

Site Trust Check

Design and Development of an Explainable Multi-Layer Browser Extension for Real-Time Website Trust Assessment and Fraud Detection

Interim Viva Presentation

Mirela Gerasimova Balabanova-Kehayova

BSc Applied Computing | Independent Project (ACCA6020)

[Presentation date: confirm exact date in week commencing 19 October]

AGENDA

What this presentation covers

- 1 Introduction & project brief
- 2 Aim & objectives
- 3 Literature review
- 4 Methodology
- 5 Design overview & architecture
- 6 Progress so far & early results
- 7 Project plan
- 8 Conclusion

A real-time second opinion on the sites you visit

A link. A login page. An urgent message. Most people now see something like this every week, with no easy way to check it.

What & why

A Chrome extension that assesses a website's trustworthiness as the user browses, combining URL analysis, live threat intelligence, and machine learning into one explainable score.

Chosen because phishing detection today is reactive and unexplained, a gap directly relevant to a future in cybersecurity and to real users who currently have no tool that tells them *why* a site looks risky, not just that it does.

Who benefits

- Everyday browser users, who currently rely on intuition alone
- Less tech-confident users, who benefit most from a plain-English explanation rather than a raw warning
- The researcher's own career direction toward cybersecurity and applied ML

One aim, four SMART objectives

Aim: design, develop, and evaluate an explainable multi-layer browser extension that integrates ML classification and threat-intelligence APIs to give real-time, interpretable risk scoring.

O1

Literature Review

15+ peer-reviewed sources on phishing detection, feature engineering, and explainable AI.

O2

Backend & Threat Intel

FastAPI backend integrating Google Safe Browsing, VirusTotal, and PhishTank.

O3

ML Classification

Train and benchmark Random Forest, XGBoost, and a neural network, targeting F1 above 0.90.

O4

Extension & Evaluation

Chrome extension with SHAP-informed explanations, plus a formative usability evaluation.

Prior approaches, and what each leaves unresolved

Approach	Strength	Left unresolved
Zaimi et al. (2024) Deep learning on URL features	Strong lexical-feature performance via permutation importance and SMOTE-Tomek balancing.	No live threat-intelligence coverage for newly registered domains.
Sheikhi & Kostakos (2024) PSO-optimised XGBoost	High accuracy on static benchmark datasets.	No real-time browser integration; no explainability layer.
NoPhish (2024) / Wu et al. (2023) Browser-extension implementations	Demonstrate feasible client-side, real-time detection.	Single classifier only: no threat-intel fusion, no explanation of verdicts.

This project's design combines the strengths of each approach while addressing what none resolves alone.

An iterative, evidence-driven build process

Sources

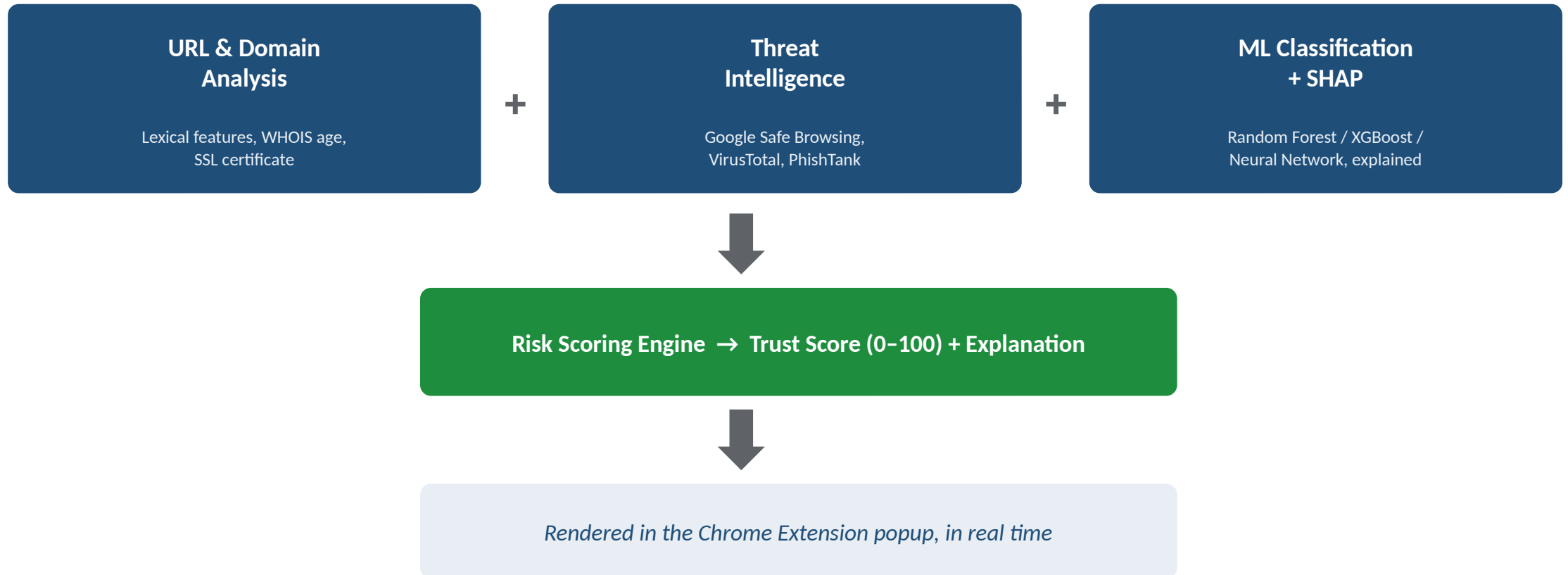
Primary: dataset built from PhishTank (phishing) and Tranco (legitimate) URLs; live testing of the trained models and API integrations.

Secondary: 15+ peer-reviewed sources on phishing detection, feature engineering, and explainable AI.

Approach

- Design, build, test, refine: each phase validated before the next begins
- Stratified 70/15/15 train/validation/test split; models compared on the untouched test set only once
- Every reported result backed by an automated test or a logged, reproducible run

A three-layer architecture, combined into one trust score

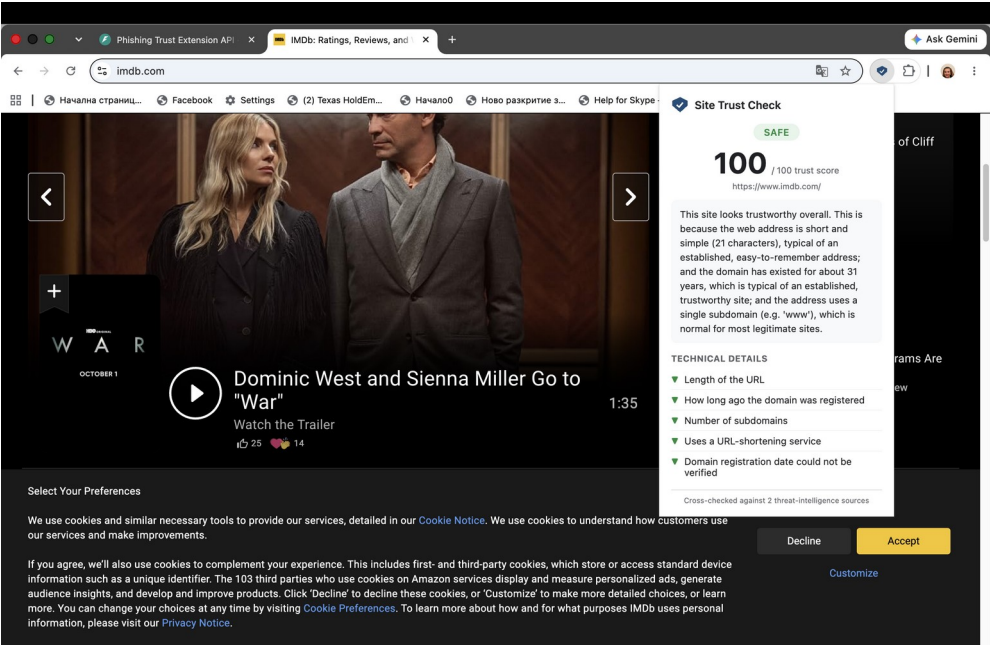


Beyond the interim milestone: a working end-to-end prototype

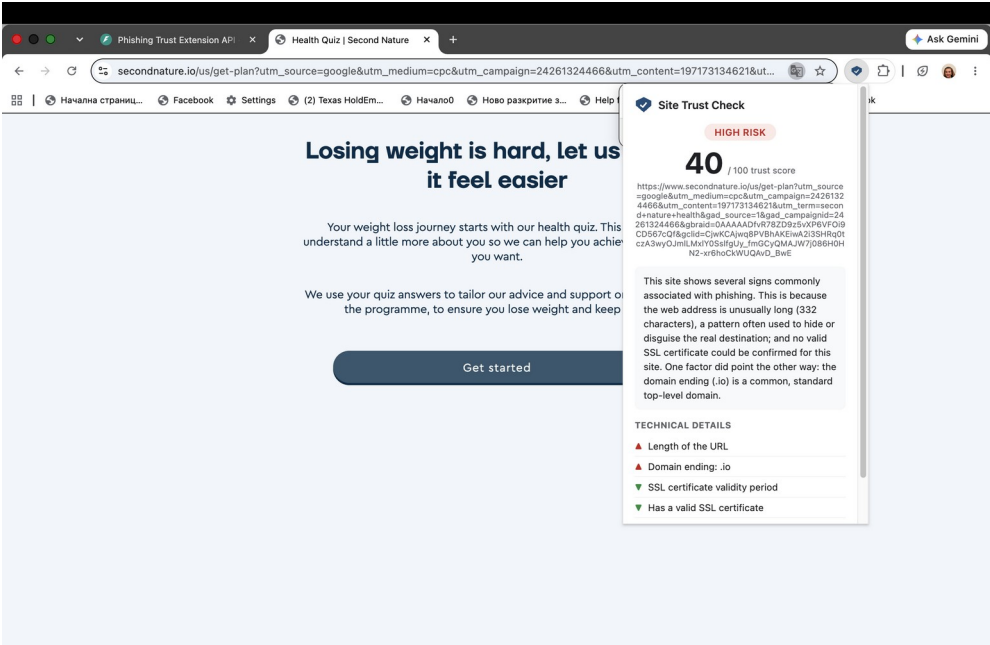
- ✓ Dataset built: 10,000 labelled URLs (PhishTank + Tranco)
- ✓ Three ML models trained & benchmarked (O3 target exceeded)
- ✓ SHAP explainability layer implemented and validated
- ✓ FastAPI backend live, with Threat Intelligence integrated
- ✓ Chrome extension (Manifest V3) functional, tested on real sites
- ✓ 51 automated unit tests passing; Docker container built

Live demonstration available now, or a short recorded walkthrough on request

Tested against real, live websites



imdb.com : correctly scored SAFE (100/100) after the retrained model



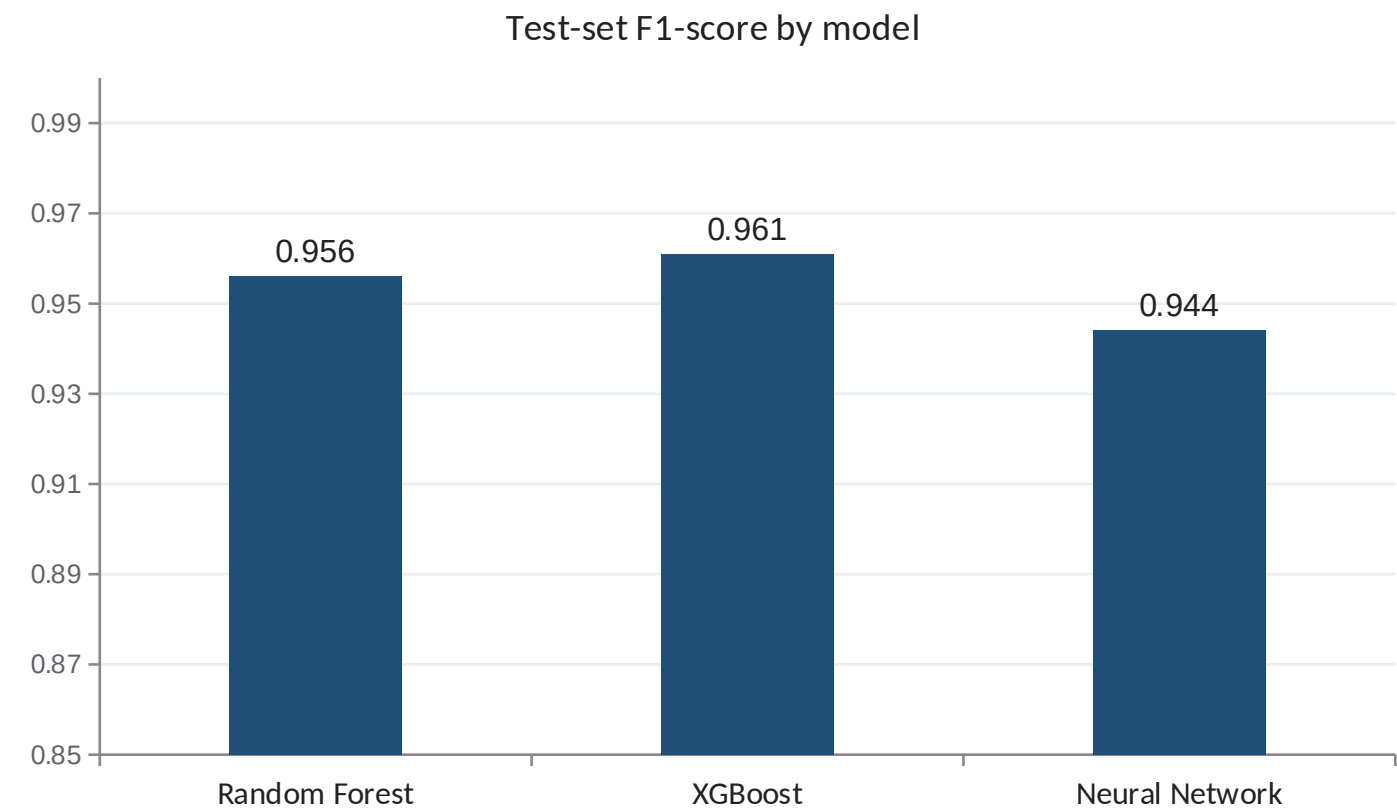
secondnature.io : flagged HIGH RISK (40/100), a legitimate site whose long marketing-tracking URL structurally resembles obfuscation

"This site shows several signs commonly associated with phishing. This is because the web address is unusually long (332 characters), a pattern often used to hide or disguise the real destination..."

- the system's own generated explanation for secondnature.io, in plain English, not a bare risk label

EARLY RESULTS

All three classifiers already exceed the O3 target (F1 > 0.90)



Final, verified numbers - measured on the held-out test set, touched only once

XGBoost

Best model

0.990

AUC-ROC

0.970

Precision

0.952

Recall

Going beyond a single test-set number

± 0.003

5-Fold Cross-Validation

F1 standard deviation across 5 folds is very tight - strong evidence the reported test-set performance reflects a genuinely stable model, not a favourable split.

2 / 10

Adversarial Robustness Test

Hand-crafted evasion URLs (typosquats, punycode, minimal-length) probed the model directly. Two clear evasion successes found and named as limitations, not hidden.

0.94 $\rightarrow$ 0.97

Baseline Comparison

Combining threat intelligence with ML initially made results WORSE, not better. Tracing why led to a real dilution bug in the scoring logic - fixed, then re-verified.

Each of these tools found something - a stability check, an honest limitation, and a real bug fixed - rather than only confirming what was already believed.

On track for a December completion



Remaining risk: usability testing requires an ethics form amendment (in progress with the module leader) before participant sessions can run.

Full week-by-week Gantt chart available in the accompanying project documentation

CONCLUSION

Summary

1

A working end-to-end prototype already exceeds interim expectations: dataset, three benchmarked ML models, SHAP explainability, live threat-intelligence, and a functional Chrome extension.

2

The design directly addresses a gap identified across the reviewed literature: no prior system unifies proactive ML detection, live threat intelligence, and user-facing explanation.

3

Evaluation goes beyond a single accuracy figure: cross-validation confirms stability, adversarial testing surfaces named limitations, and a baseline comparison caught and fixed a real design bug.

4

The remaining work (usability evaluation, final testing, and write-up) is scheduled for completion by December, with one dependency, an ethics amendment, actively being resolved.

REFERENCES

Key sources cited in this presentation

[1] Anti-Phishing Working Group, "Phishing Activity Trends Report – Full Year 2023," APWG, 2024.

[2] Oest, A., Safaei, Y., Doupe, A., Ahn, G. J., Wardman, B. & Tyers, K. "Inside a Phisher's Mind: Understanding the Anti-Phishing Ecosystem Through Phishing Kit Analysis," APWG eCrime, 2020.

[3] Hannousse, A. & Yahiouche, S. "Towards Benchmark Datasets for Machine Learning Based Website Phishing Detection," Engineering Applications of Artificial Intelligence, vol. 101, 2021.

[4] Vrbancic, G., Fister Jr., I. & Podgorelec, V. "Datasets for Phishing Websites Detection," Data in Brief, vol. 33, 2020.

[5] Zaimi, R., Hafidi, M. & Mahnane, L. "A Deep Learning Mechanism to Detect Phishing URLs Using the Permutation Importance Method and SMOTE-Tomek Link," The Journal of Supercomputing, vol. 80, 2024.

[6] Sheikhi, S. & Kostakos, P. "Safeguarding Cyberspace: Enhancing Malicious Website Detection with PSO-Optimized XGBoost and Firefly-Based Feature Selection," Computers & Security, vol. 142, 2024.

[7] Lundberg, S. M. & Lee, S.-I. "A Unified Approach to Interpreting Model Predictions," NeurIPS, vol. 30, 2017.

Q & A

Thank you for your time.