**Stage‑1 Model & Pipeline Document**

(CSE Phishing Domain Detection)

Version: 1.2

Date: October 03, 2025

Prepared by: Aditya & Ankit

# 1) Compliance Summary

* On‑prem only: All components run locally. No calls to third‑party threat‑intel / phishing‑detection APIs.
* Open‑source, non‑Chinese models in the recommended set.
* 90‑day monitoring: Recrawl scheduler + Isolation Forest for drift.
* Evidence & auditability: Store HTML, headers, redirects, screenshot, OCR text, model scores, SHAP features, and decision trail.

# 2) Model Inventory (Minimal, Stage‑1 Ready)

Classical baselines: Logistic Regression (LR), Random Forest (RF), XGBoost (XGB)

Neural baselines: mBERT / XLM‑RoBERTa (text), tiny CNN (optional visual)

Ensemble: Weighted soft vote (calibrated probabilities)

## Recommended (all open‑source, non‑Chinese):

* fastText LID (lid.176.bin) — language ID (Hindi/English/Punjabi/code‑mix)
* mBERT or XLM‑RoBERTa‑base — text phishing intent & brand mention classifier
* Tesseract OCR — extract text from screenshots/HTML
* OpenCLIP ViT‑L/14 — brand/logo similarity scoring
* DINOv2‑base — layout/template resemblance (screenshot embeddings)
* XGBoost — tabular classifier on Annexure features (URL/DNS/WHOIS/IP/TLS/HTTP/page structure)
* pHash / SSIM — fast visual clone filter
* Isolation Forest — trend/drift detection over 90‑day monitoring
* (Optional) ResNet‑18 (tiny CNN) — quick “looks‑like‑login” cue on downscaled screenshots

# 3) Annexure Feature → Model Map

|  |  |  |  |
| --- | --- | --- | --- |
| Feature Group | Concrete Features | Model / Algorithm | Output |
| URL & Lexical | length, entropy, token n‑grams, TLD, subdomain depth, '@'/'-'/'//' counts, punycode flag | Via Internet Crawler | Prob(phish)\_url |
| IDN/Homoglyph | xn-- punycode decode, Unicode confusables ratio | Via Internet Crawler | Boosted URL score |
| DNS | A/AAAA/MX/NS count, TTL stats, parked/CDN hints | Via Internet Crawler | Infra risk score |
| WHOIS/Domain Age | registrar, privacy proxy, creation/expiry deltas | Via Internet Crawler | Age/registrar risk |
| IP/ASN | ASN id, org, geo, residential/DC hint | Via Internet Crawler | Hosting risk |
| TLS/Cert | issuer, SAN count, validity, host mismatch | Via Internet Crawler | Cert risk |
| HTTP/Page Structure | redirect chain length, http→https, <form> count, password/email inputs, onsubmit JS | XGBoost (+ regex flags) | UI heuristic risk |
| OCR Text | Tesseract text from page/screenshot | mBERT/XLM‑R | Prob(phish)\_text |
| Language Mismatch | Declared vs fastText LID on OCR/HTML | Rules + mBERT/XLM‑R | Mismatch flag/score |
| Brand/Logo | favicon pHash, logo crops vs CSE allowlist | OpenCLIP + pHash | Brand similarity score |
| Layout Similarity | screenshot embeddings cosine | DINOv2‑base | Template resemblance |
| Visual (optional) | downscaled screenshot | ResNet‑18 | Prob(phish)\_vis |
| 90‑day Monitoring | feature deltas over time | Isolation Forest | Drift/novelty alerts |

# 4) End‑to‑End Pipeline (Scan → Decision → Monitor)

1. Collect — Fetch HTML/headers/redirects/TLS, resolve DNS/WHOIS/IP/ASN, render screenshot + favicon; persist raw artifacts.
2. Fast Triage (cheap) — pHash/SSIM clone check; OpenCLIP logo similarity to CSE allowlist (top‑k); DINOv2 layout similarity.
3. OCR + NLP — Tesseract text → fastText LID tags → mBERT/XLM‑R for phishing intent & brand mentions (code‑mix aware).
4. Feature Classifier (Tabular) — Build Annexure feature vector; XGBoost outputs Prob(phish)\_urltab.
5. Ensemble (Calibrated Soft Vote) — Calibrate heads (isotonic/Platt); P\_final = w1·Prob\_urltab + w2·Prob\_text + w3·Prob\_vis + α·BrandSim + β·LayoutSim; thresholds define phishing/benign/suspected.
6. Decision & Evidence Pack — Store final label + per‑head probs + top XGB SHAP features; attach HTML/headers/redirects/screenshot/OCR/logo/layout; emit JSON report + PDF/HTML evidence.
7. Monitoring (≥90 days) — Scheduled recrawl; Isolation Forest flags drift; maintain timeline and auto‑promote/demote on evidence.

# 5) Detailed Flows (What each component does)

## 5.1 URL/DNS/WHOIS/IP/TLS/HTTP (Tabular)

* Parse URL tokens, entropy, separators, subdomain depth, punycode bits.
* DNS counts (A/AAAA/MX/NS), TTL mean/variance; WHOIS age, registrar, privacy.
* IP/ASN org & geo; TLS issuer/SAN/validity/mismatch.
* HTTP/page structure: redirects, forms, credential inputs, inline JS flags.
* Model: XGBoost (RF/LR as baselines). Output: Prob(phish)\_urltab + SHAP explanations.

## 5.2 OCR + Text (Multilingual)

* Tesseract OCR on screenshot (and/or HTML visible text).
* fastText LID: Hindi / English / Punjabi / mixed.
* mBERT/XLM‑R classifier: phishing intent, brand mentions (“खाता अपडेट करें”, “verify your account”).
* Outputs: Prob(phish)\_text, language mismatch bit, top cues.

## 5.3 Visual (Logos & Layouts)

* Favicon pHash distance vs allowlist.
* OpenCLIP ViT‑L/14: logo/mark similarity to CSE references.
* DINOv2: full‑page screenshot embeddings → cosine against legit templates.
* Optional ResNet‑18 on downscaled screenshot for “login‑like” cue.
* Outputs: BrandSim [0..1], LayoutSim [0..1], Prob(phish)\_vis (if CNN used).

## 5.4 Ensemble & Thresholding

* Calibrate each head’s probability (isotonic/Platt).
* Weighted soft vote; tune weights to maximize TPR @ target FPR.
* Tiers: Phishing (≥ τ\_high), Benign (≤ τ\_low), Suspected (between).

## 5.5 Reporting & Evidence

* JSON fields: url, scan\_time, label, P\_final, head\_probs, brand/layout scores, top\_features, language\_tags.
* Evidence bundle: screenshot, OCR excerpt, redirect chain, DNS/WHOIS/TLS snapshots.
* Human narrative: short rationale (top features + cues).

## 5.6 Monitoring & Drift

* Re‑scan cadence: e.g., every 3–7 days for ≥ 90 days.
* Isolation Forest on temporal feature vectors to surface novelty.
* Auto‑update status & notify on meaningful deltas.

# 6) Deployment & Resources

## Services (Docker/K8s)

* crawler (HTML/headers/TLS/screenshot)
* dns‑whois resolver
* ocr‑nlp (Tesseract + mBERT/XLM‑R; quantized on GPU or CPU)
* visual (OpenCLIP + DINOv2; optional tiny CNN)
* xgb‑classifier (tabular)
* ensemble‑api (REST/JSON) + report‑builder
* scheduler (recrawl + monitoring)
* storage (PostgreSQL for metadata, object store for evidence)

## Compute

* CPU‑only: feasible (skip CNN; keep CLIP/DINO on CPU if needed).
* GPU (recommended): 1× 16–24 GB for mBERT/XLM‑R + CLIP/DINO (4‑bit optional).

## Security/Privacy

* No third‑party TI/phishing APIs.
* PII minimization in logs; redact credentials.
* Signed evidence artifacts, immutable bucket for chain‑of‑custody.

# 7) Evaluation & Tuning

* Splits: Stratified 5‑fold CV; time‑based dev/test if possible.
* Metrics: ROC‑AUC, PR‑AUC; report TPR at fixed FPR (e.g., 1%).
* Calibration: Isotonic/Platt on dev.
* Weights/Thresholds: Grid‑search for ensemble weights (w1..w3, α, β) and τ\_low/τ\_high to maximize scorecard.
* Ablations: (1) URL‑only, (2) add DNS/WHOIS, (3) + OCR/NLP, (4) + visual, (5) + monitoring.

# 8) API & Logging (Reference)

POST /scan — Request

{  
 "url": "https://example.tld/path"  
}

POST /scan — Response

{  
 "url": "https://example.tld/path",  
 "label": "phishing|benign|suspected",  
 "P\_final": 0.86,  
 "heads": { "urltab": 0.83, "text": 0.77, "vis": 0.41 },  
 "brandSim": 0.92,  
 "layoutSim": 0.78,  
 "language": ["hi","en"],  
 "top\_features": [["whois\_age\_days", -0.42], ["subdomain\_depth", 0.37], ["punycode", 0.31]],  
 "evidence": {  
 "screenshot\_id": "scr\_...",  
 "ocr\_excerpt": "…खाता अपडेट करें…",  
 "redirect\_chain": ["http://…","https://…"]  
 }  
}

Log schema

* scan\_id, url, ts, label, P\_final
* head\_probs {urltab,text,vis}
* brandSim, layoutSim, language[], mismatch
* top\_features[] (SHAP)
* evidence\_ids {screenshot, html, headers, dns, whois, tls}

# 9) Implementation Checklist (Ready‑to‑Ship)

* Feature extractors for URL/DNS/WHOIS/IP/TLS/HTTP
* Tesseract OCR integration + text cleaning
* fastText LID + mBERT/XLM‑R fine‑tune (multilingual, code‑mix)
* OpenCLIP + DINOv2 embedding services + reference galleries
* XGBoost trainer + SHAP exporter
* Probability calibration (isotonic/Platt)
* Ensemble weights/threshold grid‑search
* Evidence bundler (HTML/headers/redirects/screenshot/OCR)
* 90‑day scheduler + Isolation Forest
* REST API + logging + metrics dashboards

# 10) Appendix — Hyperparameters (Starters)

* XGBoost: max\_depth=6–8, n\_estimators=600–1200, learning\_rate=0.05, subsample=0.8, colsample\_bytree=0.8, scale\_pos\_weight≈(N\_neg/N\_pos).
* mBERT/XLM‑R: lr=2e‑5, batch=16–32, max\_len=256–512, epochs=3–5; focal loss if skewed.
* OpenCLIP/DINOv2: cosine retrieval; keep top‑k=5; min\_conf=0.6 to contribute to ensemble.
* Tesseract: psm=6/11; whitelist alnum + a few punctuation; run bidi for Devanagari.
* Calibration: isotonic on dev; reserve clean hold‑out for final metrics.
* Ensemble thresholds: start τ\_low=0.30, τ\_high=0.70; tune per ROC/PR curves.