WHOIS のドメイン登録情報に基づく フィッシングサイトの検出

小島 大輝[1,*] 宮地 麟[1] 齊藤 泰一[2] 佐々木 良一[2,3]

概要：フィッシングサイトは，ユーザを欺いて個人情報や資産を窃取することを目的としている．これまでは，ブラックリストに基づく方法，ドメイン名の文字列を利用する方法，URL と Web ページのソースコードを利用する方法など，様々なフィッシングサイト検知手法が提案されてきた．本研究では，ドメインの WHOIS 情報に含まれる登録者のメールアドレスの特徴を利用したフィッシングサイト検知方法を提案する．提案手法では，SSL 証明書の特徴量も利用し，教師あり機械学習を用いて，フィッシングサイトか正規サイトかを判定する．

キーワード：フィッシング，WHOIS，機械学習，SVM，XGBoost

Detection of phishing sites based on Registrant Emails in WHOIS

Daiki Kojima[1,*] Rin Miyachi[1] Taiichi Saito[2] Ryoichi Sasaki[2,3]

Abstract: Phishing sites aim to deceive users and steal their personal information and assets. Various methods have been proposed for detecting phishing sites, including blacklist-based methods, methods using domain name strings, and methods using URLs and web page source codes. In this study, we propose a phishing site detection method that uses the characteristics of the registrant's e-mail address contained in the WHOIS information of the domain. The proposed method also utilizes the features of SSL certificates and uses supervised machine learning to determine whether a site is a phishing site or a legitimate site.

Keywords: Phishing site, WHOIS, Machine learning, SVM, XGBoost

1. はじめに

近年,フィッシングサイトは増加の一途をたどっている．フィッシング対策協議会によると，2023 年における日本のフィッシングに関する報告件数は，1,196,390 件に達し、過去最多であった[1]．図 1 には、近年の日本におけるフィッシングサイトの増加の傾向が示されている[2]．フィッシングとは，実在する組織を騙って，ユーザネーム，パスワード，アカウント ID，銀行口座等の暗証番号，クレジットカード番号といった個人情報を詐取する行為であり、フィッシングサイトとはフィッシングを行うウェブサイトである．フィッシングサイトは，正規サイトと類似したページ構成を持つことによりユーザを欺いて，個人情報の入力を促し窃取することを目的としている．このような手法で個人情報が盗まれることにより，アカウントが乗っ取られ金銭的被害を受けたり，インターネットショッピングサイトで不正購入が行われたりする可能性がある[3]．これらの被害を防ぐために，フィッシングサイトか正規サイトかを判定する研究が進められている．一般的に，フィッシングサイトの判定手法は 2 つに大別できる．1 つ目はブラックリストに基づいて判定する手法である．この手法は報告されたフィッシングサイトの URL リストを使用し，リストのいずれかに一致する場合にブラウザで警告を発するものである．しかし，近年フィッシングサイトは短期間でアクセス不能になる傾向があることや，フィッシング URL が増加しているためブラックリストによる判定は必ずしも十分に機能するとは言えない[4]．2 つ目としては，Web サイトの特徴を収集し，機械学習を用いてフィッシングサイトを判定する手法がある．この手法はブラックリストによる判定と異なり，新しく作られたフィッシングサイトをリアルタイムに判定できる可能性がある．

本研究では，Web サイトの特徴として，その URL のドメインの WHOIS 情報である登録者のメールアドレスを考慮することを提案する．WHOIS 情報，SSL 証明書から得られた特徴を用いて，教師あり機械学習によるフィッシングサイト判定を行う．

本論文は以下の節で構成される．2 節では関連研究について述べる．3 節で提案手法を示し，4 節で特徴量の有効性を評価し，5 節で考察を行う．最後に 6 節で本研究をまとめる．

1 東京電機大学大学院
Tokyo Denki University
2 東京電機大学
Tokyo Denki University
3 一般財団法人日本サイバー犯罪対策センター
Japan Cybercrime Control Center
* 23kmc05@ms.dendai.ac.jp

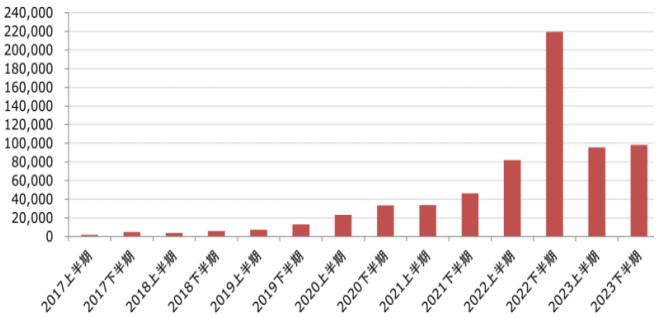

図 1. 国内のフィッシングサイト件数(出典[2])

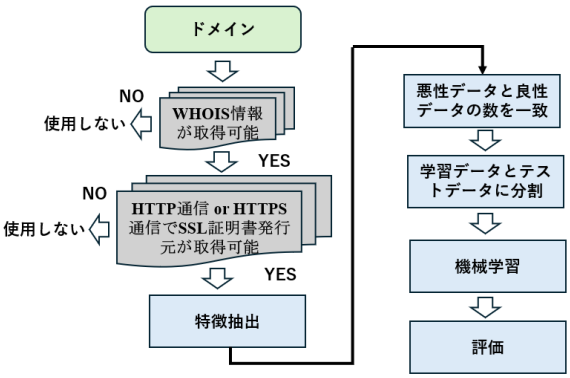

図 2. 提案手法の概要

2. 関連研究

機械学習を用いてフィッシングサイトと正規サイトを判別する手法はこれまでにいくつか提案されているが，以下に本研究に関連する代表的なものを示す.

M.Justin 氏ら[5]は，URL から得られる文字列の特徴とホストベースの特徴を特徴量とした，機械学習によるフィッシングサイトの検知を行った．この研究では URL のドットで区切られるトークンの文字列や‘/’ ,‘?‘ などの記号を bag-of-words を用いてベクトル化し，特徴量としている．さらに，ホストに関する情報として WHOIS のレジストラやレジストラントの情報を用いている．これらも同様に，bag-of-words を用いて単語をベクトル化し，特徴量としている．これらの特徴量を用いて SVM とナイーブベイズにて教師ありの機械学習を行った．その結果，高い精度での検知に成功しているが，bag-of-words の性質上，特徴量の数が膨大となり学習に時間がかかる問題があるとしている.

堺氏ら[6]は，日本をターゲットにしたフィッシングサイトおよび偽ショッピングサイトを調査対象とし，サイトの使い捨て状況や証明書の有効期限などのサイトの特性に焦点を当てた分析を行った．その結果，フィッシングサイトおよび偽ショッピングサイトでは，無料の証明書を使用していることが多い傾向にあり，証明書の有効期限が 90 日以下のものが多いことを明らかにした．また，フィッシングサイトでは短期間でアクセス不能になることや，フィッシングサイトは正規サイトに比べてドメインが長く，ドメインに含まれるドットの数が多い傾向を明らかにした.

また，C&C サーバを検知するために機械学習を用いる研究も存在している.

久山氏ら[7]は，C&C サーバの通信を検知するために，WHOIS と検索エンジンから得られた情報を特徴量として機械学習による検知モデルの作成を行った．この研究では，C&C サーバにアクセスせずに得られる情報を用いて検知を行うことを目的とし，その中で WHOIS から得られる情報のドメイン登録者メールアドレスに注目した．通常のサーバではサーバのドメインと登録者メールアドレスのドメインが一致していることが多く，C&C サーバでは登録者の詳細を隠すためフリーのメールアドレスや登録代行サービスのメールアドレスが使用されていることに着目している．これを特徴量の一つとして SVM とニューラルネットワークを用いた教師ありの機械学習を行った結果，高い精度での検知に成功している.

しかし，WHOIS 登録者メールアドレスのドメインと評価ドメインとの関係を特徴量とした機械学習による判定の研究は我々が知る限りでは行われていない.

そこで，本研究では URL から得ることのできる WHOIS 情報や SSL 証明書を用いて，フィッシングサイトを検知する機械学習の手法を提案する.

3. 提案手法

提案手法は，WHOIS 情報の一部であるドメイン登録者のメールアドレスに着目した検知手法である．本手法は，WHOIS 情報や SSL 証明書から特徴点を抽出し，機械学習を用いてフィッシングサイトかどうかを判別する．提案手法の概要を図 2 に示す.

3.1 データセット

悪性データとして，一般財団法人日本サイバー犯罪対策センター(Japan Cybercrime Control Center:以下,JC3) [8]から提供された 2019/10~2023/11/30 に収集された 405,335 件のフィッシングサイトリスト PhishHunter のうち，WHOIS 情報が取得可能であり，かつ HTTP 通信または HTTPS 通信で SSL 証明書発行元が取得可能な 51,117 件を利用した．また，良性データとして，2024/04/26 に取得した Tranco [9]に掲載されているドメインリスト上位 10 万件から，PhishHunter に出現する 163 種類の TLD に対応する 83,876 件を取得し，その中で WHOIS 情報が取得可能であり，かつ HTTP 通信または HTTPS 通信で SSL 証明書発行元が取得可能な 82,334 件を利用した．これらのドメインから WHOIS および SSL 証明書から得られる情報を抽出する.

表 1. データセットの取得日時と特徴量収集日時

データ	取得日時	特徴量収集日時	件数
PhishHunter	2019/10~2023/11/30	2024/05/11~2024/07/14	51,117
Tranco	2024/04/26	2024/05/11~2024/07/01	82,334

その後，悪性データ数と良性データ数を一致させたうえで，それぞれを学習データとテストデータに 7:3 に分割し，機械学習を行う．

表 1 にこれらの 2 つのデータの取得日時と特徴量収集を行った日時を示す．

3.2 特徴量

本研究では特徴量として，ドメイン登録者のメールアドレス，ドメインの有効日数，証明書発行元を利用する．これらを各々ラベル付する．なお，数値データである場合はそのままの値を特徴量とする．以下に，各特徴量の詳細について説明する．

3.2.1 WHOIS

WHOIS とは，インターネット上の IP アドレスやドメイン名の登録者情報を公開するサービスであり，一般的なインターネットユーザも自由にアクセスすることができる．このサービスは，主に以下の目的でレジストリやレジストラによって提供される[10].

1. ネットワークの安定運用：技術的な問題が発生した際に，迅速な連絡を取るための情報を提供することで，ネットワークの安定的な運用を支援する．

2. ドメイン名の重複確認：新たにドメイン名を申請する際に，既存の同一または類似のドメイン名の存在を確認するために必要な情報を提供する．

3. ドメイン名と商標トラブルの解決：ドメイン名と商標権に関するトラブルを自主的に解決するために必要な情報を提供する．

WHOIS では，以下の情報を提供することが ICANN(Internet Corporation For Assigned Names and Numbers) より定められている[11].

a) 登録ドメイン名
b) 登録ドメイン名のプライマリネームサーバーとセカンダリネームサーバー

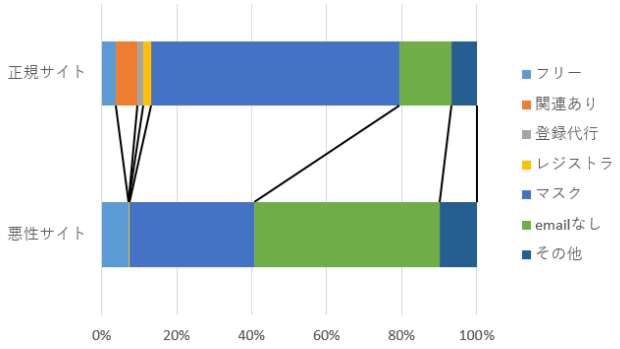

図 3. メールアドレスの比較

c) レジストラ名
d) ドメイン名の登録日
e) ドメイン名の有効期限
f) ドメイン名登録者の氏名および住所
g) ドメイン名の技術担当者の氏名，住所，電子メールアドレス，電話番号
h) ドメイン名の管理担当者の氏名，住所，電子メールアドレス，電話番号

また，一般的に f) には g) ~ h) 同様に氏名および住所以外にも電子メールアドレスや電話番号を WHOIS 情報として得ることができる．

WHOIS には，いくつかの問題が存在する．まず，WHOIS の応答形式には統一されたフォーマットが規定されていないため，WHOIS サーバによってさまざまな応答形式が使用される．その結果，WHOIS 情報の収集を自動化するためにはさまざまな応答形式に対応するようなパーサーを必要とする．また，各国で個人情報保護に関する法律が施行されることに伴い，WHOIS に掲載される情報，特に個人に関する情報の公開の見直しが進んでいる．さらに，WHOIS 情報が犯罪に悪用される可能性があるため，一部レジストリやレジストラは，WHOIS に掲載する情報の項目を減らしたり，登録者情報などを隠したりするなどの対応を取る傾向がある[10].

久山氏らが行ったボットネットの C&C サーバを検知する研究では，WHOIS 情報であるドメイン登録者のメールアドレスを特徴量として用い，それが有効であることを示した[7]．フィッシングサイトも，同様に，フィッシングに用いられるドメインは，身元を特定されないためにドメイン登録時に WHOIS 登録代行サービスを利用して登録情報を隠蔽したり，でたらめな情報を登録したりたりすると考えられる．そこで，我々はメールアドレスを特徴量として用いることとした．

メールアドレスを比較した結果を図 3 に示す．メールアドレスのドメインは以下の基準に基づいて 7 つに分類した．

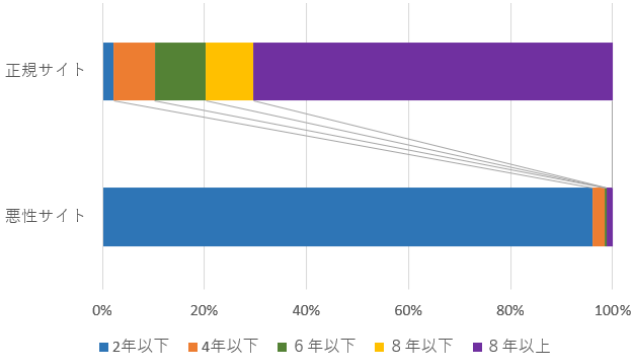

図 4. ドメイン存続期間の比較

図 5. WHOIS の応答例

A) フリー
B) 関係有
C) 登録代行
D) レジストラ
E) マスク
F) email なし
G) その他

各項目の詳細の説明を以下に示す.

A) フリー：メールアドレスのドメインが ‘gmail.com’ や ’outlook.jp’などの無料メールサービスのドメインと一致する場合に分類される.
B) 関係有：評価対象のドメインとメールアドレスのドメインが一致，または TLD を除いたドメイン部分が一致する場合に分類される.
C) 登録代行：レジストラからドメインの再販権を得てドメインを販売しているリセラーが保有するドメインと一致する場合に分類される. これらの業者はドメインの登録者を隠すために自社のメールアドレスを登録するサービスを提供している.
D) レジストラ：登録者からドメイン名の登録申請を受け付け，その登録データをレジストリのデータベースに登録する機関であるレジストラの保有するドメインと一致する場合に分類される.
E) マスク：メールアドレスのドメインが ’whoisprivacyservice.com’ や ’whoisprotect.com’ などの登録者情報を隠すために用いられると考えられるドメインを使用している場合，または WHOIS 検索結果に’REDACTED FOR PRIVACY’と表示される場合，さらに，URL が記載されて WHOIS サーバではなくサイトから申請を行う必要がある場合に分類される.
F) email なし：WHOIS の応答に登録者のメールアドレスの項目が存在しない場合に分類される.
G) その他：上記の分類に当てはまらない場合に分類される.

また，フィッシングサイトのドメイン存続期間は一般的に短いことが知られている[12][13]. そのため，WHOIS 情報から得られるドメインの有効期限年月日と登録年月日から有効日数を算出し，これを特徴量として用いた. ドメイン存続期間の比較した結果を図 4 に，WHOIS の応答例を図 5 に示す.

3.2.2 SSL 証明書

ここでは SSL 証明書の発行元を特徴量として使用することについて説明する. Web においてコンテンツを利用する際に用いられるプロトコルは主に HTTP (Hypertext Transfer Protocol) である. HTTP では Web ブラウザと Web サーバ間の通信は暗号化されず，平文でのやり取りされるため接続を監視する第 3 者によって通信の内容の盗聴や改ざん，なりすましなどのリスクがある. これを解消するために，SSL/TLS プロトコルにより提供されるセキュアな接続の上で通信を行う HTTPS 通信が普及している. HTTPS 通信を実現するには SSL 証明書を利用する必要がある[14].

SSL 証明書は SSL/TLS 技術を用いて認証局(CA)が発行する電子証明書である. 電子証明書は階層構造を持ち，下位の認証局は上位の認証局に証明書を発行してもらうことで信頼性を担保している. 証明書が必須となる HTTPS を利用しているサイトは信頼性が高いと言える[15]. 従来，企業などが運営する正規サイトでは信頼性を高めるため HTTPS を実装していたが，フィッシングサイトではその傾向が見られなかったためこれを指標として 悪性サイトであるかの判定が行われていた[5]. しかし，近年は無料で SSL 証明書を発行できる認証局の登場や，HTTPS 通信を使用していない場合にブラウザが警告を発するため，ユーザに警戒されることからフィッシングサイトにおいても HTTPS 通信を利用する割合が増加している. The PhishLabs Blog によるとフィッシングサイトにおいて HTTPS 通信の利用率は，2016 年には 5%程度であったものが 2020 年では 80%程度に増加していると報告されている 16]. また，Google が公開する透明化レポートによると，Windows 版 Chrome において HTTPS 経由で読み込まれたページの割合

は 7 月 6 日時点で 94%を超えておりサイト全体の HTTPS の利用率が増加している事がわかる[17]．これらの理由から，HTTPS の利用の有無のみではフィッシングサイトの判定が困難であると予想した．そこで我々は CA 局に注目した．前述の通り，CA 局には無料で SSL 証明書を発行するものと発行に金銭の支払いやサイト運営者の実体認証が必要であるものがある．フィッシングサイトではサイト構築のコストを下げるため証明書発行が無料もしくは低価格で行えるものを利用していると考えられる．一方，正規サイトでは信頼性向上のため，有料で SSL 証明書を発行する CA 局や実体認証を行うものを利用していると考えられる．

HTTPS 通信を利用しているフィッシングサイトと正規サイトの証明書を取得し発行者のコモンネーム（CN）を確認する．CN 毎にラベル付を行い特徴量として用いる．ただし，HTTP 通信を利用していないこともそれを特徴量とする．

使用するデータに関しては，正規サイトの証明書は自らアクセスをして収集する．また，悪性サイトの証明書は PhishHunter のデータに含まれていたものをそのまま使用する．

3.3 訓練モデル構築

収集したデータセットから学習データとテストデータを作成し，機械学習モデルを構築する．学習データとテストデータは 7:3 に分割した．機械学習アルゴリズムではグリッドサーチを用いて最適なパラメータを求めた．各機械学習アルゴリズムの詳細について説明する．

3.3.1 SVM

SVM(support vector machine)とは，得られたデータを 2 つに分類する教師あり機械学習の一種である．

主に以下の 2 つの手法によって高精度な分類および回帰分析を実現している[18].

1. マージンの最大化：SVM は，分類境界線とデータポイントとの距離，すなわち「マージン」を最大化することを目指す．これにより，SVM は分類境界をデータから最も離れた位置に設定し，分類の信頼性を高める．この手法によりモデルは新しいデータに対する汎化能力を向上させる．

2. カーネル法：SVM は，「非線形データ」を分類するためにカーネル法を用いる．カーネル法は，クラスの境界を直線で表現できない複雑なデータを，高次元空間に写像することにより線形分離可能にする手法である．この手法を用いることで，SVM は非線形な境界を持つデータに対しても高精度な分類を実現できる．

これら 2 つの手法により，過学習が起こりにくく，データの次元が増加しても高精度な分類が可能になる．

3.3.2 XGBoost

XGBoost（eXtreme Gradient Boosting）は，機械学習のアルゴリズムの中で広く用いられている手法であり，主に分類および回帰の予測タスクに対応する．XGBoost は，勾配ブースティング技術を強化するために，並列処理，木の刈り込み，欠損値の処理，正則化などの高度な手法を採用している．これにより，ソフトウェアとハードウェアの最適化を通じて，最小限のコンピュータ資源で短時間に高いパフォーマンスを発揮することができる．このモデルは，訓練データの特徴ベクトルとターゲット変数のペアを用いて学習し，各特徴に基づいた判断を行う複数の決定木を組み合わせることで，予測精度を向上させる．具体的には，データセットのパラメータを定義し，モデルが将来的にそのパラメータを基に予測を行う能力を持つようにする．さらに，XGBoost は勾配降下アーキテクチャを使用して，通常は CART（分類と回帰木）による弱学習のパフォーマンスを向上させるとともに，システムの最適化とアルゴリズムの強化により，基本的な勾配ブースティング機械（GBM）を改良している[19].

4. 評価

本節では，3 節で述べた提案手法を用いてフィッシングサイトの判定を行った．また，評価指数としては，正解率，適合率，再現率，F1 値を算出した．SVM および XGBoost で構築した機械学習モデルを評価した各評価指数の値を表 2 に示す．特徴量としては，メールアドレス，有効日数，証明書発行元の 3 種類を使用し，それぞれの組み合わせを変えて評価を行った．評価結果より，機械学習において，分類器として XGBoost を利用した場合，SVM を利用した場合よりも精度が高くなった．SVM の正解率は 96.4%であり，XGBoost の正解率は 96.7%であった．正解率は，機械学習モデルが正しく予測できた割合を示す指標である．そのため，SVM および XGBoost はともに比較的高い正解率を達成できた．

表 2 より，特徴量として単体では，証明書発行元と有効日数が同程度の正解率を示し，次いでメールアドレスという順であることがわかった．

今回，特徴量として利用したメールアドレス単体の正解率は，SVM および XGBoost ともに 70%程度であった．これらの結果から，WHOIS 情報であるドメイン登録者のメールアドレスを特徴量として用いることは，一定の有効性はあるものの，高い有効性があるとは評価できなかった．

表 2．各評価指数の値

	SVM				XGBoost			
特徴量の組み合わせ	正解率	適合率	再現率	F1 値	正解率	適合率	再現率	F1 値
メールアドレス	72.4%	78.1%	62.9%	69.6%	70.4%	79.8%	54.5%	64.8%
有効日数	89.2%	82.5%	99.2%	90.1%	89.9%	84.2%	98.2%	90.7%
証明書発行元	88.9%	84.7%	94.8%	89.4%	88.6%	84.5%	94.5%	89.2%
メールアドレス ＋ 証明書発行元	90.0%	85.3%	96.5%	90.6%	89.5%	86.3%	94.0%	90.0%
メールアドレス ＋ 有効日数	91.3%	86.1%	98.4%	91.9%	91.5%	86.7%	97.9%	92.0%
証明書発行元 ＋ 有効日数	94.4%	90.5%	99.2%	94.7%	95.2%	96.1%	94.1%	95.1%
メールアドレス ＋ 有効日数 ＋ 証明書発行元	96.4%	94.3%	98.7%	96.4%	96.7%	95.3%	98.3%	96.8%

5. 考察

本研究では, WHOIS 情報に含まれる登録者のメールアドレスの特徴を用いてフィッシングサイトを検知する手法を提案し，機械学習モデルの性能評価を行った．

表 2 より，証明書発行元単体の精度が比較的高いことが分かる．堺氏ら[6]の研究では，正規サイトとフィッシングサイトで利用されている証明書発行元に大きな差があることを明らかにされており，この差異が有効な特徴量として機能したと考えられる．また，フィッシングサイトのドメイン存続期間は一般的に短いことがこれまでの研究で明らかになっている[12][13]．本研究でも，有効日数単体の精度が比較的高いことから，この傾向が現在も続いていることが確認され，特徴量としての有効であることが分かった．

最後に, WHOIS 情報であるドメイン登録者のメールアドレスは，当初期待していたほど有効な特徴量として機能しなかった．図 1 より，悪性サイトは正規サイトと比べて，「フリー」が多く,「関連有」が少ないという傾向があるのにもかかわらず，特徴量として有効に働かなかった．その理由として, 正規サイトおよび悪性サイトともに「マスク」や「email なし」が大半を占めたため，有効性が低くなってしまったと考えられる．久山氏ら[7]の研究では，フィッシングサイトではなくボットネットの C&C サーバを対象としているが, その際にはみられなかった「マスク」や「email なし」が，本研究では大半を占めていた．このことから，各国で個人情報保護に関する法律が施行されることに伴い, WHOIS に掲載される情報, 特に個人に関する情報の公開の見直しがここ数年で非常に進んでいることが分かる[10].

6. おわりに

本論文では，ドメインの WHOIS 情報に含まれる登録者のメールアドレスの特徴に注目したフィッシングサイト検知手法を提案した．提案手法では，SSL 証明書の特徴量も利用し，教師あり機械学習を用いて，フィッシングサイトか正規サイトかの判定を行った．提案手法を実データに適用し評価した結果，正解率は SVM および XGBoost ともに約 96%と比較的高い値を達成した．しかし，WHOIS 情報であるドメイン登録者のメールアドレスは，当初期待していたほど有効な特徴量として機能しなかった．その理由としては，悪性サイトは正規サイトと比べて,「フリー」および「登録代行」が多く,「関連有」が少ないという傾向があるのにもかかわらず, 近年, WHOIS に掲載される公開情報の見直しが進んでおり，正規サイトおよび悪性サイトともに「マスク」や「email なし」が大半を占めたため，有効性が低くなってしまったと考える．

今後の課題としては，さらなる精度向上のため新たな特徴量の組み合わせなどを検討し，フィッシングサイトの被害を低減できるような実環境でのシステムの提案をしていきたいと考える．

謝辞　本研究は，一般財団法人日本サイバー犯罪対策センター様よりデータを提供していただき実施している．本研究を進めるにあたり，データ提供に協力いただいた関係者各位に深く感謝する．

参考文献

[1] 警察庁，令和５年におけるサイバー空間をめぐる情勢等について，https://www.npa.go.jp/publications/statistics/cybersecurity/data/R5/R05_cyber_jousei.pdf ，(参照:2024-08-23)
[2] フィッシング対策協議会，フィッシングレポート 2024，https://www.antiphishing.jp/report/phishing_report_2024.pdf，(参照:2024-08-23)
[3] 警察庁，フィッシングとは，https://www.npa.go.jp/bureau/cyber/countermeasures/phishing.html，(参照:2024-08-23)
[4] フィッシング対策協議会，フィッシングレポート 2023，https://www.antiphishing.jp/report/phishing_report_2023.pdf，(参照:2024-08-23)
[5] Ma, Justin, et al. "Beyond blacklists: learning to detect malicious web sites from suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009.
[6] 堺啓介, et al. "日本を対象とする偽サイトの動向分析に関する研究." コンピュータセキュリティシンポジウム 2023 論文集 (2023): 596-603.

[7] 久山真宏, 柿崎淑郎, 佐々木良一. "攻撃者に察知されにくい情報を用いた C&C サーバの検知手法の提案と評価." 情報処理学会論文誌 58.9 (2017): 1410-1418.
[8] 一般財団法人日本サイバー犯罪対策センター，https://www.jc3.or.jp/
[9] Tranco，https://tranco-list.eu/，(参照:2024-08-23)
[10] 株式会社日本レジストリサービス，Whois とは，https://jprs.jp/about/dom-search/whois/，(参照:2024-08-23)
[11] ICANN，レジストラ認定契約，https://www.icann.org/resources/unthemped-pages/raa-2001-05-17-en，(参照:2023-08-17)
[12] 千葉大紀，森達哉，後藤滋樹. "悪性 Web サイト探索のための優先巡回順序の選定法." コンピュータセキュリティシンポジウム 2012 論文集 2012.3 (2012): 805-812.
[13] 宮澤孝如，寺田真敏, and 土居範久. "フィッシングサイトの特徴を用いた検出手法の改善." 情報処理学会論文誌 49.9 (2008): 3112-3120.
[14] ネットワークエンジニアとして，SSL とは，https://www.infraexpert.com/info/server16.html，(参照:2024-08-23)
[15] NTT コミュニケーションズ，SSL 証明書とは? 発行の仕組みと種類，https://www.ntt.com/business/sdpf/knowledge/archive_91.html，(参照:2024-08-23)
[16] NTT Data，WEB サイトを安全に利用するために知りたい，SSL 証明書を取り巻く変化，https://www.nttdata.com/jp/ja/trends/data-insight/2021/0121/，(参照:2024-08-23)
[17] Google，ウェブ上での HTTPS 暗号化，https://transparencyreport.google.com/https/overview?hl=ja，(参照:2024-08-23)
[18] 一般財団法人生成 AI 活用普及協会, SVM（サポートベクターマシン）とは？特徴や仕組み，メリットや活用事例をわかりやすく徹底解説！，https://gen-ai-media.guga.or.jp/glossary/svm/，(参照:2024-08-23)
[19] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
[20] PhishTank，https://phishtank.org/，(参照:2024-08-23)

付録

付録 A サイト識別の追加評価

本付録では新たに，悪性データとして，JC3 から提供された 405,335 件のフィッシングサイトリスト PhishHunter のすべてのデータを利用し，正規データとして Tranco に掲載されている Web サイト上位 10 万件から，PhishHunter に出現する 163 種類の TLD に対応する上位 83,876 件を利用する．つまり，WHOIS 情報が取得可能かどうかは考慮せずに機械学習を行い評価を実施する．これらのドメインを利用し，WHOIS から得られる情報として登録者メールアドレスと有効日数，さらに SSL 証明書の発行元情報を抽出する．また，メールアドレスの分類に WHOIS 情報が取得不可であった場合を加えた．その後，悪性データ数と正規データ数を一致させたうえで，それぞれを学習データとテストデータに 7:3 に分割し，機械学習を行う．ただし，追加研究では，XGBoost で構築した機械学習モデルのみ評価した．

表 3 にこれらの 2 つのデータの取得日時と，特徴量収集を行った日時を示す．

これまでと同様に，3 節で述べた提案手法を用いてフィッシングサイトの識別を行った．また，評価指数では，正解率，適合率，再現率，F1 値を算出した．各評価指数の値を表 4 に示す

表 4 より，メールアドレスの有効性が向上した．その理由として，WHOIS 情報が取得不可のデータが 30 万件ほど含まれていた．そのため，WHOIS 情報が取得不可という分類が特徴量として働いたためと考える．

表 3．データセットの取得日時と特徴量収集日時

データ	取得日時	特徴量収集日時	件数
PhishHunter	2019/10~2023/11/30	2024/05/11~2024/07/14	405,335
Tranco	2024/04/26	2024/05/11~2024/07/01	83,876

表 4．各評価指数の値

	XGBoost			
特徴量の組み合わせ	正解率	適合率	再現率	F1 値
メールアドレス	85.6%	98.6%	72.3%	83.4%
有効日数	73.8%	97.3%	49.0%	65.2%
証明書発行元	89.3%	84.4%	96.5%	90.0%
メールアドレス ＋ 証明書発行元	93.8%	98.0%	89.5%	93.6%
メールアドレス ＋ 有効日数	93.3%	92.2%	94.4%	93.3%
証明書発行元 ＋ 有効日数	92.6%	96.4%	88.5%	92.3%
メールアドレス ＋ 有効日数 ＋ 証明書発行元	98.1%	97.0%	99.3%	98.1%

表 5．正規サイトと悪性サイトのコモンネーム TOP5

コモンネーム(悪性サイト)	コモンネーム(正規サイト)
ZeroSSL	Let’s Encrypt
Let’s Encrypt	Google Trust Services LLC
DigiCert Inc	DigiCert Inc
Cloudflare Inc	Amazon
Google Trust Services LLC	Sectigo Limited

付録 B SSL 証明書

本研究においてフィッシングサイトの判定に有効であった SSL 証明書の発行者について議論する．モデル作成に使用したフィッシングサイトと正規サイトで用いられていた各 SSL 証明書の Issuer のコモンネーム TOP5 を表 5 に示す．

表 5 より，正規サイトと悪性サイトの両方で ”Let’s Encrypt” などの無料で利用できる SSL 証明書を利用していた．しかし，正規サイトでは有料の SSL 証明書を利用しているものが多く見られたが，悪性サイトでは大半が無料で利用できる SSL 証明書を利用していた．

表 6. 各データセット

悪性学習データ	PhishHunter
良性学習データ	Tranco
悪性テストデータ	PhishTank
良性テストデータ	Tranco

表 7. データセットの取得日時と特徴量取集日時

データ	取得日時	特徴量収集日時	件数
PhishTank	2024/06/07~2024/07/04	2024/07/09~2024/07/12	8,641

表 8. 各評価指数の値

	XGBoost			
特徴量の組み合わせ	正解率	適合率	再現率	F1 値
メールアドレス	46.3%	47.7%	76.4%	58.7%
有効日数	50.8%	52.0%	20.1%	29.0%
証明書発行元	49.7%	49.2%	16.6%	24.8%
メールアドレス ＋ 証明書発行元	43.8%	37.6%	18.7%	25.0%
メールアドレス ＋ 有効日数	48.7%	47.0%	19.8%	27.9%
証明書発行元 ＋ 有効日数	49.7%	49.2%	16.6%	24.8%
メールアドレス ＋ 有効日数 ＋ 証明書発行元	53.4%	64.3%	15.5%	25.0%

付録 C PhishTank データによるモデル評価

本付録では，新たにフィッシングデータベースである PhishTank[20]を悪性テストデータとして用いて，機械学習モデルを再評価する．つまり，悪性学習データとして PhishHunter，良性学習データとして Tranco を用いて，機械学習モデルを構築する．その後，PhishTank を悪性テストデータ，Tranco を良性テストデータとして評価する（表 6）．

悪性テストデータとして，2024/06/07~2024/07/04 に取得した PhishTank に掲載されているフィッシングサイトから，PhishHunter に出現する 163 種類の TLD に対応する 14,108 件に限定して取得し，その中で WHOIS 情報が取得可能であり，かつ HTTP 通信または HTTPS 通信で SSL 証明書発行元が取得可能な 8,641 件を利用した．表 7 に PhishTank のデータの取得日時と特徴量収集を行った日時を示す．

これまで同様に，3 節で述べた提案手法を用いてフィッシングサイトの識別を行った．また，評価指数としては正解率，適合率，再現率，F1 値を算出した．ただし，追加研究では，XGBoost で構築した機械学習モデルのみを評価した．各評価指数の値を表 8 に示す．

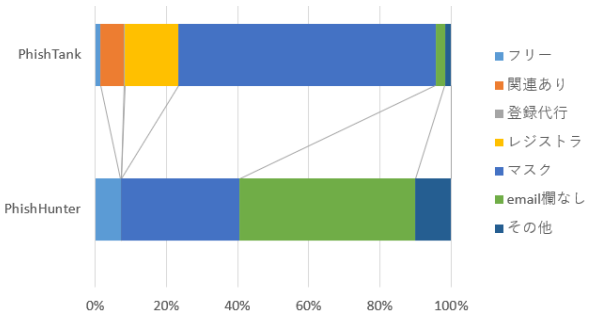

図 6. メールアドレスの比較

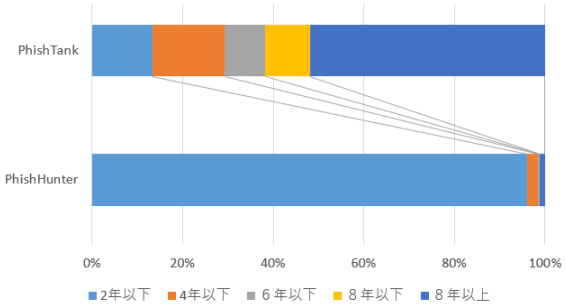

図 7. ドメイン存続期間の比較

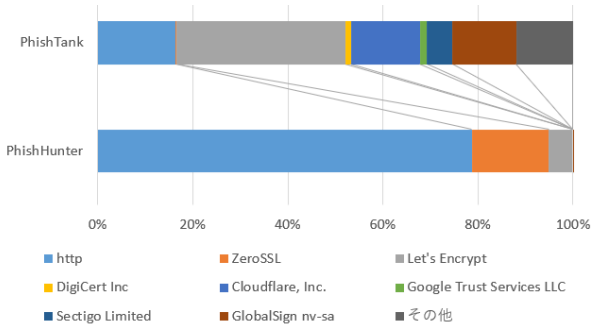

図 8. 証明書発行元の比較

表 8 より，すべての特徴量の組み合わせにおいて，正解率が 50%程度となった．また，図 6 から図 8 より，PhishHunter と PhishTank を比較した際，2 つの悪性データセットにおけるすべての特徴量において顕著な差異がみられた．この差異が精度の著しい低下を引き起こしたと考えられる．この大きな差が生じた理由として，PhishHunter は日本国内を対象としたフィッシングサイトのリストである一方で，PhishTank は主に海外を対象としていることが考えられる．フィッシングサイトのドメイン存続期間は一般的に短い傾向にあるが，図 7 より PhishTank のデータには 8 年以上のものが多く含まれている．これは一般的な傾向と異なる結果であり，PhishTank からの取得件数が少なかったことがデータの偏りをさらに助長した可能性がある．この問題を解明するには，PhishTank からの取得件数を増加させ実施することや，日本国内のフィッシングサイトと海外のフィッシングサイトについて，詳細な調査が必要であると考える．