**Comprehensive Phishing Detection Feature Architecture**

**Feature Categories with AI/Non-AI Implementation Guidelines**

**Decision Framework:**

* **AI (LLM/VLM)**: Use for semantic understanding, pattern recognition in unstructured data, contextual analysis, and complex behavioral patterns
* **NOT AI (Classical/Deterministic)**: Use for counting, parsing, mathematical calculations, lookups, regex matching, and fingerprinting

**1. LEXICAL & URL FEATURES**

**Basic Measurements (NOT AI)**

* **URL length (total characters)** - Simple counting
* **Domain length** - Simple counting
* **Path length** - Simple counting
* **Query string length** - Simple counting
* **Number of subdomains** - Simple parsing
* **Number of parameters in query string** - Simple parsing
* **Character counts** (digits/uppercase/lowercase/special chars/dots/slashes/hyphens/underscores) - Simple counting
* **Shannon entropy** (domain/subdomain/path/query) - Mathematical calculation
* **Character n-grams** (2-3 char sequences) - Statistical frequency analysis

**Similarity & Pattern Analysis (NOT AI)**

* **Homoglyph usage** (Unicode confusables, punycode) - Unicode normalization + lookup tables
* **Levenshtein/Edit distance** from target domain - Algorithmic string distance
* **Jaccard similarity** with target domain tokens - Set-based similarity metric
* **IDN skeleton distance** - Unicode normalization
* **Token similarity to target CSE** (n-gram overlap) - Statistical comparison

**Semantic Analysis (AI)**

* **Number of dictionary words** in URL - NLP tokenization + dictionary lookup
* **Number of random character sequences** - Pattern recognition for randomness detection
* **Suspicious substring detection** - Semantic understanding + context-aware detection
* **Keyword overlap with target CSE brand** - Semantic similarity + brand recognition

**Tools:** urllib.parse, unicodedata, textdistance, custom homoglyph detector

**2. DOMAIN-BASED FEATURES**

**Registration Data (NOT AI)**

* **TLD analysis** (popular vs cheap/abused .xyz, .top, .pw, .cn) - Lookup tables + reputation lists
* **Subdomain depth and longest subdomain length** - Simple parsing/counting
* **Hyphen/dot count in domain** - Simple counting
* **Domain age** (days since registration) - Date arithmetic
* **Domain expiration time left** - Date arithmetic
* **Privacy/proxy registration usage** - Pattern matching
* **Brand-in-subdomain detection** (e.g., sbi.login.example) - String parsing

**Advanced Analysis (AI)**

* **Registrar reputation scoring** - Pattern analysis across registrar behaviors
* **Registrant email domain classification** (public vs corporate) - Classification of email domain patterns
* **Registration pattern analysis** - Temporal pattern recognition across registrations

**Tools:** python-whois, RDAP queries, custom TLD reputation lists

**3. PATH & QUERY FEATURES**

**Basic Analysis (NOT AI)**

* **Path depth** (number of / segments) - Simple parsing
* **File extension analysis** (.php, .html, .exe, .zip) - Regex pattern matching
* **Encoded parameters** (Base64, hex, percent-encoding) - Pattern matching + decoding
* **Query parameter count** - Simple counting
* **Percent-encoding ratio** - Mathematical calculation
* **Long query string detection** - Length thresholds

**Semantic Analysis (AI)**

* **Path entropy** (randomness detection) - Pattern recognition for algorithmic vs human-generated paths
* **Sensitive keywords in path** ("login", "auth", "verify") - Contextual understanding + semantic analysis
* **Query parameter name analysis** - Semantic analysis of parameter meanings
* **Query value entropy and randomness** - Pattern detection in query values

**Tools:** urllib.parse, custom regex patterns, encoding detection logic

**4. HTML & DOM FEATURES**

**Structure Analysis (NOT AI)**

* **Element counts** (forms/input fields/password fields/OTP fields/iframes/scripts) - DOM counting
* **Suspicious JavaScript keywords** ("eval", "document.write", "escape", "atob") - Keyword matching
* **Script-to-HTML size ratio** - Size calculation
* **Cross-domain form submission detection** - URL parsing + domain comparison
* **Meta refresh and JS redirection detection** - Pattern matching
* **Cloaking hints** (User-Agent/IP conditional content) - Pattern matching

**Content Analysis (AI)**

* **Redirection analysis** - Contextual analysis of redirect purposes
* **Language mismatch detection** - Natural language detection + comparison
* **Content semantic analysis** - Understanding of phishing intent in text

**Tools:** Playwright + lxml/BeautifulSoup4, language detection libraries

**5. VISUAL & IMAGE-BASED FEATURES**

**Image Processing (NOT AI - Classical CV)**

* **Favicon presence and hash similarity** - Perceptual hashing algorithms
* **Favicon color histogram similarity** - Color histogram comparison
* **Screenshot perceptual hash** (pHash, SSIM) - Perceptual hashing algorithms
* **Logo detection/matching** (ORB/SIFT/template matching) - Classical computer vision
* **Layout similarity** (DOM tree hashing) - Structural comparison
* **Image metadata analysis** (EXIF tags) - Metadata extraction + parsing

**Advanced Visual Analysis (AI - VLM)**

* **OCR text extraction and analysis** - OCR + content analysis
* **Visual phishing intent detection** - Computer vision for understanding visual deception
* **Logo similarity with semantic understanding** - AI-powered brand recognition
* **Overall visual impersonation detection** - When reference gallery is limited

**Tools:** Playwright, imagehash, OpenCV, pytesseract, VLM APIs

**6. SSL/TLS & PKI FEATURES (NOT AI)**

**Certificate Analysis**

* **HTTPS presence** - Boolean check
* **Certificate details** (issuer, subject CN, SAN count) - Certificate field extraction
* **Validity period and time since issuance** - Date calculations
* **Self-signed certificate detection** - Certificate chain validation
* **Free CA usage** (Let's Encrypt, ZeroSSL) - Issuer lookup
* **Wildcard certificate presence** - Certificate pattern matching
* **Domain-certificate CN mismatch** - String comparison
* **Certificate Transparency signals** - Log parsing + matching

**Tools:** pyOpenSSL, ssl module, ZGrab2, CertStream

**7. WHOIS & REGISTRATION FEATURES**

**Data Extraction (NOT AI)**

* **Registrar, registrant details** (name/organization/country) - Data extraction
* **Privacy-protected registration usage** - Pattern matching
* **Registration dates** (creation/expiration/last updated) - Date extraction

**Pattern Analysis (AI)**

* **Registrant email domain classification** - Classification of email patterns
* **Registration temporal patterns** - Pattern recognition across multiple registrations
* **Registrant behavior analysis** - Behavioral pattern analysis

**Tools:** RDAP HTTP queries, python-whois

**8. DNS & NETWORK FEATURES**

**Basic Network Data (NOT AI)**

* **IP address, ASN, hosting ISP/country** - Database lookups
* **Number of resolved IPs** - DNS query counting
* **MX records and provider reputation** - Reputation lookup
* **Reverse DNS lookup, CNAME chains** - DNS resolution
* **TTL values** - DNS record analysis
* **Historical IP count/churn** - Historical tracking

**Advanced Analysis (AI)**

* **Geolocation mismatch detection** - Contextual analysis of claimed vs actual location
* **Fast-flux detection** - Temporal pattern analysis of DNS changes
* **ASN reputation analysis** - Historical behavior pattern analysis

**Tools:** dnspython, MaxMind GeoLite2, ipwhois

**9. HOSTING & INFRASTRUCTURE FEATURES (NOT AI)**

**Service Detection**

* **Cloud hosting detection** (AWS, GCP, Azure, DigitalOcean) - IP range/ASN lookup
* **Platform detection** (Ngrok, Vercel, Render, Netlify, Cloudflare Pages, GitHub Pages) - Header/CNAME fingerprinting
* **Free platform hosting detection** - Service fingerprinting
* **Shared vs dedicated hosting indicators** - Infrastructure fingerprinting
* **Bulletproof/risky ASN detection** - Reputation list matching

**Tools:** Custom fingerprinting via headers/CNAME analysis, service fingerprint database

**10. BEHAVIORAL & TEMPORAL FEATURES**

**Time-based Tracking (NOT AI)**

* **First/last-seen timestamps** - Timestamp tracking
* **Uptime monitoring** - Availability tracking
* **Daily change rate** - Change frequency calculation

**Pattern Analysis (AI)**

* **State transitions** (parked-to-active) - State change pattern recognition
* **DOM/screenshot change significance** - Change detection + significance analysis
* **Behavioral drift analysis** - Pattern analysis over time
* **Traffic pattern analysis** - Behavioral analysis of network patterns

**Tools:** CertStream, custom scheduler, change detection systems

**11. NETWORK TRAFFIC FEATURES**

**Traffic Metrics (NOT AI)**

* **HTTP response code distribution** - Statistical analysis
* **Page load time and request count** - Performance metrics
* **Resource types requested** - Content-type classification
* **Redirect hop patterns** - Pattern tracking

**Traffic Analysis (AI)**

* **External domain relationships** - Relationship analysis between domains
* **Suspicious endpoint detection** - Anomaly detection in network patterns
* **Traffic fingerprinting** - Advanced pattern analysis

**Tools:** Playwright with HAR capture

**12. OSINT & EXTERNAL FEATURES**

**Data Collection (NOT AI)**

* **Certificate Transparency mentions** - Log parsing + matching
* **Passive DNS dataset appearances** - Dataset querying
* **Blocklist/security feed mentions** - Feed parsing + matching
* **Traffic rank/popularity data** - Reputation lookup

**Content Analysis (AI)**

* **Social media mentions analysis** - NLP for context understanding and threat classification
* **Paste/code-sharing site analysis** - Content analysis for credential theft indicators
* **Narrative summary generation** - Human-readable rationale from features

**Tools:** CertStream, social media APIs, paste site scrapers

**Implementation Priority Framework**

**High Priority (Core Detection)**

1. **Lexical features** (brand similarity, homoglyphs)
2. **Domain analysis** (age, registrar, TLD)
3. **Visual similarity** (favicon, screenshot hashing)
4. **SSL/Certificate analysis**
5. **Basic behavioral tracking**

**Medium Priority (Enhanced Detection)**

1. **HTML/DOM analysis**
2. **DNS/Network features**
3. **Infrastructure fingerprinting**
4. **Advanced temporal analysis**

**Low Priority (Intelligence Enhancement)**

1. **OSINT integration**
2. **Social media monitoring**
3. **Advanced traffic analysis**
4. **Narrative generation**

**Key Design Principles**

1. **Prefer deterministic over AI** when accuracy and speed are equivalent
2. **Use AI for semantic understanding** and complex pattern recognition
3. **Implement proper feature engineering** before applying ML models
4. **Maintain explainable features** for analyst review
5. **Design for real-time operation** with appropriate caching
6. **Build comprehensive ground truth** datasets for training and validation

**Cost-Benefit Analysis**

**Classical Methods:**

* ✅ Fast, cheap, deterministic, easily debuggable
* ✅ No model drift, consistent results
* ❌ Limited semantic understanding

**AI Methods:**

* ✅ Superior semantic understanding, adaptive to new patterns
* ✅ Can handle obfuscation and novel attack vectors
* ❌ Higher computational cost, potential for false positives
* ❌ Requires ongoing model maintenance and retraining