偽造ウェブサイトに対する顔画像生成技術を用いた判別支援手法の設計と実装

山崎 慎治[1,a)]　宮本 大輔[1]

概要：フィッシング攻撃は，ソーシャルエンジニアリングの手口によって機密情報を不正に入手し利用する詐欺行為で，サイバー社会に対する脅威の一つである．これまでに，システムによるフィッシング攻撃の検知，ユーザーに対する教育による被害の防止，ユーザーがフィッシング攻撃と判断できるような意思決定の支援などの，多角的な手法の研究により数多くの対策が提案されている．ここで，フィッシング攻撃がユーザーを騙すことを特徴とすることから，ユーザーの意思決定を支援する手法が重要である．しかし，SSL 証明書表示のような従来の手法は一般のユーザーにとって難解であるという問題がある．そこで本研究では，ユーザーに提示する簡易な情報の形態として顔画像を用い，フィッシング判別を支援する手法として実装した．本論文では，110 人の実験参加者を対象としたアンケートを実施し，ウェブブラウザの拡張機能として実装した提案手法の効果を分析した．さらに，フィッシング検知技術と組み合わせて用いる方法について議論し，提示する顔画像の彩度を変化させることで，ユーザーにとってより判別のしやすい支援手法を提案する．

キーワード：フィッシング, ユーザー支援, ユーザブルセキュリティ

A Facial Image Generation-based Method for Supporting Users to Identify Fraudulent Websites

Shinji Yamazaki[1,a)]　Daisuke Miyamoto[1]

Abstract: Phishing attacks are insidious fraudulent actions that use social engineering techniques to obtain confidential information by stealth and are thus among one of most serious threats to the cyber society that needs countermeasures. Numerous previous studies have explored anti-phishing methods, and a variety of countermeasures have been proposed. These include developing computer software to detect phishing attacks, educating users to recognize phishing attacks, and providing users with decision-making support tools that can help them identify phishing attacks. Since user deception characterizes most phishing attacks, it is essential to develop methods that support user decision-making processes. However, conventional methods such as Secure Socket Layer (SSL) certificate display are too complex for average users to understand. Therefore, in this study, we propose using face images as a simple form of information for presentation to users. In this paper, we report on a questionnaire survey conducted on users (N=110) to analyze the effectiveness of our proposed method, the results of which show that our method reduces the percentage of wrong answers without compromising user convenience.

Keywords: Phishing, Supporting User, Usable Security

1. はじめに

フィッシングとは，ソーシャルエンジニアリングの手口によって，個人情報や金融口座の情報といった個人や団体

[1] 東京大学大学院
The University of Tokyo
[a)] 6e1a-yomogi@g.ecc.u-tokyo.ac.jp

の保有する機密情報を開示するように誘導し，それらの情報を入手，或いは利用する詐欺行為である [1,2]．ここで，ソーシャルエンジニアリングとは，攻撃対象の保有する機密情報を意図せずに漏洩させる様に仕向けるための心理学的操作手法を表す [3]．フィッシングは 1995 年の America Online (AOL) における事例の報告 [4] から始まり，20 年以上経った現在においても尚，脅威となっている．

我々は，フィッシングへの対策としてユーザーがフィッシング攻撃と判断できるように意思決定を支援するための手法を研究している．先行研究 [5] では，ユーザーに提示する簡易な情報の形態として顔画像を用い，ユーザーを支援するシステムを提案した．初期実装および実験では，8 名の実験参加者を対象として実験を行った．ただし，参加者は全て本学学生であり，サイバーセキュリティの講義の受講者を対象としていたために，実験結果の偏りを否定できない．そこで本論文では，この研究を発展させることを目的とする．実験では 10 代から 50 代までの 110 人に参加してもらい，より実世界に近い集団を対象に実施した．実装についても擬似コードによるモデル化および今後の利用を見据えた検討を行い，フィッシング検知技術との組み合わせといった試験的な機能の実装を行った．

本論文の構成を以下に記す．2 節では関連研究について述べ，本研究で着目した問題点について説明する．3 節では著者の提案するシステムの設計と実装を行い，4 節において調査及び評価を行う．5 節では提案システムの問題点や応用について議論し，最後にまとめと今後の課題について 6 節に述べる．

2. 関連研究

2.1 ユーザーの意思決定支援手法についての既存の研究

本項では，フィッシングの攻撃に直面しているユーザーが正しく意思決定をした上で攻撃を回避できるように，警告や情報の提示によって支援することを目的としたユーザーインターフェースの様々な既存の研究を扱う．

2.1.1 ツールバー

ウェブブラウザに対してブラウザ開発元以外の第三者が提供し，ユーザーがインストールすることで機能を拡張するインターフェースがツールバーである．

Wu ら [6] は，このツールバーを用いたセキュリティ機構が実際にユーザーをフィッシングから保護しているのかを検証するために，実験環境下で被験者にメールが送られ，その一部がシミュレートされたフィッシングサイトに誘導するものである，というシナリオでユーザー調査を行った．ここで多くのユーザーは，ウェブサイトのコンテンツが正規のウェブサイトのものと捉えると，ツールバーの警告を見過ごしてしまうことが報告されている．他にも，ツールバー上の消極的な警告ではなくポップアップ画面のような積極的に割り込む警告のほうが効果的だとも指摘している．

2.1.2 アトラクター

セキュリティに関する警告画面が頻発すると馴化によってユーザーは警告を信頼しなくなり，効果的に被害を防止できなくなる事態は問題である．この問題に対して Bravo-Lillo ら [7] は，最も重要な情報に注意を向けるような UI としてアトラクターを設計し，意思決定を支援する手法として提案した．馴化の下での実験や馴化そのものの影響を調査した結果，表示された重要な情報をユーザー自身が入力する方式やユーザーがマウスポインタでなぞる方式の効果が高いことが報告されている．また一連の調査を踏まえ，馴化の影響を受けた中でのユーザーテストを実施することを提案している．

2.1.3 SSL 証明書およびそのインジケーター

SSL 証明書はウェブサイトのアイデンティティを示し，正しく使用すれば正規のウェブサイトかフィッシングサイトかを判別するための鍵となる要素である．ウェブブラウザには SSL 証明書に記載された情報をユーザーに提示する UI がある．

Felt ら [8] は，ブラウザが不審な SSL 証明書を検知した時に表示する警告を改善することで，よりユーザーが危険性を理解し安全な選択するような UI を目指した．警告を簡潔かつ具体的かつ技術的用語を排することで改良した結果，ユーザーはより安全な選択をしフィッシングを回避したが，警告が意味する SSL 証明書についての危険性の理解を向上させられなかった．

また Felt ら [9] は，アドレスバーの横に表示される接続の安全性を示すアイコンであるインジケーターについての，多様な形状や色を比較するサーベイ調査を行うことにより，インジケーターの改良を目指した．調査結果を元に，状態によって色や形状がはっきりと変化し，簡潔に説明する単語を付随させたインジケーターのセットを提案した．この研究では，インジケーターを設計する際に，スマートフォンなどの小さなデバイスでも表示できること，色覚特性の影響を受けないように形状のみで理解可能であること，アイコンの意味を推定できないユーザーを考慮して言語で説明可能であることが要請されているとも指摘している．

ここで Thompson ら [10] は，EV-SSL 証明書の緑色を用いて強調表示するブラウザの UI がどれだけユーザーに影響を及ぼしているか調査した．インジケーターは改良されたもののユーザーの SSL 証明書に対する理解に影響を及ぼさず，EV-SSL 証明書の緑色の強調表示を無効化しても影響がなかったことが報告されている．

2.2 問題分析

前述の通り，ユーザーがフィッシング攻撃を回避できるように意思決定を支援する様々な研究が行われてきた．しかし，Felt らの研究ではユーザーをより安全な選択へ誘導させることができたものの，SSL 証明書についての危険性

図 1　StyleGAN2 で生成した顔画像例
Fig. 1　Example of a face image generated by StyleGAN2.

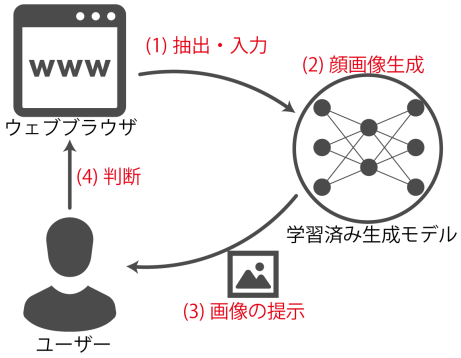


図 2　提案システムの設計図
Fig. 2　Schematic of our proposed system.

への理解を向上させられなかった [8]．また，Wu らが指摘したようにブラウザの提示する情報が無視されることがあり [6]，Thompson らの研究においても報告されている [10] 問題である．

著者は，既存の手法で用いられてきた SSL 証明書や Uniform Resource Locator（URL）の概念が一般のユーザーには難解であることや判別が難しいことで，これらの問題が生ずると考える．

ここで，フィッシング攻撃は実空間ではなく，サイバー空間におけるなりすましであることを再考する．Web ブラウザに表示される通信相手についての情報は，URL や SSL 証明書など文字情報に限られている．しかし，電話によるコミュニケーションでは通話相手の声色のような音声情報が，対面によるコミュニケーションでは相手の表情や背格好といった映像情報がある．実空間における取引やコミュニケーションは，サイバー空間に比べ豊富な情報が複数の形態で提供されており，なりすましについての判断材料となっている．

このことを踏まえ，本研究では通信相手の情報を文字列の形態ではなく映像情報の一種である顔画像を提示する．

3.　提案手法

本研究では，ユーザーが閲覧したウェブサイトの URL から，敵対的生成ネットワーク（Generative adversarial network, GAN）を利用しヒト顔画像を生成し，生成された顔画像をユーザーに提示するシステムを提案する．ユーザーは普段利用するサービスの正規のウェブサイトに対応する顔画像を予め記憶していることを前提条件とする．そのうえで，ユーザーが閲覧しているウェブサイトが正規のウェブサイトかフィッシングサイトであるかを判断する際に，正規のウェブサイトと関連させて既にユーザー自身が

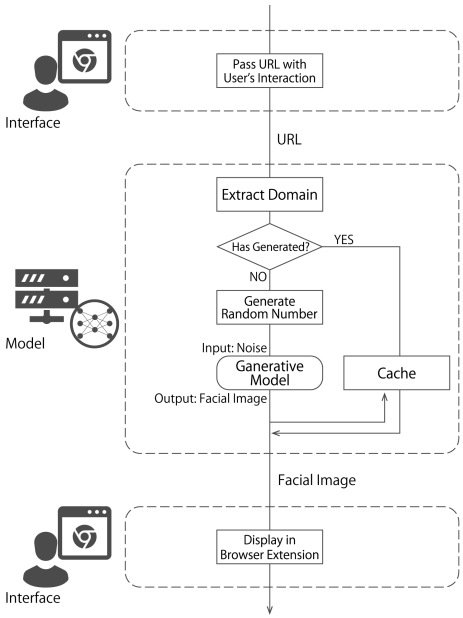


図 3　提案システムの実装図
Fig. 3　Implementation schematic of our proposed system.

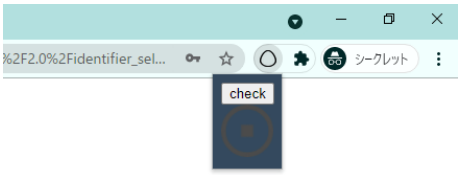


図 4　起動時の拡張機能の様子．check のボタンをクリックすることで次の手順に進む．
Fig. 4　The extension at startup; click on the check button to proceed to the next step.

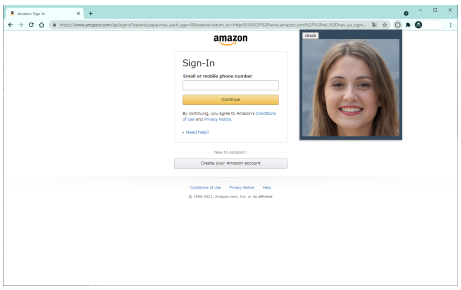


図 5　提案システムでの画像表示例．FQDN(www.amazon.com) に対応した画像はブラウザウィンドウの右上エリアに表示される．
Fig. 5　An example image display by our proposed system. The image associated with the FQDN (www.amazon.com) is presented in the upper right area of the browser window.

記憶している画像と，システムから提示された画像とが，一致するかを判断材料とすることを狙いとしている．

3.1 GAN を利用したヒト顔画像生成

近年機械学習分野の発展は目覚ましいものであり，特にニューラルネットワーク及びディープニューラルネットワークによって諸分野への応用が進んでいる．中でも，互いに競合する 2 つのニューラルネットワークによって実装される GAN では，学習済み生成モデルにノイズを入力することで，学習に用いられたデータに近い疑似データを生成するものである [11]．フィッシングの検知に機械学習を応用した先行研究は数多くあり [12–14]，一部は GAN を用いている [15,16]．

本システムでは，ヒトの顔画像の生成に GAN による学習済み生成モデルを用いる．GAN を用いる利点として 2 つの理由が挙げられる．一つには，生成モデルへの入力値であるノイズを変化させることで数に制限なく顔画像を生成できるために，無数に存在するウェブサイトに個別に対応した顔画像を確保できることがある．もう一つには，疑似データが生成されることにより実在する人に帰属する権利との競合しないことがある．例えば Karras らによる StyleGAN では，図 1 のように実際の顔画像と判別のつかない疑似顔画像を生成されることが報告されている．

3.2 システムの設計

本論文で提案するシステムは図 2 に示すような構造を持つものである．このシステムは次のような手順で動作する．(1) ウェブブラウザで閲覧しているウェブサイトから URL，ドメイン名といったウェブサイトの同一性を示す情報を抽出し，ノイズを生成する．(2) 生成されたノイズを GAN による学習済みモデルに入力し，ヒトの顔画像を生成する．(3) 学習済みモデルはウェブサイトに対応付けられて生成されたヒトの顔画像をユーザーに提示する．(4) ユーザーは提示された画像を元に，現在閲覧しているウェブサイトが正規のウェブサイトであるか，フィッシングサイトであるかを判断する．

3.3 システムの実装

提案したシステムを図 3 のように実装した．画像の生成にあたっては一定水準のリソースを要求されるため，ブラウザの拡張機能としてユーザーインターフェース部分を担当するインターフェース部分と，生成済みモデルによる顔画像生成を担当するモデル部分の，大きく 2 つに分解して提案したシステムを実装した．

このシステムは具体的に次のような手順で動作し，ユーザーに顔画像を提示する．

(1) ユーザー操作によって，現在閲覧しているウェブサイトの URL がインターフェース部分からモデル部分に送られる．(2) モデル部分では，インターフェース部分から送られた URL を元に顔画像を生成し，インターフェース部分に顔画像を返す．(3) インターフェース部分は，モデル部分から送られた顔画像をユーザーに提示する．

インターフェース部分とモデル部分の詳細な実装は以下に記す．

Algorithm 1 Interface Part

```
if check_button.clicked then
    URL ← ExtructUrl(browser.currentPage)
    send URL to ModelPart
    recieve face_image from ModelPart
    display face_image
end if
```

3.3.1 インターフェース部分

ブラウザの拡張機能としてユーザーインターフェースに関する機能を担当するインターフェース部分は，まずユーザーが拡張機能のエリアに表示されるボタンをクリックすることで一連の動作が始まる（図 4）．これにより，インターフェース部分はユーザーが現在閲覧しているウェブサイトの URL を取得しモデル部分に送る．その後，モデル部分から顔画像が返されると，図 5 のようにブラウザの拡張機能の領域上に顔画像を表示する．インターフェース部分の動作を擬似コードで示すとアルゴリズム 1 のようになる．

3.3.2 モデル部分

顔画像を生成するモデル部分は次のように動作する．まず，インターフェース部分から送られた URL から Fully Qualified Domain Name (FQDN) を抽出する．顔画像はこの FQDN を元に生成される．FQDN を参照し対応した顔画像が生成済みでなければ，次の手順として FQDN から顔画像生成モデルに入力する乱数列を生成する．ここでは，可変長である FQDN を取り扱いやすいように SHA-256 関数にかけることで固定長とする．この FQDN のハッシュ値を乱数生成器のシード値として，疑似乱数列を生成する．そして，生成された乱数列をノイズとして，学習済み顔画像生成モデルに入力する．今回の実装では学習済みモデルに StyleGAN2 によって事前に学習されたものを用いた．[*1] 出力された顔画像は FQDN と対応付けて保存し，最後に生成した顔画像をインターフェース部分に受け渡す．一方で，FQDN を参照し対応した顔画像が生成済みであれば，保存された生成済み顔画像をキャッシュのように利用しインターフェース部分に受け渡す．このモデル部分の動作を擬似コードで示すとアルゴリズム 2 のようになる．

このような実装により，FQDN を元に乱数を生成し顔画像の生成モデルに入力することによって，文字情報として

[*1] 利用した学習済みモデルは Karras らによって利用可能となっている．`https://nvlabs-fi-cdn.nvidia.com/stylegan2/networks/stylegan2-ffhq-config-f.pkl`(2021/08/20 現在)

Algorithm 2 Model Part

```
recieve URL from InterfacePart
if image.associated_with(URL.domain) exsits then
    face_image ← cached_image[URL.domain]
else
    noise ← GenerateRandomNumbers(URL.domain)
    face_image ← GenerativeModel(noise)
end if
send face_image to InterfacePart

function GenerateRandomeNumbers(domain)
    hashed_domain ← SHA256(domain)
    for i = 0, 1, 2, 3 do
        seeds[i] ← hashed_domain[32i: 32i + 31].to_int
    end for
    for i = 0, 1, . . . , NoiseSize − 1 do
        random_numbers[i] ← Random(seeds)
    end for
    return random_numbers
end function
```

類似したドメインであっても，全く異なる顔画像が生成されることになる．これによって，IDN ホモグラフィック攻撃を含むホモグラフィック攻撃に対して，全く違った画像をユーザーに提示することを可能としている．

3.3.3 提案システムの検証環境

提案システムのプロトタイプが動作するか検証した際の動作環境を次に示す．

- GPU: NVIDIA GeForce GTX 1070 Ti
- OS: Windows10 20H2
- ブラウザ: Windows 版 Google Chrome 87
- Python3.7.8/CUDA10.0/cnDNN7.5/Tensorflow1.14

この動作環境において，Alexa の TOP50 [17] のドメインについて顔画像を生成したところ，一つの顔画像生成にかかる時間は平均 24.34 秒（23.75 s – 26.05 s）であった．なお，既にドメインに対応する画像が生成済みであった場合，先に述べた画像の流用によって画像生成に要する時間を短縮することが可能である．

4. アンケート調査

著者は，株式会社マクロミルを通じ，12 歳〜59 歳の男女合わせて 110 名に協力いただき，2021 年 3 月に WEB アンケート調査を実施した．なお，本調査を実施するにあたって東京大学ライフサイエンス研究倫理支援室を通じて研究倫理審査申請を行っており，承認を得ている．また，調査を依頼した企業では親権者の同意の上で，未成年者による回答が行われており，アンケート調査においても個人を特定可能な情報を収集せず，すべての年代に対して適切に実施できるように配慮した．18 歳以下のユーザーであっても，フィッシング攻撃の対象となり個人情報を窃盗される可能性があるため，アンケート調査の対象から排除しなかった．

4.1 調査方法

参加者は 55 名ずつの二群に分けられ，それぞれのグループに対してアンケートへの回答を要請した．参加者の年齢層ごとの人数は表 2 に示すとおりである．

提案システムの実装を用いたグループをグループ A とし，通常のブラウザを用いたグループをグループ B とする．

4.2 アンケートの質問内容

グループ A に対しては提案システムの実装を用いた画面を，グループ B に対しては提案システムを用いず通常のブラウザの画面を使用した．グループ A の参加者のみに対して，アンケートの最初に提案システムの説明を図 5 とともに行った．アンケートは，(1) 正規のウェブサイトの記憶，(2) ウェブサイトの判別のテスト，(3) ユーザーによる評価の 3 つで構成される．

4.2.1 正規のウェブサイトの記憶

正規のウェブサイトのログイン画面を表示したブラウザの画像を参加者に 4 枚提示し（図 6），それぞれの画像を記憶するように指示をした．

4.2.2 ウェブサイトの判別のテスト

ウェブサイトのログイン画面を表示したブラウザの画像を参加者に提示し，その画像が正規サイトのものであるか，フィッシングサイトのものであるかを判別させた．このテストは各アンケートにつき 10 問ある．各問題で提示される画像は，正規のウェブサイトを表示した (1) と同一の画像，もしくは実際のフィッシングサイトを Web ブラウザに表示した画像である．画像を用いた理由は，参加者が誤って個人情報を入力してしまい漏洩させてしまうことを防ぐためである．

4.2.3 ユーザーによる評価

(1)，(2) を通じて参加者の主観評価を調査するために表 1 に示す質問を行った．質問 1 では判別の難易度を，質問 2 では正規のウェブサイトの記憶にあたっての心理的負担を，質問 3 では判別した際の自信についての評価を，質問 4 では判別にあたっての心理的負担を問うた．これらの質問では，各記述にどの程度同意できるかを 5 段階のリッカート尺度で質問した．

4.3 調査結果

提案法を利用したグループ A（以下，提案群）と，既存の方法を利用したグループ B（以下，対照群）のそれぞれから 55 名ずつの回答が得られた．なお，本項で参照する図中のエラーバーは標準誤差を示す．

4.3.1 正答率及び誤答率の比較

提案群と対照群から得られた回答を比較することにより，提案システムの意思決定支援の効果について検証した．

正規のウェブサイトを正規のウェブサイトと判別するか，フィッシングサイトをフィッシングサイトと判別した

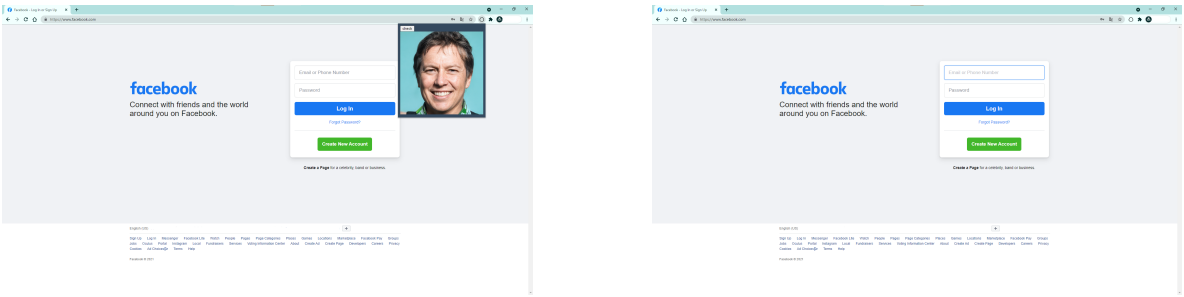


図 6　実験参加者に正規ウェブサイトとして提示した画像の例．これらは同じウェブサイトであり，グループ A には左図をグループ B には右図を提示した．

Fig. 6　Examples of images presented to participants as legitimate websites. The same website was presented to Group A (left) and Group B (right).

表 1　ユーザー評価におけるアンケート項目

Table 1　Questionnaire items for user evaluation.

ユーザー評価の質問項目
Q1. 判別は容易であった．
Q2. 覚えやすかった．
Q3. 自身を持って判別できた．
Q4. 判別に時間や手間はかからなかった．
5 段階リッカート尺度: (1) 全くそう思わない – (5) 強くそう思う

表 2　実験参加者の各年代層における人数

Table 2　Number of people in each age group for the survey.

年齢区分	A (提案群)	B (対照群)
12 – 19	11	11
20 – 29	11	11
30 – 39	11	11
40 – 49	11	11
50 – 59	11	11
合計	55	55

割合を正答率とし，提案群と対照群での正答率の平均値の比較を図 7(a) に示す．正答率の平均値について，提案群が対照群よりも大きい値を示したが，いずれの場合も統計的有意差は認められなかった．

また，正規のウェブサイトをフィッシングサイトと判別するか，フィッシングサイトを正規のウェブサイトと判別した割合を誤答率とし，提案群と対照群での誤答率の平均値の比較を図 7(b) に示す．誤答率の平均値について，提案軍が対照群よりも小さい値を示し，正規のウェブサイトについて有意水準 5 %で，フィッシングサイトについて及びその両方合わせた場合に有意水準 1 %で統計的有意差が認められた．

4.3.2　ユーザーの主観評価

提案群と対照群のユーザーによる主観評価の比較を図 7 に示す．なお，質問内容は表 1 に示したとおりである．各質問の平均スコアは，リッカート尺度の選択肢に 5（強くそう思う）から 1（全く思わない）までのスコアを割り振った上で質問ごとのスコアの平均値を算出した値であり，スコアが大きいほど，質問項目に対して同意できる傾向が強いことを表している．各設問について提案群が対照群よりもより同意できる値を示したが，いずれの場合も提案群と対照群で統計的有意差は認められなかった．

5.　議論

5.1　検知技術との組み合わせによる応用

閲覧しているウェブサイトがフィッシングサイトであるかどうかを判別する既存の技術と組み合わせ，フィッシングサイトである確率といった怪しさに関する数値を視覚的な情報としてユーザーに提示できるか，応用を検討する．生成される顔画像が示す属性，例えば年齢や性別をフィッシングサイトであるかどうかに対応付けて操作する方法が考えられる．しかし，人の属性とフィッシング行為の善悪を対応付けることに繋がるため，好ましくない．同様に，色相や明度を対応付けることも好ましくない．ここで，彩度を対応付けることを考える．例えば，フィッシングサイトであるか確からしさを判別する既存の技術[*2]と組み合わせ，その値によって生成された顔画像の彩度を減少させる．図 9 は怪しいウェブサイトであるとして彩度が大きく下げられた画像の調整例である．他にも，フィッシングサイトである確からしさに応じてぼかし等の画像効果を加える方法が考えられる．

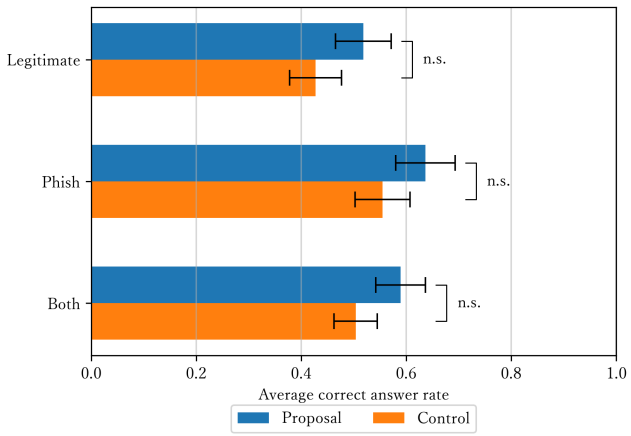
他方で，Allow List や Deny List といった技術との組み合わせを考えた場合，それらの検知技術に許可/拒否されたウェブサイトについては，特定の顔画像を統一して表示することが考えられる．

5.2　提案システムの設計・実装について

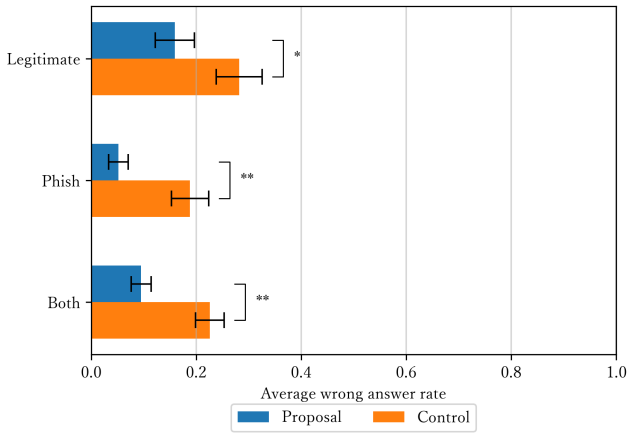
5.2.1　生成された顔画像の共有

提案システムの実装では，特定のドメイン名に対応する顔画像は全てのユーザーであるため，生成済み顔画像のキャッシュを共有することで，画像表示までの時間を短縮

[*2] 今回の実装では GitHub 上に公開されている実装を利用した．`https://github.com/zpettry/AI-Deep-Learning-for-Phishing-URL-Detection`(2021/08/20 現在)



(a) 正規・フィッシング・両方の正答率．有意差はなかった．



(b) 正規・フィッシング・両方の誤答率．
提案群は対照群に比べて全ての項目で優れたスコアとなった．

図 7 提案群と対照群の比較．(n.s.: 有意差なし, *: $p < 0.05$, **: $p < 0.01$)

Fig. 7 Comparison between the proposal and control groups.
(n.s.: no significant differences, *: $p < 0.05$, **: $p < 0.01$.)

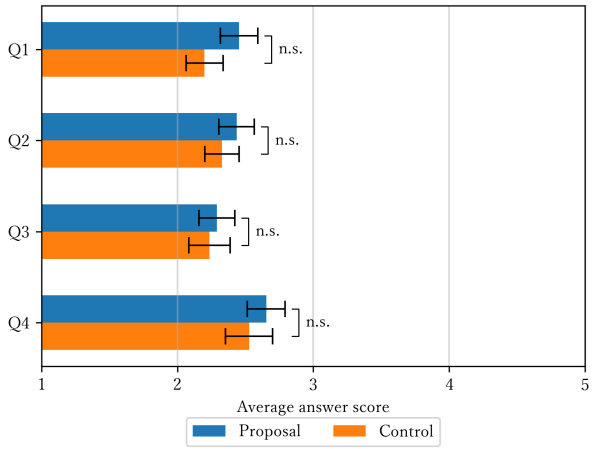


図 8 ユーザー評価の提案群と対照群の比較（n.s.: 有意差なし）

Fig. 8 Comparison of user evaluation between the proposal and control groups. (n.s.: no significant differences.)

図 9 生成された顔画像（左図）とフィッシングの怪しさの度合いに合わせて彩度を調整した画像（右図）

Fig. 9 Generated face image(left) and with saturation adjusted to the level of suspiciousness of phishing(right).

するような応用が可能である．一方で，他のユーザーが一度利用したウェブサイトのキャッシュも機能するため，少数の人数が提案システムを利用していたときに，ユーザーが提案システムを利用した際の画像表示までの時間から，別のユーザーがそのウェブサイトを利用していたという閲覧履歴が間接的に漏洩する可能性が想定される．現在の実装では，生成済みの顔画像の共有はなされないが，画像の共有については今後の課題である．

5.2.2 FQDN ではなくドメイン名を元に画像生成する手法について

提案システムの実装では FQDN を元に顔画像を生成しているが，サブドメイン名を含めずにドメイン名を元に顔画像を生成する手法が考えられる．長所としてはロードバランサーや複数サービスを展開するためにサブドメイン名が変化する場合であっても，システムはブランドに紐付いた共通の顔画像を提示することができ，ユーザーにとって識別が容易になることが考えられる．一方で，短所としてクラウドサービス上にフィッシングサイトを設置されていた際のシステムの利用が考えられる．仮に，正規ウェブサイト `legit.cloud.example.com` とフィッシングサイト `phish.cloud.example.com` に対して，ドメイン名 `cloud.example.com` を元に顔画像の生成がなされると，双方が同一の顔画像となってしまい判別できなくなってしまう．このようなクラウドサービスを利用したフィッシングの調査は今後の課題とし，本研究の現在の実装では FQDN を元に画像生成を行っている．

5.3 ユーザーのばらつきについて

アンケート調査の回答結果を分析すると，提案法を用いても全く判別できず，全ての設問に対してわからないと回答したユーザーが存在した．一方で既存の方法においても，十分な精度で適切に判別できるユーザーも存在する．このようにユーザーの能力にばらつきがあるため，画一的な方法では限界が生じてしまう．ここから，ユーザーの傾向に

合わせた意思決定支援手法の開発や，より適した方法で意思決定支援手法を提案するシステムの提案が考えられる．

5.4 提案システムに対する攻撃について

提案システムに対する攻撃として，攻撃者側が偽の顔画像を提示することでユーザーに正規のウェブサイトであると誤認させることが考えられる．本研究の実装では，ウェブサイトに対応する画像はすべてのユーザーで同一のものであり，攻撃者に対しても同じである．この正規のウェブサイトに対応する画像を攻撃者側がフィッシングサイトで提示することで，ユーザーを騙すという攻撃が考えられる．ただし，提案システムでは拡張機能の領域上で画像を提示しており，フィッシングの攻撃者がこの領域上に干渉することは難しいと考えられる．

6. まとめ

本研究では，偽造されたウェブサイトの判別にあたってのユーザーの意思決定を支援する既存の手法が難解であるとして，ヒトの顔画像を用いてユーザーの意思決定を支援するシステムを提案した．提案システムの実装では，閲覧しているウェブサイトの FQDN を乱数シードとして GAN による学習済みモデルに入力し生成された顔画像を用い，これをユーザーに提示することでウェブサイトと顔画像を対応付けた．また，提案システムの効果を検証するためにアンケート調査を実施し結果の分析をしたところ，提案システムが利便性を損なわずに誤答率を軽減できたことが示された．また，フィッシング検知技術との融合について議論を行い，提案手法の改善及び実装について改善策の検討を行った．

参考文献

[1] Anti-Phishing Working Group (APWG). Phishing attack trends report – fourth quarter 2020, 2021.

[2] Elmer E.H. Lastdrager. Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Science*, Vol. 3, No. 9, pp. 1–16, 2014.

[3] Kaspersky. What is social engineering?, 2017.

[4] Koceilah Rekouche. Early Phishing. pp. 1–9, 2011.

[5] 山崎慎治, 宮本大輔. 顔画像生成技術を用いた偽造ウェブサイト判別支援手法の提案. 暗号と情報セキュリティシンポジウム, 2021.

[6] Min Wu, Robert C. Miller, and Simson L. Garfinkel. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, p. 601–610, New York, NY, USA, 2006. Association for Computing Machinery.

[7] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W. Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. Your attention please: Designing security-decision uis to make genuine risks harder to ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, New York, NY, USA, 2013. Association for Computing Machinery.

[8] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. Improving ssl warnings: Comprehension and adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 2893–2902, New York, NY, USA, 2015. Association for Computing Machinery.

[9] Adrienne Porter Felt, Robert W. Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Emre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking connection security indicators. In *Proceedings of the Twelfth USENIX Conference on Usable Privacy and Security*, SOUPS '16, p. 1–13, USA, 2016. USENIX Association.

[10] Christopher Thompson, Martin Shelton, Emily Stark, Maximilian Walker, Emily Schechter, and Adrienne Porter Felt. The web's identity crisis: Understanding the effectiveness of website identity indicators. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, p. 1715–1732, USA, 2019. USENIX Association.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, Vol. 63, No. 11, pp. 139–144, 2020.

[12] Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. Lexical feature based phishing url detection using online learning. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security*, AISec '10, p. 54–60, New York, NY, USA, 2010. Association for Computing Machinery.

[13] Rami M. Mohammad, Fadi Thabtah, and Lee McCluskey. Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, Vol. 25, No. 2, pp. 443–458, 2014.

[14] Mohammed Nazim Feroz and Susan Mengel. Phishing url detection using url ranking. In *2015 IEEE International Congress on Big Data*, pp. 635–638, 2015.

[15] Ankesh Anand, Kshitij Gorde, Joel Ruben Antony Moniz, Noseong Park, Tanmoy Chakraborty, and Bei-Tseng Chu. Phishing url detection with oversampling based on text generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1168–1177, 2018.

[16] Ahmed AlEroud and George Karabatis. Bypassing detection of url-based phishing attacks using generative adversarial deep neural networks. In *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, IWSPA '20, p. 53–60, New York, NY, USA, 2020. Association for Computing Machinery.

[17] Alexa Internet. Alexa - top sites, 2021.