## Appendix A: Complete Feature Specifications

**Table A1.** Core structural features for BiLSTM-based sequential analysis, detailing eleven URL characteristics including IP address hosting, subdomain patterns, brand impersonation indicators, and entropy measurements for comprehensive phishing detection.

| Feature | Parameter | Description |
|---|---|---|
| IP Address Host | *is_ip_host*(*host*) | Checks if host is an IP address instead of a domain name. |
| Subdomain Count | *count_subdomains*(*host*) | Number of subdomains in the URL. |
| Double Slash in Path | *has_double_slash_in_path*(*url*) | Checks for "//" in URL path. |
| TLD in Path | *has_tld_in_path*(*url*) | Checks if top-level domain appears in path. |
| Symbols in Domain | *has_symbols_in_domain*(*host*) | Special characters in the domain name. |
| Brand Mimicry with Hyphens | *domain_prefix_suffix_like_brand*(*host*) | Brand keywords with hyphens (e.g., "pay-pal"). |
| Brand in Path/Subdomain | *brand_in_path_or_subdomain*(*host*, *url*) | Brand names in path or subdomain. |
| URL Length | *url_length*(*url*) | Total character count of URL. |
| HTTPS Scheme | 1 if *scheme* == '*https*' else 0 | Whether URL uses HTTPS. |
| Digit Count | *digit_count*(*url*) | Number of digits in URL. |
| URL Entropy | *url_entropy*(*url*) | An entropy score. |

**Table A2.** An extended feature set for rule-based analysis, comprising six additional indicators, including domain age, URL shortening services, suspicious top-level domains, homoglyph detection, and threat intelligence matches from PhishTank and OpenPhish databases.

| Feature | Description |
|---|---|
| Domain Age | How long the domain has been registered. |
| Shortened URL | Uses URL shortening services. |
| Suspicious TLD | Uncommon top-level domains. |
| Homoglyph Detection | Characters that look like letters. |
| PhishTank Match | Found in PhishTank database. |
| OpenPhish Match | Found in OpenPhish feed. |

**Rule-Based Scoring System.** The framework implements a weighted scoring mechanism that categorizes features based on their discriminative power for phishing detection:

**High-Risk Features (2 points each):**

1. **IP Address Host**: Detection of IP address formats (e.g., 192.168.1.1) instead of domain names, strongly indicative of evasion techniques.
2. **Brand Mimicry with Hyphens**: Identification of brand names segmented by hyphens (e.g., "pay-pal", "face-book"), suggesting intentional impersonation.

3. **Homoglyph in Domain**: Recognition of character substitution attacks using visually similar characters (@, !, $, numeric replacements).
4. **Blacklist Matches**: Verification against established threat intelligence feeds (PhishTank, OpenPhish), with confirmed matches receiving maximum penalty scores.

**Medium-Risk Features (1 point each):**

5. **Excessive Subdomain Count**: Threshold of >2 subdomains, potentially indicating obfuscation attempts.
6. **Abnormal URL Length**: URLs exceeding 75 characters, often associated with generated or obfuscated addresses.
7. **High Digit Density**: Presence of >5 numerical digits, suggesting automated generation.
8. **Elevated URL Entropy**: Entropy values >4.0, indicating random character sequences typical of malicious domains.
9. **Suspicious Symbol Usage**: Detection of special characters beyond standard domain conventions.
10. **Brand Terms in Structural Elements**: Identification of brand keywords within paths or subdomains.
11. **Non-Standard Path Syntax**: Presence of double slashes within URL paths.
12. **TLD Replication in Paths**: Suspicious repetition of top-level domain patterns within path segments.
13. **URL Shortener Usage**: Employment of known shortening services that obscure final destinations.
14. **Suspicious TLD Associations**: Domains using recently established or reputation-challenged top-level domains.
15. **Domain Age Anomalies**: Recently registered domains (<6 months) are commonly associated with phishing campaigns.

**HTML Content Features (1-2 points):**

16. **Abnormal Link Protocols**: Detection of non-standard link types (javascript:, mailto:, data:) that may facilitate malicious actions.
17. **Suspicious Form Actions**: Identification of form submissions pointing to external domains, potentially harvesting credentials.
18. **External Anchor Dominance**: Analysis showing >50% of links directing to unrelated external domains.
19. **Meta Keyword Inconsistencies**: Mismatches between declared meta keywords and actual domain content.

**Risk Classification Schema.** The framework establishes a multi-level risk assessment protocol:

- **Low Risk**: Cumulative score <3 points
- **Medium Risk**: Cumulative score 3-6 points
- **High Risk**: Cumulative score 7-15 points

**Multi-Modal Decision Fusion.** This detection verdict integrates outputs from all analytical components through a sophisticated fusion strategy:

**BiLSTM Classification:** The deep learning component generates probability scores [safe_prob, phishing_prob], with a classification threshold set at phishing_prob > 0.5. Confidence metrics are derived from the probability differentials.

**Decision Logic**: A URL is classified as phishing through multiple convergent indicators:

- Rule-based score exceeding high-risk threshold (>7 points)
- BiLSTM phishing probability surpassing conservative confidence level (>0.7)
- Immediate blacklist verification from threat intelligence feeds
- Consensus-based ensemble voting with LLM explanatory validation