**For the AI-based phishing domain detection work, the features and data to extract from URLs and domains can be categorized as follows**

**1. Lexical Features (String-based on URL text)**

* **URL length** (total characters)
* **Domain length**
* **Path length**
* **Query string length**
* **Number of subdomains**
* **Number of parameters in query string**
* **Number of digits**
* **Presence of repeated digits**
* **Number of uppercase vs lowercase letters**
* **Number of words (dictionary tokens) in URL**
* **Number of random character sequences** (non-dictionary tokens)
* **Character n-grams** (frequency of 2-3 char sequences)
* **Shannon entropy** of:
  + Domain string
  + Subdomain string
  + Path string
  + Query string
* **Presence of suspicious substrings**
  + “login”, “secure”, “update”, “verify”, “account”, “bank”, “admin”
* **Keyword overlap with target brand name**
* **Homoglyph usage** (Unicode confusables, punycode)
* **Levenshtein/Edit distance from target domain**
* **Jaccard similarity with target domain tokens**

**2. Domain-Based Features**

* **TLD (Top Level Domain)**
  + Popular TLD (.com, .org) vs cheap/abused (.xyz, .top, .pw, .cn, etc.)
* **Subdomain depth** (levels)
* **Longest subdomain length**
* **Hyphen count in domain**
* **Dot count in domain**
* **Presence of suspicious TLDs** (new gTLDs often abused)
* **Age of domain** (days since registration)
* **Domain expiration time left** (short-lived vs long-lived)
* **Registrar reputation score** (benign vs risky registrars)

**3. Path & Query Features**

* **Path depth** (number of / segments)
* **Path entropy** (random-looking folder names)
* **File extension in path** (.php, .html, .exe, .zip)
* **Presence of sensitive keywords in path** (“login”, “auth”, “verify”)
* **Presence of encoded parameters** (Base64, hex, percent-encoding)
* **Query parameter count**
* **Suspicious parameter names** (id=, session=, token=, password=, otp=)
* **Query value entropy** (random tokens often used for tracking/malware)
* **Presence of long query strings** (>200 chars)

**4. HTML & DOM Features**

* **Number of forms** (<form> tags)
* **Number of input fields** (<input>)
* **Presence of password input field** (type=password)
* **Presence of OTP/2FA input field**
* **Number of iframes** (hidden malicious frames)
* **Number of external scripts loaded**
* **Suspicious JavaScript keywords** (“eval”, “document.write”, “escape”, “atob”)
* **Number of inline scripts**
* **Ratio of script size to HTML size** (heavy JS = suspicious)
* **Redirection tags** (<meta refresh>, JavaScript redirects)
* **Suspicious HTML keywords** (“bank”, “verify”, “update”, “otp”, “password”)
* **Language mismatch** (declared language vs content)
* **Forms submission destination mismatch** (submits to different domain)

**5. Visual & Image-Based Features**

* **Favicon presence** (True/False)
* **Favicon hash similarity** with known CSE favicon
* **Favicon color histogram similarity**
* **Number of images on page**
* **OCR extracted text from images** (check for brand/keywords)
* **Logo similarity detection** (compare extracted logo with genuine logo)
* **Screenshot perceptual hash (pHash, SSIM)** compared to CSE reference site
* **Page layout similarity** (structural DOM tree comparison)
* **Image metadata analysis (EXIF tags)**

**6. SSL/TLS Features**

* **HTTPS presence**
* **Certificate issuer organization**
* **Certificate subject CN**
* **SAN (Subject Alternative Names) count**
* **Validity days of certificate**
* **Time since issuance**
* **Self-signed certificate flag**
* **Use of free CA (Let’s Encrypt, ZeroSSL, etc.)**
* **Wildcard certificate presence**
* **Mismatch between domain and certificate CN**

**7. WHOIS / Registration Features**

* **Registrar name**
* **Registrant name**
* **Registrant organization**
* **Registrant country**
* **Registrant email domain type** (public like Gmail/Yahoo vs corporate)
* **Use of privacy-protected registration** (whoisguard, proxy)
* **Creation date**
* **Expiration date**
* **Last updated date**

**8. DNS & Network Features**

* **IP address of domain**
* **ASN (Autonomous System Number)**
* **Hosting ISP name**
* **Hosting country**
* **Geolocation mismatch** (domain claims to be bank in India, hosted in Russia)
* **DNS TTL values** (short TTLs for fast-flux hosting)
* **Number of resolved IPs** (load balancing vs fast-flux)
* **MX records presence**
* **MX provider reputation**
* **Reverse DNS lookup name**
* **CNAME chains (pointing to tunneling/CDN services)**

**9. Hosting / Infrastructure Features**

* **Cloud hosting detection** (AWS, GCP, Azure, DigitalOcean, OVH)
* **Tunneling/CDN detection**
  + Ngrok
  + Vercel
  + Render
  + Netlify
  + Cloudflare Pages
  + GitHub Pages
  + Pages.dev
* **Hosting on free platforms** (Google Sites, Wix, Weebly, WordPress.com)
* **Dedicated server vs shared hosting**
* **ASN reputation** (known bad hosting ASNs)

**10. Behavioral & Temporal Features**

* **First-seen timestamp (CT log, DNS discovery)**
* **Last-seen timestamp**
* **Uptime monitoring (days active)**
* **Changes over time**
  + DOM snapshot hash changes
  + Screenshot similarity changes
* **State transition** (parked → active phishing)
* **Number of redirects before landing**
* **Presence of cloaking (different content for bots vs browsers)**
* **Traffic fingerprinting** (Alexa rank, popularity score if available)

**11. Network Traffic Features *(if active interaction allowed)***

* **HTTP response code distribution** (200 vs 3xx vs 404)
* **Page load time**
* **Number of network requests made**
* **External domains contacted**
* **Resource types requested** (JS, CSS, fonts, trackers)
* **Suspicious endpoints in requests** (IP addresses instead of domains)

**12. OSINT & External Features**

* **First mention in Certificate Transparency logs**
* **First mention in passive DNS datasets**
* **Appearance in social media posts**
* **Appearance in paste/code-sharing sites**
* **Mentions in security feeds / blocklists (if self-maintained, no external APIs)**

**Summary**

This is the **most exhaustive list possible** for phishing URL/domain detection.  
It covers **lexical → domain → path → HTML/DOM → visual → SSL → WHOIS → DNS → hosting → behavioral → traffic → OSINT**.

This way, your ML/ensemble models can consume **rich multi-modal features**:

* String-based
* Tabular metadata
* Visual features
* Temporal/graph features