는 방식**이므로, 라벨과 특징이 완전히 독립인 자연 관측 자료는 아니다. 이러한 인공 설정은 실환경 적용의 중요한 제약이다.

원본 CSV(data/raw/logs.csv), 구간표(windows.csv), 정답표(labels.csv), 생성 매니페스트와 SHA-256 을 각각 보존했다. 원자료 이벤트 수는 총 **6,551 건**이다. 표본 수(구간)는 사건 수와 다르다.

3.3 평가 및 추가 분석

각 구간의 실패 횟수를 추출한 뒤 7 개 기준에 대해 TP(공격 탐지), FN(공격 누락), FP(정상 오탐), TN(정상 판정)을 집계했다. 정확한 분수로 계산하고, 인쇄시 비율은 소수점 둘째 자리, F1 은 셋째 자리로 반올림했다. 통계적 유의성이나 인과 일반화는 주장하지 않는다.

확인적 평가와 별도로, 가정한 공격 기저율 $p \in \{1\%, 5\%, 30\%\}$와 미탐/오탐 비용비 $cFN:cFP \in \{1:1, 5:1\}$에 따라 $\text{기대 비용} = p \times cFN \times FNR + (1-p) \times cFP \times FPR$ 을 계산한다. 이는 실제 측정 비용이 아닌 가정 기반 민감도 분석이다.

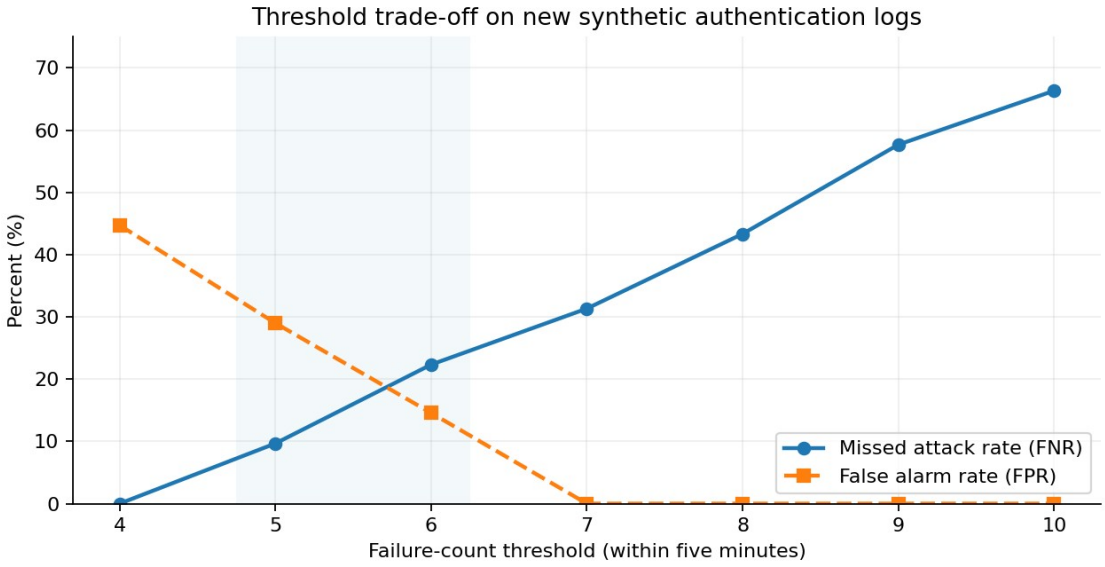
기본 실험 결과를 본 뒤 정상 실패 횟수의 상한 6 으로 인해 7 회 이상 임계값에서 FPR=0 이 된다는 문제를 인식하여, 별도 난수 시드로 **사후 탐색적 강건성 분석** 1,000 구간을 추가했다. 이 분석에서는 공격 2~12 회, 정상 0~10 회로 범위를 넓혔다. 변경 시점과 사유를 docs/03_사후탐색_강건성계획.md 에 남겼으며, 확인적 H1 재검증이라고 주장하지 않는다.

4. 결과

4.1 기본 실험: 임계값 4~10 회 비교

표 1. 1,000 구간의 누적 혼동행렬 및 성능 (비율 단위 %, F1 무단위)

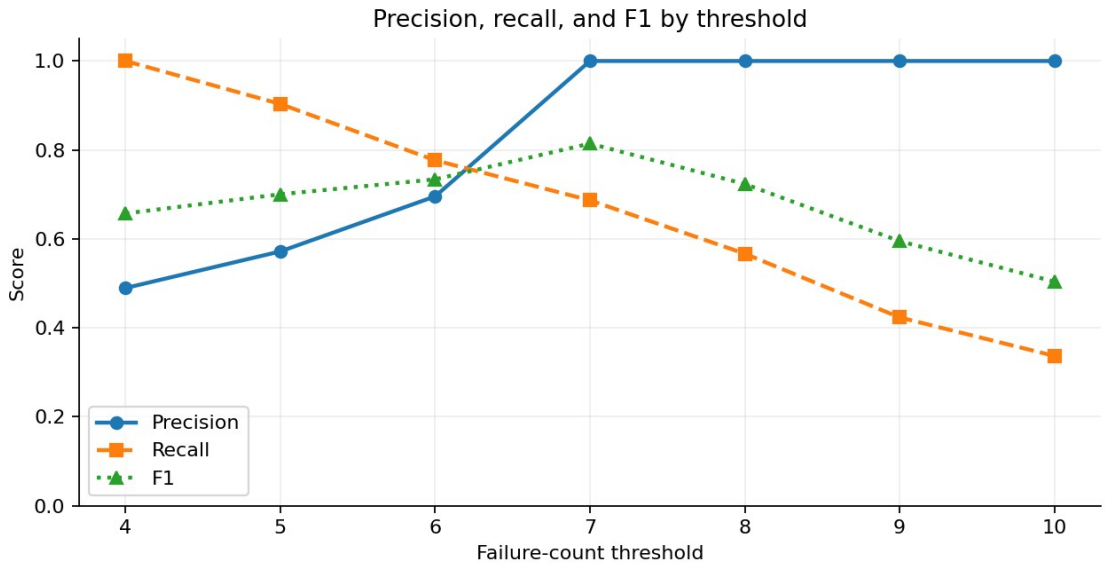
기준(회)	TP	FN	FP	TN	FNR(%)	FPR(%)	Precision(%)	F1
4	300	0	313	387	0.00	44.71	48.94	0.657
5	271	29	203	497	9.67	29.00	57.17	0.700
6	233	67	102	598	22.33	14.57	69.55	0.734
7	206	94	0	700	31.33	0.00	100.00	0.814
8	170	130	0	700	43.33	0.00	100.00	0.723
9	127	173	0	700	57.67	0.00	100.00	0.595
10	101	199	0	700	66.33	0.00	100.00	0.504



4.2 사전 가설 H1 의 판정

6 회 기준은 TP 233, FN 67, FP 102, TN 598 였고, 5 회 기준은 TP 271, FN 29, FP 203, TN 497 였다. 누락률은 22.33% → 9.67%로 **12.67%p 감소**하여 첫 번째 조건을 충족했다. 그러나 오탐률은 14.57% → 29.00%로 **14.43%p 증가**하여 사전 허용폭 5.00%p 를 넘겼다. 따라서 두 조건을 동시에 요구하는 **H1 은 지지되지 않았다.**

5 회에서 새로 감지한 공격은 38 구간이지만, 정상 오탐도 101 구간 추가되었다. 이 결과는 단순히 탐지를 강화하면 항상 좋은 운영 결과가 나온다는 주장을 뒷받침하지 않는다.



4.3 탐색적 정책 비교와 사후 강건성 분석

기본 실험에서는 7 회 기준 FPR=0%, 정밀도=100%, F1=0.814 으로 계산되었다. **이 수치가 실제 환경에서 오탐이 없음을 의미하지는 않는다.** 합성 정상 구간의 실패 횟수를 최대 6 회로 제한했기 때문에 7 회 이상 오탐이 구조적으로 발생하지 않는 데이터 생성 효과다. 공격 구간 또한 균등 분포이므로 F1 의 우열을 실제 조직에 일반화할 수 없다.

이 우려를 반영해 사후 탐색적 분포 변경을 실시했다. 정상 구간도 7 회 이상 실패할 수 있고 공격 구간도 4 회 미만일 수 있게 만든 별도 1,000 구간에서 측정한 결과는 다음과 같다.

표 2. 사후 탐색적 분포 변경 실험의 결과

기준(회)	TP	FN	FP	TN	FNR(%)	FPR(%)	F1
4	250	50	440	260	16.67	62.86	0.505
5	218	82	385	315	27.33	55.00	0.483
6	194	106	322	378	35.33	46.00	0.475
7	167	133	272	428	44.33	38.86	0.452
8	140	160	209	491	53.33	29.86	0.431
9	117	183	148	552	61.00	21.14	0.414
10	86	214	73	627	71.33	10.43	0.375

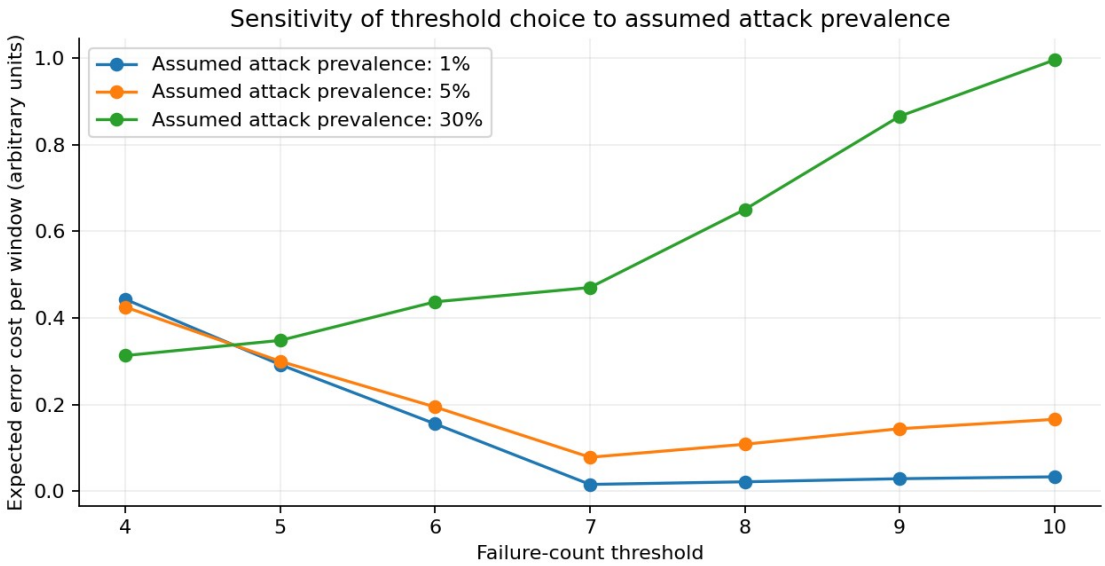
이 조건에서 7 회 기준의 FPR 은 38.86%로 0 이 아니었다. 이 분석은 정상·공격 실패 횟수의 분포 가정이 바뀌면 탐지 정책의 평가도 크게 바뀔 수 있음을 보여주는 예시이며, 원래 H1 의 독립적인 추가 검증이 아니다.

4.4 비용·발생률 가정에 따른 임계값 민감도

같은 기본 실험의 FNR 과 FPR 이라도 기저율과 오분류 비용 가정을 다르게 대입하면 최소 기대 비용을 갖는 임계값이 달라질 수 있다.

표 3. 사전 정의한 가정값에 따른 계산상 최소 비용 기준(탐색적)

가정 공격 비율	FN:FP 가정 비용비	최소 비용 기준	기대 비용/구간
1%	1:1	7 회	0.0031
1%	5:1	7 회	0.0157
5%	1:1	7 회	0.0157
5%	5:1	7 회	0.0783
30%	1:1	7 회	0.0940
30%	5:1	4 회	0.3130



이 비용표는 관측된 사고 비용이 아니라 가정값을 대입한 계산 결과다. 실제 정책 설정을 제안하는 근거로 활용하기에는 부족하다.

5. 논의

5.1 가설과 일치하는 부분, 일치하지 않는 부분

기준을 6 회에서 5 회로 낮추면 탐지 누락률이 낮아진다는 예측은 이번 데이터에서 관찰되었다. 그러나 함께 고정한 오탐 증가 5%p 이내 조건은 충족하지 못했다. 따라서 연구자는 사후에 '누락률 감소만 성공'이라고 가설을 바꾸지 않았다. 실제 H1 의 최종 판단은 **지지되지 않음**이다. 확인적 비교 외의 정책별 F1 최대값과 사후 강건성 결과는 탐색적 결과로 분리했다.

5.2 관찰 결과의 원인과 의미

실패 횟수 임계값을 낮추면 동일 자료에 대하여 경보 구간 집합이 커지므로 정상과 공격 구간이 모두 추가로 탐지될 수 있다. 기본 합성 데이터는 정상 실패가 5 회인 구간을 포함하고 있었으므로, 6→5 변경에서 정상 오탐도 크게 증가했다. 높은 기준에서는 경보량이 감소하지만 공격 누락이 증가하는 경향이 나타났다. 동시에 기본 합성 분포의 정상 최대 실패 횟수 6 이라는 인위적인 제약이 7 회 이상의 오탐률 0 을 만들었다. 실제 7 회 정책이 완벽하다는 증거는 아니다.

관련 실증 연구(Javed, 2016; Singh et al., 2024)는 실제 로그에서 인증 실패 패턴과 탐지 방법을 평가하며 운영 환경에서는 사건 분포가 더 다양할 수 있음을 보여준다. 본 결과는 학습 기반 공격 탐지나 분산 IP 공격에 대한 성능을 검증하지 않았다. 또한 MITRE 의 탐지 전략은 고정된 하나의 수치를 강제하지 않으므로, 본 실험의 값이 그 전략에 대한 우열 판단을 대신하지 않는다.

5.3 연구 한계

첫째, 데이터는 전부 합성이며 정상·공격 라벨 비율 70:30 과 각각의 실패 횟수 범위가 인위적으로 정해졌다. 특히 높은 기준에서 FPR=0 이 나온 것은 생성 규칙의 상한과 직접 관련된다. 둘째, 구간별 이벤트가 실제 네트워크 트래픽의 장기 자기상관, IP 순환, 사용자 행태 변화를 나타내지 않는다. 셋째, 본 정책은 같은 IP·5 분 범위만 검사하므로 분산 공격·저속 공격·정상 사용자 로그인 실수의 실제 분포를 다루지 못한다. 넷째, 10 회는 서로 다른 난수 시드이지만 같은 모형의 반복이어서 다른 조직·기간으로 일반화할 근거가 아니다. 다섯째, 정밀도와 비용 계산에서 실제 공격 발생률과 운영 비용을 측정하지 않았으며 민감도 분석의 1·5·30%는 **가정**이다. 여섯째, 강건성 실험은 결과를 본 뒤 설계했으므로 독립적인 확인 검증이 아니다.

6. 결론

실패 횟수 임계값을 6 회에서 5 회로 낮추면 합성 로그에서 공격 누락률이 12.67%p 낮아졌지만 오탐률은 14.43%p 증가했다. 사전에 설정한 오탐 증가폭 5%p 이내 조건을 충족하지 못했으므로 H1 은 지지되지 않았다. 4~10 회 탐색적 비교와 사후 강건성 분석에서도 어느 하나의 임계값을 실제 보안 운영의 보편적인 최적 기준이라고 결론 내릴 수 없었다. **관측 분포, 오탐 대응 부담, 공격 누락 비용을 함께 고려해야 한다**는 것이 본 연구의 제한적인 시사점이다. 실제 적용을 논하려면 정답 라벨이 검증된 독립적인 운영 로그와 여러 환경에서의 추가 검증이 필요하다.

참고문헌

1. Axelsson, S. (2000). The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security, 3*(3), 186–205. https://doi.org/10.1145/357830.357849

2. Javed, M. (2016). *Detecting Credential Compromise in Enterprise Networks* (UCB/EECS-2016-216) [Doctoral dissertation, University of California, Berkeley]. https://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-216.html
3. Khandait, P., Tiwari, N., & Hubballi, N. (2021). Who is trying to compromise your SSH server? An analysis of authentication logs and detection of bruteforce attacks. *ACM International Conference Proceeding Series*, 127–132. https://doi.org/10.1145/3427477.3429772
4. MITRE ATT&CK. (2026). *Brute Force (T1110); Detection Strategy DET0463*. https://attack.mitre.org/techniques/T1110/ ; https://attack.mitre.org/detectionstrategies/DET0463/
5. National Institute of Standards and Technology. (2025). *Digital Identity Guidelines: Authentication and Authenticator Management* (SP 800-63B-4). https://doi.org/10.6028/NIST.SP.800-63B-4
6. Singh, S. K., Gautam, S., Cartier, C., Patil, S., & Ricci, R. (2024). Where the wild things are: Brute-force SSH attacks in the wild and how to stop them. In *Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation*. https://www.flux.utah.edu/paper/singh-nsdi24

재현 방법과 검증

이 논문과 같은 ZIP 의 최상위 디렉터리에서 Python 3.10 이상 환경으로 다음 순서로 실행한다.

```
python scripts/experiment.py --root .
python scripts/stress_test.py
python -m unittest discover -s tests -v
python scripts/figures.py
```

data/raw/의 세 CSV 및 생성 매니페스트, results/의 실험 기록과 모든 파생 지표를 보존했다. README.md 에 실행 방법, SHA256SUMS.txt 에 파일 해시, docs/01_사전연구계획.md 에 본 실험 전 확정한 가설, docs/03_사후탐색_강건성계획.md 에 탐색적 추가 분석의 시점을 기록했다. 최종 ZIP 은 별도 임시 디렉터리에서 압축을 풀고 재실행해 원자료와 결과 파일을 검증한다.

부록 A. 반복 회차별 확인적 비교

각 회차는 공격 30 구간과 정상 70 구간으로 구성했다. FNR·FPR 은 회차 내 모수 기준 값이다.

회차	5 회 TP	5 회 FN	5 회 FP	5 회 FNR%	5 회 FPR%	6 회 TP	6 회 FN	6 회 FP	6 회 FNR%	6 회 FPR%
1	28	2	22	6.67	31.43	26	4	9	13.33	12.86
2	24	6	19	20.00	27.14	20	10	6	33.33	8.57
3	26	4	20	13.33	28.57	23	7	12	23.33	17.14
4	27	3	21	10.00	30.00	22	8	10	26.67	14.29
5	26	4	21	13.33	30.00	25	5	11	16.67	15.71
6	29	1	20	3.33	28.57	26	4	10	13.33	14.29
7	28	2	20	6.67	28.57	25	5	13	16.67	18.57
8	27	3	22	10.00	31.43	18	12	11	40.00	15.71
9	29	1	17	3.33	24.29	25	5	8	16.67	11.43
10	27	3	21	10.00	30.00	23	7	12	23.33	17.14

부록 B. AI 협업 및 연구 설계 변경

AI 가 담당한 일: 새 연구 가설 후보와 측정 변수의 구체화, 합성 로그 생성 코드·평가 코드·검사 코드 작성 보조, 원자료 기반 표·그래프 산출, 논문 초안 구성.

학생이 선택한 방향: 기존 정규화 비교를 유지하거나 후속 실험을 더하는 대신 C 안(임계값 자체를 비교하는 별도 연구)를 선택함. AI 가 생성한 표와 결론은 재현 코드 및 원자료와 대조했으며, 사용자가 직접 수치를 검산했다는 주장은 하지 않음.

AI 제안 및 변경 기록: 5%p 허용폭은 AI 가 제안한 임의적 연구 조건이며 학술적 권고값이 아니다. 기존 논문의 '정규화의 효과' 가설을 새 데이터 없이 재해석하지 않음. 새로운 확인적 H1 과 합성 규칙을 결과 생성 전에 문서로 고정했으며, 결과를 확인하고 실시한 강건성 분석은 **사후 탐색적**이라고 표시함.