Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 3, 2025

A Knowledge Distillation-Based Triple-Branch Framework for Malicious Domain Detection via Cross-Source Feature Fusion

Pu Huang *

School of Cyberspace Security, Cyberspace Security Programme, Nanjing University of Science and Technology, Nanjing, Jiangsu, China
*** Corresponding author Email:** huangpuzaozao060215@njust.edu.cn

Abstract: With the continuous expansion of internet applications, malicious domains have become a significant entry point for cyberattacks, posing severe threats to user privacy and corporate security. In practical scenarios, the limited number and complex distribution of malicious domain samples render traditional detection methods—relying on large-scale labelled datasets—ineffective under smallsample conditions, struggling to adapt to rapidly evolving attack techniques. To address this, this paper proposes a malicious domain detection model integrating knowledge distillation with tri-branch cross-source features. Designed for small-sample learning scenarios, this model comprehensively utilises domain semantic features (BERT vectors), manually derived statistical features, and SSL certificate features. It achieves deep semantic representation through branch encoders and employs attention mechanisms for adaptive weighted fusion of multi-source information. Notably, certificate features exhibit significant discriminative power between malicious and benign domains, providing structured security clues that enhance the model's recognition of novel attack patterns. To further improve generalisation performance and deployment efficiency, this paper introduces a knowledge distillation mechanism. A pre-trained large model (RoBERTa) serves as the teacher model, guiding the student model to learn high-level semantic relationships with limited data through semantic representations and soft labels. Concurrently, Triplet Loss constraints are applied to optimise intra-class cohesion and inter-class separation within the feature space. Experimental results demonstrate that the proposed model achieves outstanding performance on balanced few-shot datasets, attaining Precision, Recall, F1, AUC, and Accuracy scores of 0.9857, 0.9988, 0.9922, 0.9968, and 0.9921 respectively, significantly outperforming conventional methods. This research validates the effectiveness and practical feasibility of a lightweight architecture integrating knowledge distillation and cross-source features for malicious domain detection with small samples, offering novel insights for rapid deployment and continuous protection in cybersecurity systems.

Keywords: Malicious Domain Detection; Knowledge Distillation; Few-shot Learning; Cross-Source Feature Fusion; Triplet Loss; Attention Mechanism.

1. Introduction

With the rapid advancement of the internet and accelerated digitalisation, cyberattack methods have become increasingly diverse and covert, with malicious domains emerging as a critical enabler in such activities. Attackers frequently register vast numbers of automatically generated or legitimately disguised domains to execute phishing attacks, disseminate malware, facilitate command-and-control (C&C) communications, and conduct data theft. Such malicious domains typically exhibit characteristics including short lifespans, high registration volumes, complex structures, and character obfuscation, posing severe threats to user privacy and the stable operation of the network ecosystem. Consequently, developing an efficient, accurate, and generalisable malicious domain detection method holds significant practical importance for strengthening cyber defence systems and establishing a trustworthy internet environment.

Traditional malicious domain detection methods primarily rely on expert knowledge and manual rules, identifying potential threats through feature design and pattern matching—such as domain length, character ratios, or inclusion of suspicious keywords. While these approaches offer low computational complexity and high interpretability, making them suitable for rapid detection scenarios, their heavy reliance on human expertise limits their effectiveness against automatically generated domains (DGA) and obfuscation-based attacks. With advances in machine learning, researchers have adopted statistical features and classification models (such as support vector machines, decision trees, and random forests) to achieve automated detection. While detection accuracy has improved, model performance remains constrained by the completeness of feature engineering and struggles to effectively capture the deep semantic characteristics of domains.

In recent years, the introduction of deep learning technologies has significantly advanced the intelligent detection of malicious domains. Models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers can automatically learn semantic representations of domains within an end-to-end framework, markedly enhancing detection capabilities against unknown threats. However, singlemodal deep learning models retain certain limitations: firstly, the inherent dimensionality constraints of domain name strings make it challenging to comprehensively capture behavioural characteristics and security attributes; secondly, malicious domains constitute only a minuscule fraction of actual network traffic, resulting in severe class imbalance across datasets [1][2]. Previous studies have indicated that this imbalance causes models to favour the majority class (benign samples) during training, thereby weakening their ability to identify malicious samples. This issue is particularly

pronounced in small-sample scenarios, where generalisation performance significantly deteriorates [3]. Furthermore, existing research has insufficiently addressed the joint modelling of cross-source security information, such as domains and digital certificates, making it difficult to capture the semantic and behavioural correlations of potential attack chains from multiple perspectives.

To address these challenges, this paper proposes a malicious domain detection model integrating knowledge distillation with tri-branch cross-source features. This model comprehensively addresses the scarcity of malicious domain samples and data imbalance in real-world scenarios by combining few-shot learning principles with large-model knowledge distillation mechanisms, thereby enhancing the model's learning capability and practical deployment value under limited labelled data. The model architecture incorporates three information branches: the domain's BERT semantic representation, manually engineered structural features, and SSL digital certificate characteristics. Within a teacher-pupil framework, the teacher model learns high-level semantic distributions of domains using RoBERTa, while the pupil model acquires latent knowledge through BERT encoding and feature distillation, enabling lightweight semantic transfer and rapid training.

During feature fusion, the model employs an attention mechanism for cross-source weighted integration, enabling adaptive prioritisation of feature dimensions contributing most significantly to malicious classification. Concurrently, Triplet Loss constraints reinforce intra-class cohesion and inter-class separation within the semantic space, thereby enhancing the model's discriminative capability and robustness. Finally, the fused features undergo nonlinear mapping and classification via a Multi-Layer Perceptron (MLP), achieving precise identification of malicious domains.

The principal innovations and contributions of this paper are as follows:

1) • **Proposing a Malicious Domain Detection Framework Integrating Knowledge Distillation with Three-Branch Cross-Source Information** combines domain semantic vectors, manual features, and certificate characteristics to achieve complementary multi-source information, significantly enhancing the detection's generalisation capability and interpretability.

2) • **Design of a Feature Fusion Module Based on Attention Mechanisms** Adaptively adjusts the weights of various feature sources, effectively enhancing the model's stability and robustness within complex feature spaces;

3) • **Introducing a joint training strategy combining knowledge distillation and triplet loss**, which preserves semantic expressiveness close to that of the teacher model while ensuring the deployability of lightweight models;

4) • **Experimental validation on small-sample datasets**, which demonstrated that the proposed method achieved outstanding performance across metrics including Precision, Recall, F1 score, and AUC. Specifically, Accuracy and AUC reached 0.9921 and 0.9967 respectively, significantly outperforming conventional approaches. This underscores the practical value of the proposed model in resource-constrained scenarios.

This paper employs a combined modelling approach integrating multi-source feature fusion, attention mechanisms, and knowledge distillation to effectively enhance the generalisation capability and few-shot adaptability of malicious domain detection. The research findings provide novel insights and technical support for constructing lightweight, intelligent, and deployable cybersecurity detection systems.

2. Related Work

This section discusses the current mainstream methods for detecting malicious domains and their respective advantages and disadvantages. The classification criteria section draws upon this literature.[4]

2.1. Expert Knowledge-Based Detection Method

This approach determines whether a domain is malicious by combining security experts' experience with existing threat intelligence. Its core methods include: directly comparing the domain under scrutiny against known malicious domain lists; applying expert-defined matching rules such as regular expressions, character patterns, and brand impersonation detection; analysing syntactic characteristics like length, special characters, and subdomain hierarchy; and identifying specific attack patterns including Fast-Flux IP rotation and Domain Generation Algorithms (DGA).

This detection method offers high interpretability and rapid processing speed, achieving high accuracy against known threats. However, its reliance on blacklists, whitelists, or predefined rules renders it ineffective against unknown threats. Furthermore, the requirement for continuous updates to domain lists and matching rules results in relatively high maintenance costs. This approach is also susceptible to circumvention by attackers through minor modifications, such as substituting similar characters.

Holz et al. [5] proposed a detection method for Fast-Flux Service Networks (FFSNs) relying on security expert experience and interpretable DNS characteristics. Through multiple DNS queries, the method extracted intuitive metrics such as the number of distinct IP addresses, the number of autonomous systems to which they belong, and the number of NS records. Based on these features, simple and interpretable decision rules were constructed to identify fast-flux domains. In experiments, this approach achieved exceptionally high detection accuracy on labelled datasets and revealed widespread fast-flux behaviour during long-term monitoring of spam domains. This work demonstrates the advantages of expert-knowledge-based methods—clear feature origins, easily understandable rules, and low implementation costs—while also indicating that such approaches require continuous rule updates to maintain effectiveness as attacker tactics evolve.

2.2. Machine Learning-based Detection Method

This approach first manually extracts features from domain names, then employs traditional machine learning models for classification to determine the classification results of the domains under test. Commonly used manually extracted features include length, vowel-to-consonant ratio, n-gram statistics, and Shannon entropy. Frequently utilised machine learning models encompass decision trees, random forests, support vector machines (SVM), as well as XGBoost and LightGBM.

Antonakakis et al. [6] proposed a method utilising manual features and traditional classification algorithms to detect

malicious domains. This approach extracts domain characteristics such as length, character distribution, n-gram frequency, and entropy values, then trains and classifies them using traditional machine learning models like random forests and naive Bayes. Experiments on real-world network data demonstrate that this feature engineering-driven machine learning approach effectively distinguishes legitimate domains from malicious ones generated by domain name generation algorithms (DGAs).

2.3. Deep Learning-based Detection Method

Deep learning constitutes a subcategory of machine learning. This paper addresses it separately from machine learning due to their fundamental divergence in data processing: deep learning autonomously learns features directly from raw domain name sequences without relying on manually extracted features.

This approach processes domain names into character sequences or subword sequences before feeding them into deep learning models to derive classification outcomes. Commonly employed models include: - CNNs, for extracting local patterns (e.g., character combinations); - RNNs/ LSTMs/ GRUs, for capturing character sequence dynamics; - Pre-trained models such as Transformers/BERT, for global context modelling. Processed data may further be combined with auxiliary data sources, such as DNS resolution logs and WHOIS information.

Woodbridge et al.[7] proposed a malicious domain detection method based on long short-term memory (LSTM) networks. This approach directly inputs domain name character sequences into a deep learning model, which autonomously learns structural and pattern features of domains without requiring manually designed features. Experimental results demonstrate that this method achieves excellent detection performance across multiple datasets, proving the advantages of deep learning in complex pattern modelling and unknown domain detection.

2.4. Hybrid-based Detection Method

In addition to the aforementioned three detection methods, this paper refers to detection approaches that combine expert knowledge with machine learning or deep learning for model construction as hybrid-based detection methods.

Typical methods include: pre-screening using expert rules; integrating manually extracted features with those automatically learned by the model; and incorporating blacklists or threat intelligence as part of the model's features.

Vinayakumar et al.[8] proposed the I-DGA-DC-Net framework, which combines deep learning with traditional machine learning for malicious domain detection. This approach utilises deep learning to automatically extract high-level features, subsequently feeding these features into traditional classifiers such as support vector machines and random forests for final judgement. Experimental results demonstrate that this hybrid architecture achieves superior accuracy across multiple datasets compared to standalone deep learning or machine learning models, combining the strengths of both approaches.

2.5. Association-based Detection Method

This method employs structural relationships between domain names, IP addresses, hosts, CNAME records, and other entities to infer and identify potentially malicious domains. Typical approaches include: utilising domain-IP bipartite graphs to infer connections via shared IP addresses; employing domain-host bipartite graphs to correlate host query behaviours; leveraging heterogeneous graphs to integrate multiple relationship types; and applying graph algorithms such as belief propagation, graph neural networks, and path inference.

Bilge et al.[9] proposed the EXPOSURE system, which employs long-term analysis of passive DNS data to uncover potentially malicious domains by leveraging associations between domain names and various entities such as IP addresses, autonomous systems, and geographical locations. The approach first constructs an association graph linking domain names to related network entities from DNS resolution records. It then infers potential connections between domains based on relational features such as shared IP addresses and resolution behaviour patterns, thereby identifying unknown domains exhibiting high association with known malicious ones. Experimental results demonstrate the method's effectiveness in detecting large numbers of previously unknown malicious domains within extensive network environments, validating the practical value of association-based reasoning in detecting unknown threats.

2.6. Discussion of Advantages and Disadvantages

Expert-knowledge-based detection methods rely on manually designed rules and empirical features, offering advantages such as high interpretability, low computational overhead, and straightforward implementation. They effectively detect malicious domains conforming to known patterns. However, such approaches exhibit poor adaptability to unknown or variant attacks, necessitate manual rule maintenance and updates, and possess limited scalability. Consequently, they are better suited as foundational detection measures or for integration with other methodologies. Holz et al.[5] demonstrated the interpretability and real-time capabilities of such approaches through systematic analysis of Fast-Flux service networks. Their research also highlighted limitations including inadequate adaptation to novel or variant attacks and high rule maintenance costs.

Machine learning-based detection methods exhibit a degree of generalisation capability for unknown domains while retaining interpretable features, enabling analysis of attack patterns behind malicious domains. However, feature design remains reliant on human expertise, entailing high feature engineering costs, and model performance remains sensitive to feature quality, thereby constraining generalisation capabilities. The domain reputation system proposed by Antonakakis et al. [6] exemplifies both the strengths and limitations of such approaches: research demonstrates that machine learning models can effectively identify malicious domains using passive DNS data and multidimensional features. Nevertheless, challenges persist, including reliance on manual feature design, high maintenance costs, and constrained generalisation capabilities.

By contrast, deep learning-based detection methods can automatically extract complex non-linear features, effectively reducing feature engineering costs, and demonstrate strong generalisation capabilities when detecting unknown and variant domains. However, such methods incur substantial training and inference overhead, exhibit significant dependence on large-scale samples, and possess weaker

model interpretability. The research work proposed by Woodbridge et al.[7] clearly exemplifies these characteristics: their deep learning model significantly outperforms traditional methods in detection accuracy and generalisation capability, yet it also suffers from 'black-box'issues, lacking interpretability and consuming substantial computational resources.

To further balance generalisation performance and interpretability, researchers have proposed hybrid detection methods. Such approaches combine the strengths of expert knowledge, machine learning, and deep learning, enhancing result stability and interpretability while maintaining high detection accuracy. However, multi-model integration increases system complexity and maintenance costs. The research by Vinayakumar et al.[8] validates this observation: whilst their approach outperforms single models in terms of accuracy and robustness, it exhibits the typical characteristics of hybrid methods, namely complex system architecture and significant challenges in module coordination and updates.

Moreover, detection methods based on associative reasoning possess unique advantages in identifying unknown malicious domain groups. Such approaches leverage correlations between multi-source data to identify potential threats, as attackers find it difficult to fully fabricate these associations, thereby demonstrating strong detection capabilities. The research by Bilge et al.[9] epitomises the characteristics of such approaches: whilst their detection framework excels at revealing relationships within large-scale malicious domain groups, it relies heavily on substantial passive DNS data and high computational resources. System deployment proves complex, and determining the maliciousness of individual domains remains challenging.

It is evident that the five detection methods each possess distinct advantages and limitations in terms of interpretability, generalisation capability, implementation complexity, and computational overhead. The expert knowledge approach emphasises rule transparency and high interpretability, yet struggles to address unknown threats; machine learning methods enhance detection accuracy and automation levels, but are constrained by feature engineering costs; deep learning methods offer strong generalisation capabilities and automatic feature extraction advantages, yet suffer from insufficient interpretability; hybrid methods balance performance and explainability, albeit with higher system complexity. Correlation reasoning excels in detecting collective threats, yet relies on multi-source data and complex computational frameworks. For diverse scenarios, integrating or applying multiple detection methods in a layered approach remains the primary direction for enhancing malicious domain detection capabilities.

3. Methodology

Through systematic analysis and synthesis of existing malicious domain detection methods, it is evident that current research exhibits multiple shortcomings. Firstly, the majority of approaches—whether based on expert rules or traditional machine learning—still concentrate their core features at the domain string level. These primarily encompass character statistical features, n-gram segments, or shallow semantic embeddings. Such approaches prove vulnerable when confronting adversaries employing techniques like obfuscation, morphological variants, or character substitution to generate adversarial domains, resulting in inadequate model generalisation and robustness. Secondly, existing research often overlooks the potential correlations between domains and external security intelligence, failing to leverage multi-source contextual features (such as digital certificates or DNS records). This limitation constrains detection performance in real-world complex scenarios.

Moreover, while recent years have seen extensive research incorporating deep learning and large-scale pre-trained language models to enhance detection accuracy, such models are often structurally complex and computationally intensive, hindering rapid deployment and online detection within practical network environments. The field currently lacks a lightweight approach that balances **detection performance with computational efficiency**, capable of meeting the practical demands of malicious domain detection for **real-time processing, transferability, and deployment cost** while maintaining accuracy.

To address the aforementioned issues, this paper proposes a **lightweight attention-enhanced malicious domain detection model integrating knowledge distillation with cross-source information**. (As shown in Figure 1)

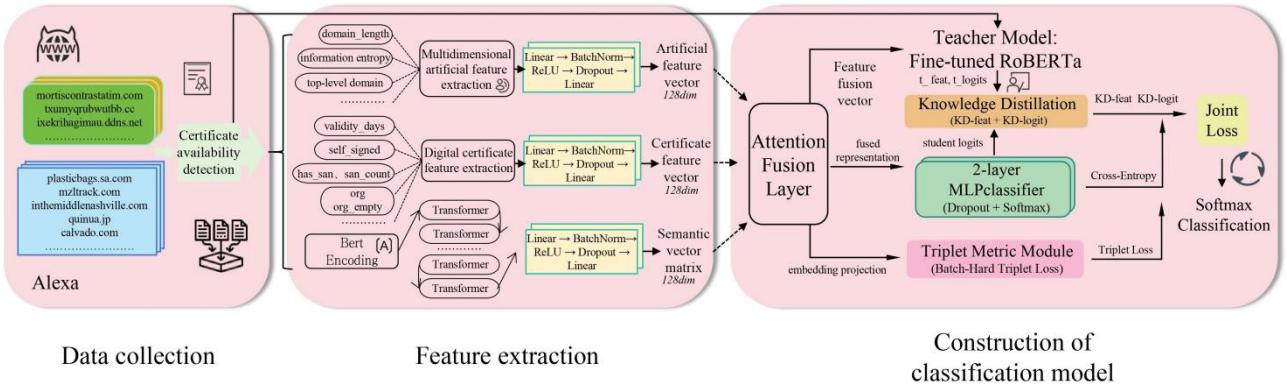


Figure 1. Model Flowchart

The design objective of this study is to achieve efficient, robust, and rapidly deployable malicious domain detection under conditions of small sample sizes and resource-constrained environments. Its core concept involves employing a knowledge distillation mechanism to transfer the semantic comprehension capabilities of large pre-trained language models (specifically RoBERTa in this paper) to simpler, computationally more efficient student models. This approach significantly reduces model complexity while maintaining high detection performance and generalisation

capabilities. Concurrently, the model incorporates a cross-source feature fusion strategy, comprehensively utilising three heterogeneous information types: domain semantic features, manually engineered structural features, and digital certificate features. This approach characterises domain security attributes across three dimensions: semantics, structure, and trustworthiness. Through an adaptive attention-weighted mechanism between features, the model dynamically captures the most discriminative feature combinations, thereby enhancing its capability to identify complex malicious domains.

In the model design, this paper draws upon recent research combining knowledge distillation with few-shot learning. Addressing the challenge of model training under resource-constrained conditions, the literature [10] proposes a knowledge distillation method for limited labelled data. By integrating few-shot learning with distillation techniques, this approach enables lightweight models to retain semantic discrimination capabilities and cross-domain generalisation performance comparable to large models, even with minimal data. This approach inspired the present work to enhance the semantic learning efficiency and generalisation capability of student models through distillation mechanisms in malicious domain detection tasks. Separately, the literature [11] integrates meta-learning with distillation learning, proposing a distilled meta-learning framework to achieve high-precision modelling of complex radar signals with fewer samples. Its core approach involves first utilising meta-learning to acquire generalised initialisations in multi-task settings, then compressing complex models into lightweight student models via knowledge distillation, thereby achieving parameter reduction while preserving performance. Building upon this, this paper introduces distillation principles to enhance the transferability and deployment efficiency of student models, making them better suited to the‘few-shot, low-computational-power’ application scenarios prevalent in practical cybersecurity detection environments.

Moreover, the metric learning module design in this paper draws inspiration from the deep metric learning approach based on Triplet Networks proposed in [12]. This research achieves metric modelling of the relationship between attack traffic and normal traffic in the feature space by integrating Multi-Scale Convolutional Neural Networks (MSCNN) with Triplet Networks in intrusion detection tasks. The employed soft-margin Triplet Loss effectively reduces distances between similar samples while distinguishing dissimilar ones, thereby enhancing the model's discrimination capability and detection accuracy. Building upon this foundation, this paper adopts a Batch-Hard Triplet Loss strategy to strengthen the separability of malicious and benign domains within the embedding space. This approach further improves the model's robustness when handling borderline samples and adversarial domains.

The overall architecture comprises the following three core modules:

1) • Data Collection and Preprocessing Module, for extracting raw domain name and digital certificate data;

2) • Feature extraction module, comprising the encoding and normalisation of BERT semantic vectors, manual features, and certificate features;

3) • The classification model construction module employs a multi-branch student network with an attention fusion mechanism, combined with feature distillation and adversarial metric learning (Triplet Loss) as joint optimisation objectives, to achieve efficient and robust malicious domain detection.

3.1. Data Collection and Processing

First, the original domain name data undergoes protocol filtering. Given that this study primarily investigates domains based on HTTPS, only samples supporting the HTTPS protocol are retained. Specifically, each domain's port 443 is sequentially checked for availability, and an attempt is made to establish a TLS/SSL connection via HTTPS. If the domain successfully completes the handshake and returns a server certificate, it is deemed to support the HTTPS protocol. If the connection times out, the port is closed, or only the HTTP protocol is supported, the domain is discarded.

Building upon this, the filtered domains undergo further certificate availability checks. This process utilises Python's ssl and socket libraries to validate connections and request corresponding digital certificates. Domains failing to establish secure connections or returning anomalous certificates (e.g., incomplete certificate chains, expired certificates, invalid issuing authorities) are deemed certificate-unavailable and excluded. This step retains only domain samples capable of successfully returning valid digital certificates, ensuring the reliability of subsequent experimental data.

Following screening, certain domains could establish TCP connections but failed the handshake due to server misconfiguration or forged certificates. Such samples exhibited uncertainty in security attributes and were therefore excluded from the final dataset. The retained domain samples all support the HTTPS protocol and possess valid digital certificates, providing the foundational data for subsequent extraction of certificaterelated features (such as issuer, validity period, and signature algorithm).

This screening step significantly enhances the dataset's quality and consistency. Firstly, it ensures each experimental sample possesses verifiable secure communication attributes. Secondly, it filters out samples failing protocol compliance or exhibiting certificate anomalies, thereby reducing noise interference and providing more stable input data for feature extraction and classification training.

3.2. Feature extraction

3.2.1. Artificial features

To characterise patterns in domain names concerning surface structure and character distribution, this paper extracts manual features from three aspects: domain composition, character statistical properties, and potential suspicious patterns. These features are categorised into numerical, binary, and categorical types. Specifically:

Numeric features

Numeric features are primarily employed to measure the statistical characteristics of domain names in terms of length and character composition:

• Overall domain length: Malicious domains often increase length to create complexity, evading detection or confusing users.

• Number of numerals and their proportion: Domains with an excessively high numeral ratio are more likely to be maliciously generated (DGA), typically lacking semantic readability.

• Number of letters and their proportion: Combined with the numeral ratio, this reflects a domain's semantic nature; legitimate domains predominantly use letters.

• Hyphen count and proportion: Malicious domains may employ excessive hyphenation to enhance obfuscation or mimic multiple subdomain combinations.

• Dot count and subdomain layers: Excessive subdomain layers may conceal malicious sites or increase detection difficulty through complex structures.

• Information entropy: Calculated using Shannon entropy, domains with higher entropy exhibit more random character distribution, typically indicating lower readability and potential for automated generation.

• Number of suspicious keywords: The presence of certain sensitive keywords is often associated with phishing attacks. This paper employs empirical judgement using the following vocabulary:‘login’,‘verify’ ,‘update’,‘secure’,‘account’,‘bank’,‘support’,‘billing’,“confirm”,‘password’,‘pay’,‘apple’,‘google’,‘microsoft’,‘amazon’,‘dhl’,“fedex”,‘paypal’.

• Maximum tag length and average tag length: Abnormally long tags or uneven distribution may indicate structural anomalies in the domain name, a common characteristic of malicious domains.

Binary features

Binary features are primarily employed to detect whether specific suspicious patterns exist within domain names:

• Whether Punycode encoding is employed: Some attackers utilise Punycode to execute internationalised domain name homograph attacks (IDN Homograph Attacks), misleading users through visually similar characters.

• Whether IP addresses are included: Using IP addresses as part of a domain name often indicates circumvention of standard registration mechanisms and may be employed to conceal true identity.

• Does it contain suspicious keywords: Keywords such as‘secure’,“account”, or‘paypal’directly suggest phishing risks.

• Does it contain common prefixes (e.g., www): Whilst legitimate domains may also use these, attackers may employ prefixes to disguise their domains.

Categorical features

Category-based characteristics are primarily reflected in the selection of top-level domains (TLDs). Significant variations exist in registration costs and regulatory oversight across different TLDs, with malicious domain registrants tending to favour low-cost, loosely regulated TLDs. Consequently, the distribution of top-level domains across categories can serve as a potential indicator for discerning malicious intent.

The extraction of these manual features effectively captures differences in the surface structure and character patterns of malicious domains. For instance, numerical proportion and information entropy reflect domain readability and randomness, aiding in the identification of automatically generated DGA domains; suspicious keyword features help uncover phishing domains targeting user sensitive information; while TLD characteristics reveal registration distribution patterns, supporting the detection of low-cost bulk registration malicious activities. By synthesising these features, effective modelling of a domain’s potential maliciousness can be achieved even in the absence of contextual or semantic information, providing a robust discriminative foundation for subsequent feature fusion and classification.

3.2.2. Bert Semantic Vector Extraction

This paper employs a pre-trained BERT model to construct semantic vector representations for domain names. The specific steps are as follows: First, the domain name is treated as an input sequence and segmented using character-level or subword-level tokenisers to obtain a subword sequence compliant with BERT’s input format. Subsequently, the sequence is fed into the BERT model. After processing through multiple layers of the Transformer encoder, a contextual semantic representation of the sequence is obtained. To facilitate subsequent feature fusion, the output vector corresponding to the [CLS] token is adopted as the semantic vector representation for the entire domain name.

Unlike manual features and certificate-based features, which primarily rely on explicit character statistics and structural information, BERT’s advantage lies in its ability to automatically learn contextual relationships between characters and subwords through large-scale unsupervised pre-training, thereby capturing deep underlying semantic patterns. This enables the model to identify characteristics of malicious domains in character arrangement, semantic disguise, and auto-generation patterns. For example:

• Character manipulation: Attackers frequently insert, replace, or duplicate characters within domain names. BERT can identify such unnatural alterations through contextual analysis.

• Semantic Disguise: Some malicious domains exploit highly similar word forms to legitimate brands (e.g., micros0ft-support.com). BERT can detect semantic discrepancies between these and genuine brand terminology.

• Generative Patterns: Domain Generation Algorithms (DGAs) tend to produce meaningless random character sequences. BERT’s contextual modelling capabilities for natural language grant it an advantage in identifying such patterns.

In this manner, BERT vectors supplement deep semantic information that manual features struggle to capture, thereby providing more comprehensive expressive capabilities for subsequent feature fusion and classification tasks.

3.2.3. Digital Certificate Feature Extraction

In traditional domain-based detection methods, models primarily rely on string structure and semantic features to identify potential malicious patterns. However, attackers frequently circumvent single-dimensional detection by disguising domain names or constructing counterfeit terms. Therefore, this paper introduces digital certificate-related features alongside manual features and BERT semantic vectors, further enhancing detection capabilities from a cross-source information perspective.

The certificate feature categories extracted in this paper correspond to manual features, specifically encompassing the following: Numeric features

• Validity Days: Calculated by examining the certificate’s not__before and not__after fields to determine its lifecycle duration. Short-term certificates are frequently issued rapidly by attackers for brief offensive operations, whereas legitimate websites typically procure certificates with extended validity periods.

• SAN count: Tally the number of domains bound to the Subject Alternative Name (SAN) field within the certificate. Certificates with excessive SAN entries may be misused for bulk registration and wildcard domain coverage, increasing potential risks.

Binary features

• CN Exact Match: Determines whether the certificate's Common Name (CN) precisely matches the domain name. If mismatched, the certificate may be misused or improperly configured.

• CN Wildcard Match: Assesses whether the CN employs wildcards and can match the domain name. Wildcard certificates serve legitimate business purposes but may also be exploited by attackers to cover multiple subdomains.

• SAN Field Presence: Records whether the certificate contains the SAN extension. Absence of this field may indicate non-compliance or anomalies.

• SAN Exact Match: Determines whether the domain name exactly matches entries within the SAN. Discrepancies may suggest the certificate was not issued specifically for that domain.

• SAN wildcard matching: Determines whether the domain is covered by a wildcard entry in the SAN, used to detect the generic characteristics of wildcard certificates.

• Support for server authentication: Checks whether the Extended Key Usage (EKU) includes server authentication. Absence indicates the certificate is not a standard HTTPS service certificate.

• Client Authentication Support: Determines whether the EKU includes client authentication, aiding supplementary scenario identification.

• Self-Signed Certificate: Checks whether the issuer and subject are identical. Self-signed certificates lack authoritative endorsement and frequently appear in untrusted or malicious contexts.

Categorical features

• Signature Algorithm Type: Extracts the certificate's signature algorithm, e.g., 'sha256WithRSAEncryption'. Weak or obsolete algorithms are more likely to appear on malicious sites.

• Organisation Name: Extracts the certificate's registered organisation details. Legitimate corporate websites typically include comprehensive organisational information, whereas attackers often employ lowvalidation-level DV certificates with missing or invalid organisation fields.

• Missing Organisation Information: Derived from the preceding field, this serves as a binary indicator for organisation information absence.

• Country Code: Records the country or region where the certificate was registered. Certificate distribution varies by region, with certain areas prone to frequent abuse.

This paper constructs a comprehensive certificate feature set across three dimensions: certificate lifecycle, domain name matching relationships, and issuance attributes. These features effectively capture anomalous states of malicious domains during certificate application, binding, and usage. When combined with manual features and BERT semantic vectors, digital certificate features supplement detection information from a crosssource security data perspective, significantly enhancing the comprehensiveness and robustness of malicious domain detection.

4. Classification Model Construction

This study addresses the application requirements for malicious domain detection in scenarios with small samples and resource constraints by designing a lightweight multi-source feature detection model that integrates knowledge distillation with metric learning mechanisms. Centred on an attention fusion layer, a multi-layer perceptron (MLP) classifier, and Triplet metric constraints, the model aims to achieve low latency and high deployability while maintaining high detection performance.

During feature fusion, the model introduces an attention mechanism to adaptively weight multi-source features. Contributions from different feature sources to classification decisions vary across samples; for instance, some malicious domains exhibit distinct character distribution patterns, while others rely on certificate or registration characteristics. Consequently, fixed weighting or simple concatenation often fails to adequately capture the nonlinear relationships between features. To address this, this paper designs a lightweight attention fusion module that dynamically weights outputs from each feature branch via a learnable weight matrix. Specifically, the model first stacks the embedding vectors from three branches, then computes attention scores through two layers of linear transformation, followed by Softmax normalisation to obtain the weight distribution. The final fused representation is obtained through weighted summation, forming the optimal feature combination for the current sample. This mechanism significantly enhances the model's ability to synergistically express heterogeneous features, enabling it to adaptively select the most discriminative information based on the input sample. Consequently, robustness in recognising complex domain names is strengthened.

During the classification phase, the fused feature representation is input into a Multi-Layer Perceptron (MLP) classifier for non-linear mapping and final discrimination. The classifier comprises two fully connected layers, each incorporating batch normalisation, ReLU activation, and Dropout operations to prevent overfitting and accelerate training convergence. The output layer is linear, with dimensions corresponding to the number of classes, yielding probability distributions for malicious versus benign domains via the Softmax function.

The primary rationale for selecting MLP as the classifier lies in its lightweight nature and efficient inference capabilities. Compared to complex architectures such as convolutional networks or Transformers, MLP features fewer parameters and a more streamlined computational graph, significantly reducing inference latency and memory consumption. As the semantic, structural, and security attributes of domains are fully abstracted during encoding, the classification stage requires no sequence-level modelling or context capture. Consequently, the MLP achieves an optimal balance for the discriminative task under low complexity constraints. This design not only fulfils the real-time and transferability requirements for malicious domain detection but also enhances the model's deployability, enabling rapid operation within security gateways or edge devices.

To further enhance the model's separability and robustness in feature space, this paper introduces a Triplet Loss metric learning module based on the Batch-Hard strategy within the joint training objective. This module selects the most challenging positive and negative sample pairs in each training batch, maximising inter-class distances while minimising intra-class distances, thereby forming compact class cluster distributions in the embedding space. Specifically, for each sample, the model calculates the farthest positive sample within its own class (hard positive) and the

nearest negative sample from the opposing class (hard negative), optimising the embedding representation through a triplet constraint function. This mechanism enables the model to maintain stable discrimination capabilities even when confronted with domains featuring blurred boundaries and diverse attack techniques, significantly enhancing its generalisation performance against adversarial samples and distribution drift scenarios.

4.1. Teacher Model

In this study, the teacher model serves to provide the student model with robust semantic comprehension capabilities and knowledge transfer support. The teacher model employs a pre-trained RoBERTa model, endowed with strong contextual understanding and semantic capture abilities. During the knowledge distillation process, knowledge from the teacher model is transferred to the student model, aiding the latter in achieving higher accuracy and generalisation capabilities under small-sample conditions. Through the distillation mechanism, the student model not only receives supervision signals from labels (i.e., standard classification loss) but also learns the teacher model's classification outputs and intermediate feature representations via KD-Logit Loss and KD-Feat Loss respectively, thereby maintaining strong discriminative capabilities for complex domains.

The teacher model provides the student model with more refined feature abstraction and higher-level semantic expression through its rich semantic understanding. Particularly when faced with a lack of abundant labelled data, knowledge distillation from the teacher model effectively compensates for the student model's training deficiencies. Specifically, the teacher model comprehensively encodes input domain name data through its deep semantic reasoning and feature abstraction capabilities. By transmitting its outputs (including logits and intermediate features) to the student model via distillation loss, it helps the latter avoid overfitting and local optima during training, thereby enhancing robustness and generalisation.

Concurrently, the teacher model undergoes training and fine-tuning on large-scale datasets, ensuring robust domain knowledge and strong discrimination capabilities. Although the teacher model itself presents challenges in computational complexity and inference speed, the knowledge distillation mechanism transfers its essential components to a lightweight student model. This ensures the final model maintains high accuracy while adapting to practical demands for low-resource, efficient inference, and real-time deployment.

The incorporation of the teacher model not only enhances the accuracy of the student model but also accelerates its training process through knowledge distillation. This enables the student model to maintain robust performance even under data scarcity conditions, further validating the effectiveness and practicality of the proposed malicious domain detection model based on knowledge distillation and multi-source feature fusion.

4.2. Training Optimisation Level

At the training optimisation level, the model employs a multi-objective joint loss function for end-to-end learning. The overall loss function is defined as follows:

$$\mathrm{L} = L_{ce} + \lambda_{kd_logit} L_{kd_logit} + \lambda_{kd_feat} L_{kd_feat} + \lambda_{triplet} L_{triplet} \quad (1)$$

Here, L_{ce} denotes the cross-entropy loss, optimising classification performance; L_{kd_logit} and L_{kd_feat} respectively represent knowledge distillation losses based on logits and feature layers, transferring semantic information and intermediate representations from the teacher model; $L_{triplet}$ is the metric learning loss, constraining structural consistency within the embedding space. Through this multi-objective joint training strategy, the student model achieves semantic discriminative capabilities approaching those of the teacher model while maintaining lightweight properties.

Through this joint optimisation strategy, the student model achieves semantic understanding capabilities and discriminative accuracy approaching those of the teacher model while maintaining a lightweight structure. The knowledge distillation mechanism enables the model to fully learn high-level semantic information even under small-sample conditions, while the Triplet metric constraints further ensure sample separability and stability within the embedding space. Ultimately, the constructed classification model achieves a unified balance of high accuracy, robustness, and low latency in malicious domain detection tasks, validating the effectiveness and practical value of the lightweight multi-source feature fusion framework in cybersecurity detection.

5. Experiments and Analysis of Results

5.1. Dataset

The dataset employed by this research partially originates from the publicly available dataset cited in this paper [13], wherein malicious domains primarily derive from the 360 Netlab Opendata Project, while benign domains stem from the Alexa Top domain list; the remainder was gathered from the internet. From these data, a total of 5,066 malicious domains supporting the HTTPS protocol were extracted, alongside several benign domains.

5.2. Experimental Environment

The experimental environment for this study is as detailed in the table below.

Table 1. Experimental Environment Configuration Table

Project	Configuration
Operating System	macOS 15.3.1 (24D70)
Hardware	Chip Apple M4, Memory 24GB
Experimental Tools	PyCharm 2025.1.3
Software Environment	Python 3.9.23, PyTorch 2.2.2, Transformers 4.57.1
Additional Dependencies	sklearn, pytorch, pandas, numpy, tqdm and other libraries

5.3. Evaluation Indicators

To comprehensively evaluate the performance of the proposed model in the task of detecting malicious domains, this paper employs accuracy, precision, recall, F1-score as evaluation metrics.

In a binary classification task, let: TP (True Positive): Malicious domains correctly identified as malicious; TN (True Negative): Benign domains correctly identified as benign; FP (False Positive): Benign domains incorrectly identified as malicious; FN (False Negative): Malicious domains incorrectly identified as benign.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Among these metrics, accuracy reflects the overall correctness of classification; precision measures the reliability of predictions for malicious domains; recall indicates the coverage rate of detected malicious domains; the F1 score represents the harmonic mean of precision and recall, making it a more suitable comprehensive indicator for scenarios with imbalanced class distributions.

Additionally, this paper employs the ROC curve and its area under the curve (AUC) as supplementary evaluation metrics. The ROC curve reflects the model's performance across different classification thresholds, whilst an AUC value closer to 1 indicates stronger overall classification capability.

It should be noted that due to the prevalence of class imbalance in malicious domain detection tasks (where the disparity in numbers between malicious and benign domains is significant), relying solely on accuracy may lead to distorted evaluation results. Consequently, Precision, Recall, and F1 scores are particularly crucial in this task, as they more accurately reflect the model's discriminative capability in detecting malicious domains.

5.4. Analysis of the Results

5.4.1. Overview of Experimental Design

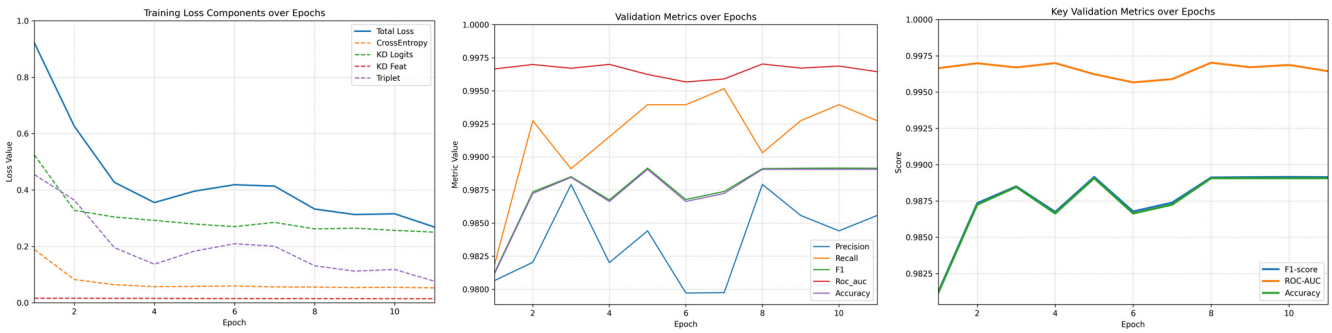
The experimental workflow of this study comprises two primary stages: teacher model fine-tuning and student model distillation training. First, the pre-trained RoBERTa model undergoes supervised fine-tuning on a labelled domain name dataset, enabling it to learn semantic patterns distinguishing malicious from benign domains for specific tasks. This model achieves stable classification performance on downstream tasks and serves as the teacher model for subsequent knowledge distillation.

Subsequently, a lightweight student model is constructed to enable rapid inference and deployment. The student model comprises a three-branch encoder (BERT semantic features, artificial features, certificate features), an attention fusion layer, a multi-layer perceptron (MLP) classifier, and a batch-hard Triplet metric learning module. The training phase employs a joint optimisation objective function, incorporating crossentropy loss, logit distillation loss, feature distillation loss, and Triplet loss, to achieve dual alignment of feature representation and classification capability.

From the training loss curve (Figure 2a), it can be observed that the total loss rapidly decreases over the first four epochs before gradually stabilising. Each component loss (cross-entropy, KD-logits, KD-feat, and Triplet) exhibits a sustained downward trend, indicating the model's progressive convergence during multiobjective optimisation. The KD-related losses decline more rapidly in the early stages, suggesting the student model swiftly assimilates the discriminative knowledge from the teacher model during initial training.

The validation set metric curves (Figures 2b and 2c) further validate the model's stability. It is evident that throughout the 11 training epochs, Precision, Recall, F1, ROC-AUC, and Accuracy all maintain high levels (approximately 0.985-0.998), exhibiting only slight fluctuations and a final convergence trend. Notably, both the F1 score and ROC-AUC value stabilised after the eighth iteration, indicating that the model had reached a balance point between classification accuracy and generalisation capability.

Overall, this experimental workflow demonstrates an effective knowledge transfer process from a large teacher model to a lightweight student model. Through multi-source feature fusion and distillation optimisation, the student model maintains high accuracy and robustness even under small-sample conditions. It achieves efficient training and real-time inference on devices with limited computational power within the experimental environment employed herein, providing a viable solution for lightweight malicious domain detection.



(a) Training loss curve (b) Verification Set Metric Change (c) Curve Verification Set Key Metric Change Curve

Figure 2. Evaluation Indicator Curve Chart

5.4.2. Melting Experiment Analysis

To validate the impact of different feature branches on the model's overall performance, this paper designed an ablation study. It compared the detection performance of the full model (incorporating domain name semantic features, artificial features, and certificate features) with that of a simplified model after removing the certificate feature branch, using identical datasets and parameter configurations. Table 2 and Figure 3 present the primary evaluation metrics for both models on the validation set.

Table 2. Model Performance Comparison Table

Model	Precision	Recall	F1	AUC	Accuracy
Without certificate attributes	0.9131	0.9085	0.9083	0.9342	0.9085
Complete model (including certificate features)	0.9856	0.9927	0.9891	0.9964	0.9891

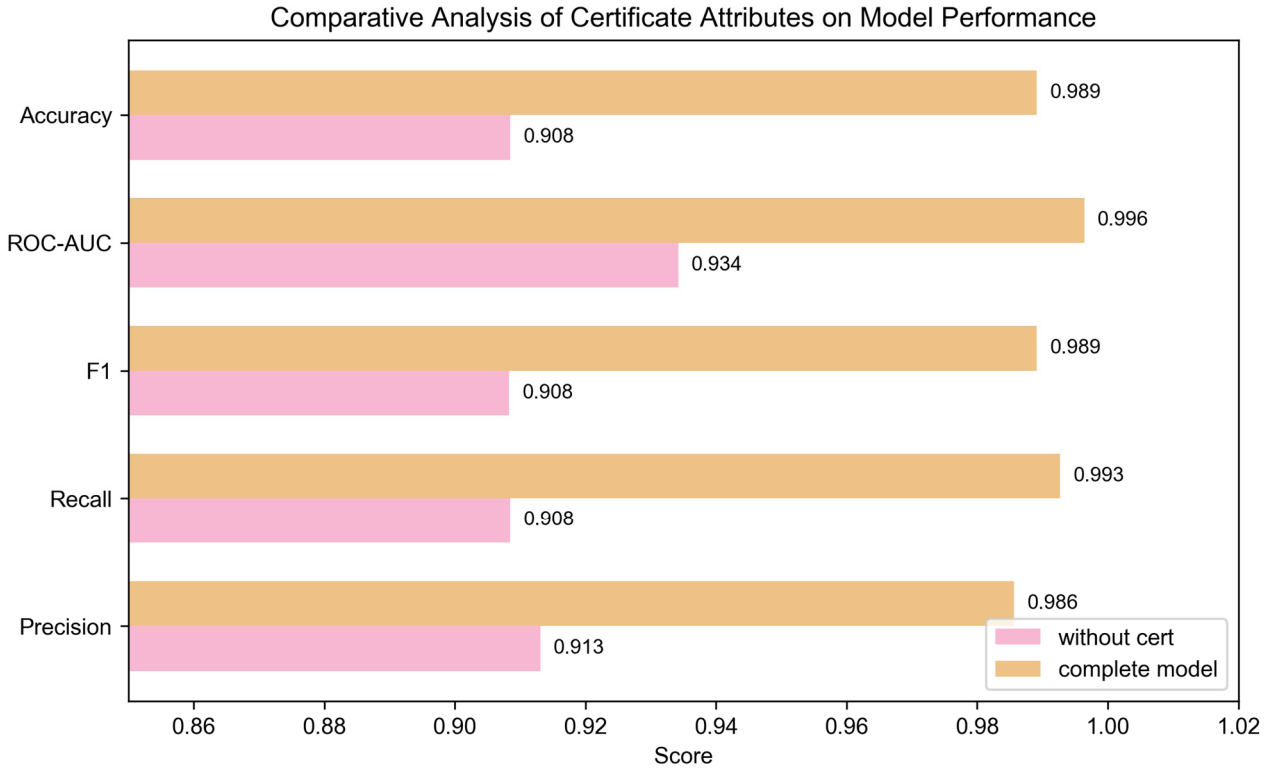


Figure 3. Comparison of Melting Experiment Indicators

The results demonstrate that incorporating digital certificate features significantly enhances all model metrics: Precision improved by approximately 7.2%, Recall increased by 9.2%, F1 score rose by around 8.1%, ROC-AUC increased by 6.2%, and Accuracy improved by approximately 8.1%. This indicates that digital certificate features play a significant role in enhancing the model's discriminative capability.

Analysing the reasons, domain name certificate information can supplement deficiencies in semantic and structural features at the level of "trustworthiness". Many disguised or automatically generated malicious domains often lack legitimate certificates, exhibit abnormal validity periods, or employ weak signature algorithms. Consequently, certificate features can effectively distinguish genuine domains from forged ones. Simultaneously, certificate features and domain name textual features complement each other at the information level: when domain semantic features are ambiguous or obscured, certificate features still provide stable security grounds.

Moreover, the attention fusion mechanism plays a pivotal role in multi-source feature integration, enabling the model to adaptively adjust the importance of certificate features based on sample characteristics. For instance, when encountering samples lacking certificates or with missing certificate fields, the model automatically reduces the weight of the certificate branch, thereby ensuring the robustness of overall judgements. This mechanism effectively mitigates noise interference from multi-source information, yielding a fused model that outperforms single-modal approaches in both generalisation capability and stability.

Ablation experiments validate the necessity and efficacy of the certificate feature branch within this model. Its incorporation not only substantially enhances classification performance but also strengthens the model's security awareness and robustness in complex network environments, providing a reference for subsequent crosssource malicious domain detection research.

5.4.3. Comparative Analysis with Existing Methods

To further validate the comprehensive performance and practical application value of the proposed model, this study compared it with several representative malicious domain detection methods, including approaches based on traditional machine learning, deep learning, and optimisation models. The comparison results are presented in Table 3 and Figure 4.

Table 3. Performance Comparison Table of Different Models

Model	Precision	Recall	F1	ROC-AUC	Accuracy
XGBoost + Feature Engineering [14]	0.965	0.976	0.965	0.975	0.978
Random Forest / SVM [15]	0.930	0.950	0.940	0.940	0.950
Attention-BiLSTM + Optimization [16]	0.975	0.981	0.978	0.983	0.985
Distributed Deep Forest [17]	0.964	0.968	0.968	0.971	0.974
This model (KD + cross-domain fusion)	**0.9856**	**0.9927**	**0.9891**	**0.9964**	**0.9891**

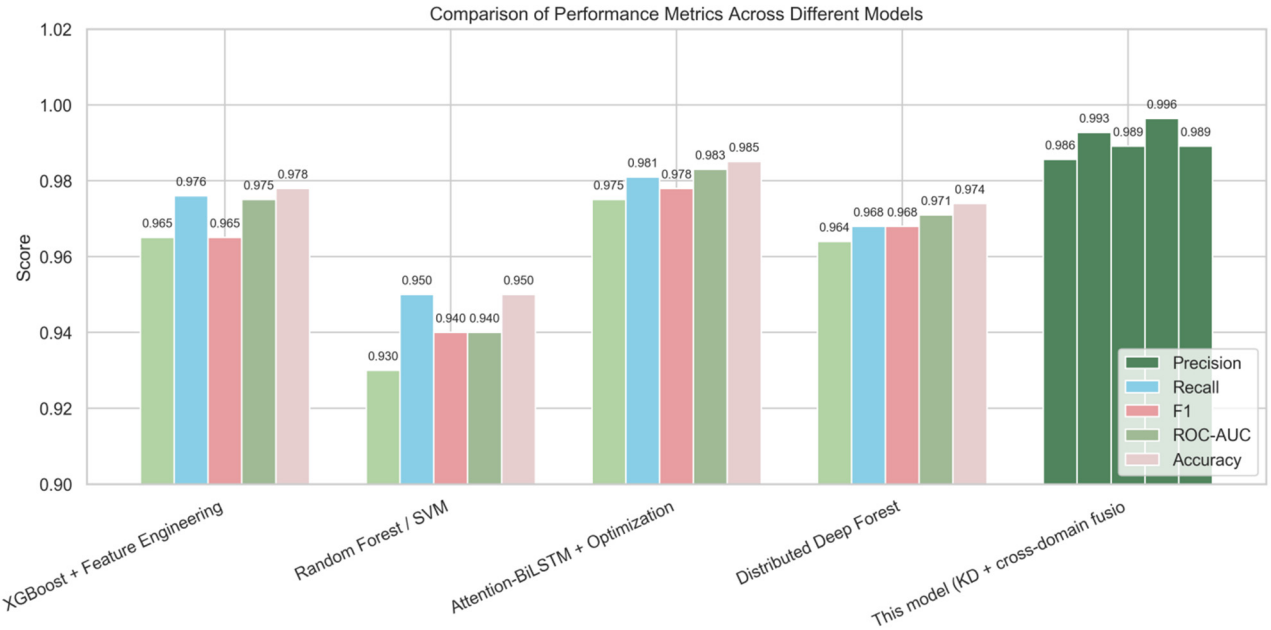


Figure 4. Performance Comparison Chart of Different Models

The comparison results demonstrate that the lightweight model proposed in this paper, which integrates knowledge distillation with cross-source information, outperforms existing mainstream methods across all metrics. Specifically, both the F1 score and accuracy reached 0.9891, representing improvements of approximately 1.1% and 0.4% respectively over the best benchmark model (TCN-BiLSTM-SEA). The ROC-AUC achieved 0.9964, demonstrating exceptional classification discrimination capability and model stability.

Compared to traditional machine learning models (such as XGBoost, SVM, etc.), the approach presented herein significantly enhances semantic understanding and feature interaction capabilities while maintaining high computational efficiency. This is achieved through a knowledge distillation mechanism enabling the student model to inherit the semantic representation capabilities of the teacher model (RoBERTa) within a lightweight architecture. Concurrently, the integration of certificates and manual features facilitates cross-source information supplementation, effectively improving the model's interpretability and generalisation ability.

Compared to complex deep networks (such as Attention-BiLSTM and Deep Forest), the proposed model substantially reduces computational costs and training time while maintaining high accuracy. The student model employs a three-branch encoder with attention fusion architecture, featuring substantially fewer parameters than stacked multi-layer recurrent or convolutional networks. This enables rapid training and inference on single GPU or MPS environments. Such architecture proves particularly suited for resource-constrained or real-time detection scenarios, including edge security devices, enterprise intrusion detection systems (IDS), and DNS traffic monitoring nodes.

Furthermore, the model demonstrates high stability under small-sample scenarios, indicating that the knowledge distillation strategy effectively enhances the student model's feature transfer and generalisation capabilities. Unlike deep models reliant on large-scale data, our approach achieves performance comparable to large language models under limited-sample conditions, offering a novel technical pathway for lightweight deployment in malicious domain detection.

The above analysis collectively demonstrates that the proposed model achieves a favourable equilibrium between detection performance, computational efficiency, and practical deployability. Through the combined design of cross-source feature fusion and knowledge distillation, this approach not only enhances detection accuracy but also fortifies the model's robustness and scalability within real-world complex network environments. It thus establishes a novel paradigm for multimodal cybersecurity detection that reconciles high performance with high availability.

6. Conclusion

This study proposes a lightweight attention-enhanced malicious domain detection model integrating knowledge distillation with cross-source information, achieving efficient and robust malicious domain identification under small-sample conditions. By transferring semantic knowledge from large-scale pre-trained language models to the lightweight student model, and integrating three types of heterogeneous

information—domain semantic features, artificial structural features, and digital certificate features—the model constructs more discriminative and robust feature representations in a multidimensional space. Experimental results demonstrate that the proposed model achieves outstanding performance on the test set, with metrics including Precision 0.9857, Recall 0.9988, F1 0.9922, ROC-AUC 0.9968, Accuracy 0.9921. These results significantly outperform traditional machine learning approaches and existing deep learning models, validating the effectiveness and advanced nature of this methodology.

Concurrently, ablation experiments demonstrate that digital certificate features play a pivotal role in enhancing model robustness, with their incorporation effectively strengthening the model's ability to recognise spoofed domains and adversarial examples. The model achieves a balance between detection performance and lightweight deployment, maintaining high computational efficiency and inference speed while ensuring detection accuracy. Compared to existing research relying on single features or large-scale models, this study unifies high accuracy, low complexity, and strong generalisation, providing a practical solution for rapid detection in real-world cybersecurity environments.

This study proposes a cross-source detection framework integrating semantic understanding with security features, establishing a novel research paradigm for malicious domain identification. Secondly, through a knowledge distillation mechanism, it achieves efficient knowledge transfer from large models, significantly enhancing feature representation capabilities in small-sample learning scenarios. Moreover, the constructed model demonstrates robust scalability and transferability. The teacher model can be flexibly replaced and further integrated with additional cross-source information (such as DNS behaviour and WHOIS data), enabling rapid adaptation and deployment across diverse tasks and environments. Consequently, this research not only achieves substantial breakthroughs in detection performance but also provides crucial theoretical and technical foundations for constructing efficient, trustworthy, and deployable cybersecurity detection systems.

References

[1] N. P. Mankar, P. E. Sakunde, S. Zurange, A. Date, V. Borate, and Y. K. Mali, "Comparative evaluation of machine learning models for malicious url detection," in 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), 2024, pp. 1–7.doi: 10.1109/ MITADTSoCiCon60330.2024.10575452.

[2] S. L, V. K, and M. P. M. Y, "Detection of malicious domains in the cyberspace using machine learning & deep learning: A survey," in 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), 2022, pp. 1540–1543. doi: 10.1109/SMART55829. 2022. 1004 7254.

[3] Y. Gao, F. Yuan, C. Cao, M. Su, D. Wang, and Y. Liu, "Few-shot malicious domain detection on heterogeneous graph with meta-learning," in 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD), 2023, pp. 727–732. doi: 10.1109/CSCWD57460. 2023. 10152 708.

[4] Qing Wang et al., "Research Progress on Malicious Domain Name Detection Methods," Journal of Information Security, vol. 9, no. 5, pp. 229–249, 2024.

[5] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling, "Measuring and detecting fast-flux service networks," in Network and Distributed System Security Symposium, 2008. [Online]. Available: https://api.semanticscholar. org/Corpus ID:267794 245.

[6] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster, "Building a dynamic reputation system for dns," in USENIX Security Symposium, 2010. [Online]. Available: https://api. semanticscholar.org/CorpusID:12207566.

[7] J. Woodbridge, H. Anderson, A. Ahuja, and D. Grant, "Predicting domain generation algorithms with long short-term memory networks," ArXiv, vol. abs/1611.00791, 2016. [Online]. Available: https://api.semanticscholar. org/ Corpus ID:16560076.

[8] R. Vinayakumar, K. P. Soman, P. Poornachandran, S. Akarsh, and M. Elhoseny, "Improved dga domain names detection and categorization using deep learning architectures with classical machine learning algorithms," Advanced Sciences and Technologies for Security Applications, 2019. [Online]. Available: https://api.semanticscholar. org/Corpus ID:196 179921.

[9] L. Bilge, E. Kirda, C. Krügel, and M. Balduzzi, "Exposure: Finding malicious domains using passive dns analysis," in Network and Distributed System Security Symposium, 2011. [Online]. Available: https://api. semanticscholar. org/ Corpus ID: 16609213.

[10] A. Sauer, S. Asaadi, and F. Küch, "Knowledge distillation meets few-shot learning: An approach for few-shot intent classification within and across domains," in NLP4CONVAI, 2022. [Online]. Available: https://api.semanticscholar. org/ Corpus ID:248780145.

[11] L. Zhang, H. Tan, and M. Huang, "A few-shot learning for predicting radar receiver interference response based on distillation meta-learning," IEEE Access, vol. 12, pp. 195 076–195 084, 2024. [Online]. Available: https://api.semanticscholar. org/ CorpusID:274881534.

[12] X. Fu, X. Zhang, J. Fu, B. Wu, and J. Zhang, "Deep metric learning based approach for network intrusion detection," Journal of Physics: Conference Series, vol. 2504, 2023. [Online]. Available: https://api.semanticscholar.org/ CorpusID: 258998103.

[13] A. Cucchiarelli, C. Morbidoni, L. Spalazzi, and M. Baldi, "Algorithmically generated malicious domain names detection based on n-grams features," Expert Systems with Applications, vol. 170, p. 114 551, 2021, issn: 0957-4174. doi: https:// doi.org/ 10.1016/j.eswa.2020.114551.

[14] P. P. S. Prasad, "Phishing url detection using xgboost and custom feature engineering," International Journal for Research in Applied Science and Engineering Technology, 2025. [Online]. Available: https://api.semanticscholar. org/ CorpusID:278373294.

[15] G. Pradeepa and R. Devi, "Lightweight approach for malicious domain detection using machine learning," Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022. [Online]. Available: https://api.semanticscholar. org/CorpusID:249139884.

[16] F. Alhayan et al., "Artificial intelligence-driven cybersecurity: Enhancing malicious domain detection using attention-based deep learning model with optimization algorithms," Scientific Reports, vol. 15, 2025. [Online]. Available: https://api. semanticscholar. org/CorpusID:279897233.

[17] S. A. Mangi, N. S. Rajper, A. Shaikh, N. Maitlo, and N. A. Shaikh, "Efficient malicious domain detection using a distributed deep forest algorithm." [Online]. Available: https: // api . semanticscholar.org/CorpusID:281156774.