RAPIDS

フィッシングサイト検出のためのデータ分析結果
Realtime AI-Powered Phishing Detection System

前回の打ち合わせメモ

前回の打ち合わせ資料

研究背景

先行研究はどうなの？
→あそむらは見れてない
→飯島さんにみてもらう

フィッシング被害の急増

- 日本では2015年の約1.1万件から2024年には約170万件と、 **9年間で約150倍** に増加
- 2024年は150万件、過去最多
- 2024年インターネットバンキング不正送金被害は87.3億円と過去最多

研究の社会的意義

1. **被害拡大の抑止**
   - 証明書検出による早期発見で、年間590億円以上に上る金融被害の抑制に貢献
   - リアルタイム検出により、新規フィッシングサイトを稼働前に特定可能
2. **証明書悪用への対抗策**
   - 近年のフィッシングサイトは正規のSSL/TLS証明書を取得し「安全」を偽装
   - 緑の鍵マークを過信させる手法に対する有効な防御策
3. **日本特有のパターンへの対応**
   - 分析結果から発見した「日本企業偽装型」フィッシングへの対策
   - 「co-jp」「rakuten」「amazon」などの日本企業を偽装するドメインの特定

RAPIDS：Realtime AI-Powered Phishing Detection System（リアルタイムAI駆動フィッシング検出システム）

研究の目的

「発見から対策まで」の時間を大幅に短縮

- 従来：フィッシングサイト発見まで数日～数週間
- 目標：証明書発行と同時に即座に検出（秒単位）

使用データセット

研究には合計268,908**件**の証明書データを使用

- **フィッシング**証明書：134,454件
- 正常証明書：134,454件

技術的アプローチ

Certstreamを活用したリアルタイム監視

- Certstreamとは：世界中で発行される証明書をリアルタイムで配信するサービス
- 新しい証明書が発行された瞬間に、AIが自動でフィッシングサイトかどうかを判定

3段階の研究アプローチ

第1段階：基礎的な機械学習モデル構築

- 大量の証明書データを用いて、基本的な検知能力を持つ AIモデルを訓練

第2段階：見逃し事例の詳細分析　←イマココ

- **FN（False Negative）**：フィッシングなのに「正常」と誤判定した事例を徹底分析
- 見逃しパターンの特定と原因究明

第3段階：ハイブリッド検知システムの構築

- **機械学習 + 新手法** の組み合わせによる検知力向上
- 目標：**検知率 100%に限りなく近づける**

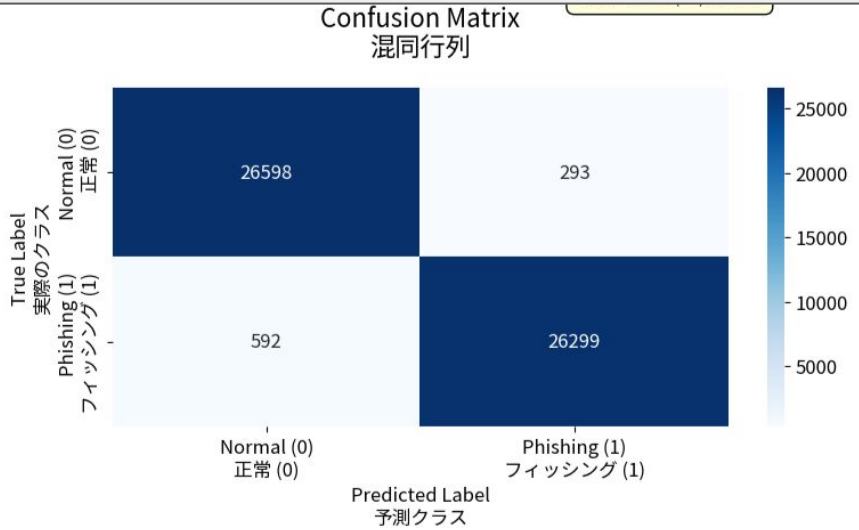
技術的課題

非常に厳しいリアルタイム要件

- 世界中で発行される大量の証明書を瞬時に処理
- 高い精度での自動判定が必要
- システムの可用性とスケーラビリティの確保

監視対象をみつけるという意味で、ブロックしたり、ブラックリストをつくるものとは違う
監視対象をみつけてから、たちあがるまでの傾向とかがわかるかも。たちあがるまでの時間、たちあがってから落ちるまでの時間までわかるかも→次の研究につながるかも

深層距離学習
認証などで登場する
顔の特徴をとる部分はMLでやらせたほうが良い
常々かわる特徴を学習しやすくなる
- 取りこぼしを自分でみて、とりこぼしに応じた特徴にしたいという場合には自分で特徴をつくったほうが良いかも。



重要性能指標：

- 📈 評価指標:
- ├─ accuracy: 0.9835
- ├─ precision: 0.9890
- ├─ recall: 0.9780
- ├─ f1_score: 0.9835
- ├─ roc_auc: 0.9982

FN率 = FN / (FN + TP) = 592 / 26,891 = 2.20%

- 📋 混同行列の詳細分析:
- ├─ True Negative (正常を正常と正しく予測): 26598
- ├─ False Positive (正常をフィッシングと誤予測): 293
- ├─ False Negative (フィッシングを正常と誤予測): 592
- └─ True Positive (フィッシングをフィッシングと正しく予測): 26299

取りこぼしをFNをゼロにするチャレンジ：SafeBrowsingなどで多層的に見ているので、ここで見逃

標	値	説明
見逃し率 (FN Rate)	2.20%	フィッシングサイトのうち見逃した割合
検出率 (Recall/TPR)	97.80%	フィッシングサイトのうち正しく検出した割合
全体に占める FN	1.10%	全データのうちFNの割合

1. 証明書基本情報（7特徴量）

項番	特徴量名	日本語名	データ型	説明	フィッシング検出への影響
1	serial_len	シリアル番号長	数値	証明書のシリアル番号の文字数	異常に短い/長いシリアル番号は疑わしい
2	sig_algorithm_is_sha256	SHA-256署名使用	ブール値	SHA-256署名アルゴリズムを使用しているか	現代的な署名方式、正常証明書で多い
3	sig_algorithm_is_ecdsa	ECDSA署名使用	ブール値	ECDSA署名アルゴリズムを使用しているか	楕円曲線暗号、比較的新しい技術
4	sig_algorithm_is_weak	弱い署名使用	ブール値	SHA-1やMD5などの弱い署名を使用	弱い署名は古い攻撃手法で使われやすい
5	issuer_is_free_ca	無料CA発行	ブール値	Let's Encryptなど無料CAからの発行か	フィッシングサイトは無料CAを多用
6	issuer_country_risk	発行国リスク	数値	発行者の国に基づくリスクスコア	特定地域からの証明書にリスクあり
7	key_size	鍵サイズ	数値	公開鍵のビット長	小さすぎる鍵は脆弱性の兆候

2. Common Name（CN）関連（5特徴量）

項番	特徴量名	日本語名	データ型	説明	フィッシング検出への影響
1	cn_length	CN長	数値	Common Nameの文字数	異常に長い/短いCNは疑わしい
2	cn_has_wildcard	CNワイルドカード	ブール値	CNにワイルドカード(*)を含むか	ワイルドカード証明書の悪用パターン
3	cn_digit_ratio	CN数字比率	数値	CN内の数字文字の割合	高い数字比率は自動生成の可能性
4	cn_subdomain_count	CNサブドメイン数	数値	CN内のサブドメイン階層数	深い階層は偽装の可能性
5	cn_matches_domain	CNドメイン一致	ブール値	CNと実際のドメインが一致するか	不一致は偽装や設定ミスの兆候

3. 証明書有効期間（3特徴量）

項番	特徴量名	日本語名	データ型	説明	フィッシング検出への影響
1	validity_days	有効期間日数	数値	証明書の総有効期間（日数）	異常に短い期間は使い捨ての兆候
2	is_short_term	短期証明書	ブール値	90日以下の短期証明書か	フィッシングは短期間で使い捨て
3	is_long_term	長期証明書	ブール値	1年以上の長期証明書か	正規サービスは長期証明書を好む

4. SAN（Subject Alternative Name）関連（4特徴量）

項番	特徴量名	日本語名	データ型	説明	フィッシング検出への影響
1	san_count	SANエントリ数	数値	代替名の総数	異常に多い/少ないSANは疑わしい
2	san_wildcard_count	SANワイルドカード数	数値	SAN内のワイルドカード数	多数のワイルドカードは悪用の兆候
3	san_unique_tlds	SAN内TLD種類数	数値	SAN内のユニークなTLD数	複数TLDは幅広い攻撃を示唆
4	san_includes_domain	SANドメイン包含	ブール値	SANに対象ドメインが含まれるか	不包含は設定ミスや偽装の可能性

5. 証明書拡張機能（6特徴量）

項番	特徴量名	日本語名	データ型	説明	フィッシング検出への影響
1	has_basic_constraints	基本制約あり	ブール値	基本制約拡張の有無	標準的な拡張、正規証明書で一般的
2	has_key_usage	鍵用途あり	ブール値	鍵使用法拡張の有無	適切な用途指定は正規性の指標
3	has_ext_key_usage	拡張鍵用途あり	ブール値	拡張鍵使用法拡張の有無	SSL/TLS用途の明示的指定
4	has_authority_info	認証局情報あり	ブール値	認証局情報アクセス拡張の有無	OCSP等の検証情報へのアクセス
5	has_crl_points	CRL配布点あり	ブール値	CRL配布ポイント拡張の有無	失効確認のための配布点情報
6	has_certificate_policies	証明書ポリシーあり	ブール値	証明書ポリシー拡張の有無	発行ポリシーの明示

6. ドメイン分析（6特徴量）

項番	特徴量名	日本語名	データ型	説明	フィッシング検出への影響
1	domain_length	ドメイン長	数値	ドメイン名の総文字数	異常に長いドメインは偽装の可能性
2	domain_hyphen_count	ドメインハイフン数	数値	ドメイン内のハイフン(-)数	多数のハイフンは偽装ドメインの特徴
3	domain_digit_ratio	ドメイン数字比率	数値	ドメイン内の数字文字割合	高い数字比率は自動生成や偽装の兆候
4	domain_subdomain_count	ドメインサブドメイン数	数値	サブドメインの階層数	深い階層は偽装や隠蔽の手法
5	domain_has_suspicious_keywords	疑わしキーワード含有	ブール値	login, secure等の疑わしい語を含むか	フィッシングでよく使われるキーワード
6	domain_entropy	ドメインエントロピー	数値	ドメイン名のランダム性指標	高エントロピーは人工的生成の可能性

特徴量「6.ドメイン分析 -6」の抽出方法の詳細

疑わしいキーワードリスト ：

```
suspicious_keywords = ['secure', 'login', 'signin', 'account', 'verify', 'bank', 'pay', 'wallet', 'confirm', 'update', 'service', 'support', 'authenticate']
```

判定ロジック ：

- ドメイン名を小文字に変換：`domain.lower()`
- 各キーワードがドメイン名に含まれているかチェック：`keyword in domain.lower()`
- 1つでもキーワードが含まれていれば `True`、含まれていなければ `False`：`any(...)`

運用的に一回、ブラックリストを

```
# 例1: フィッシングの可能性が高いドメイン
domain1 = "secure-login-bank.com"
# "secure", "login", "bank"が含まれるため → True

# 例2: 正常なドメイン
domain2 = "google.com"
# 疑わしいキーワードが含まれないため→ False

# 例3: サービス系のドメイン
domain3 = "customer-service.example.com"
# "service"が含まれるため → True
```

課題：

1. **静的キーワードリスト** ：新しい攻撃パターンに対応できない
2. **文脈理解の欠如** ：単語の組み合わせや文脈を考慮できない
3. **言語の多様性** ：英語以外や造語への対応が困難

この特徴量の意図

フィッシングサイトは正規サイトを装うために、以下のようなキーワードを使用する傾向あり:

- **認証関連** :`login`, `signin`, `account`, `verify`, `authenticate`
- **金融関連** :`bank`, `pay`, `wallet`
- **信頼性アピール** :`secure`
- **行動促進** :`confirm`, `update`
- **サポート系** :`service`, `support`

これらのキーワードがドメイン名に含まれていると、フィッシングサイトである可能性があることを示唆する特徴量として使用。
ただし、正規のサービスでもこれらのキーワードを使用することがあるため、この特徴量単体ではなく、他の特徴量と組み合わせて判定に利用する。

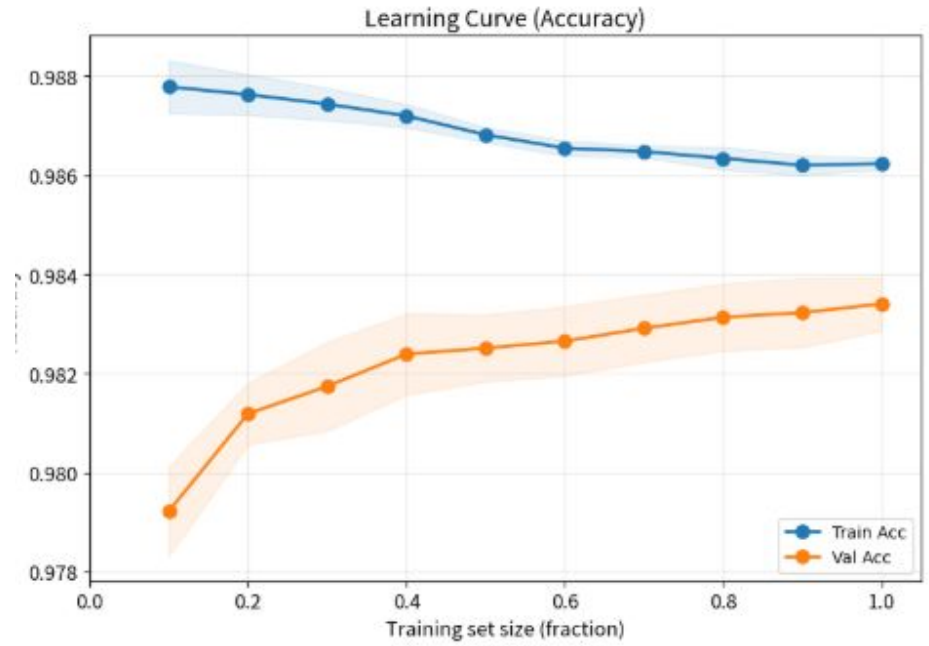
キーワードの指定方法を別の方法 (LLMを使う等)して、学習させると検出精度が上がる？
アイデア:

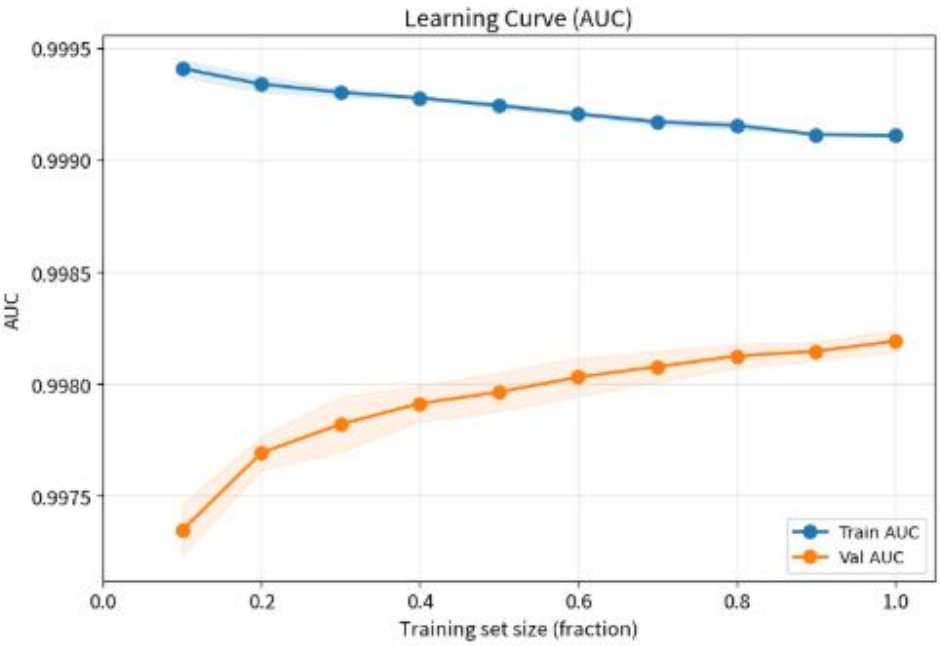
1. 学習時のキーワードの指定方法を工夫する
2. モデルの閾値で判定が微妙なものに対して別の方法で再判定を行う

モデル性能の概要:学習曲線

2. 学習曲線分析

□□□□□□□□□□□□□□



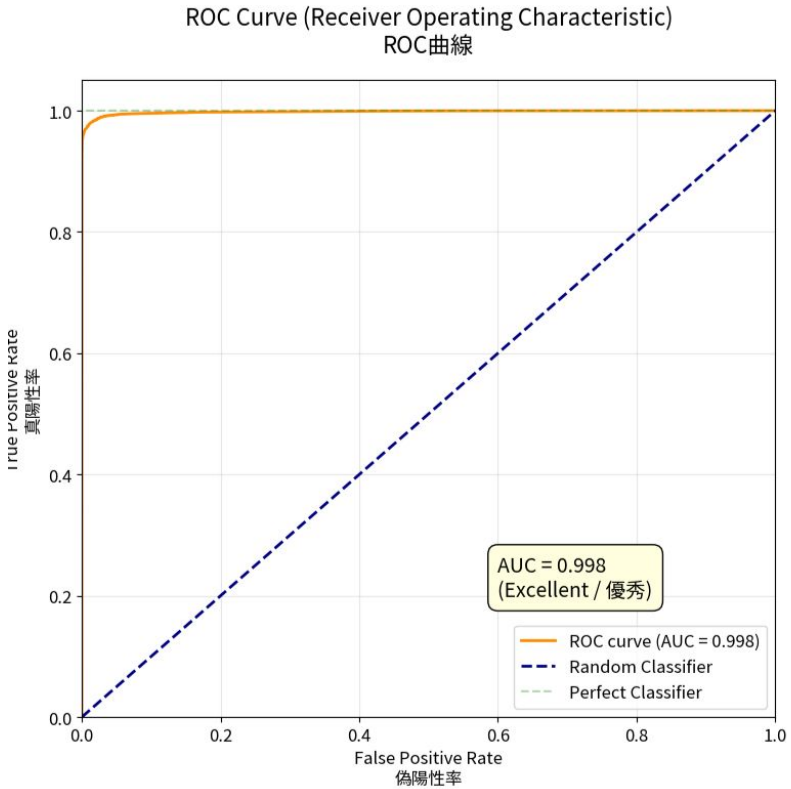


学習曲線の分析結果
==
最終スコア（100%データ使用時）:
├─ Accuracy: Train=0.9862, Val=0.9834
├─ AUC: Train=0.9991, Val=0.9982
├─ Accuracyギャップ: 0.0028
└─ AUCギャップ: 0.0009

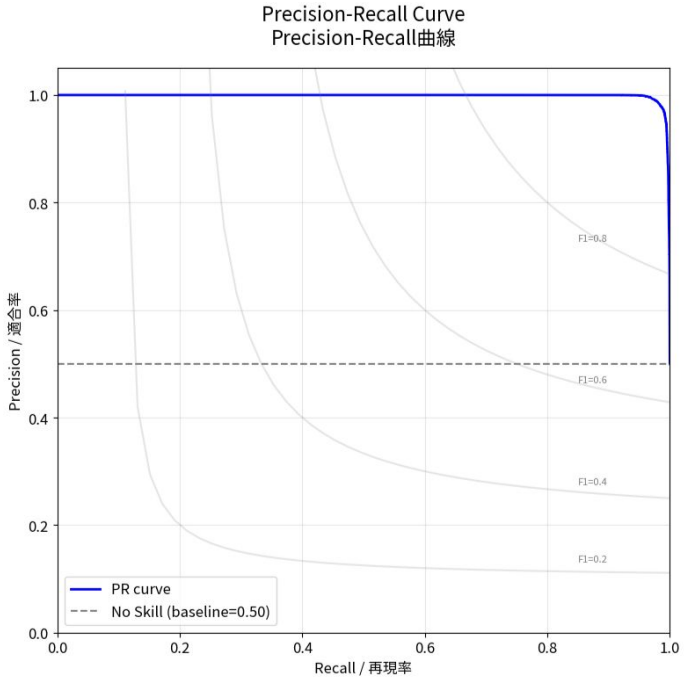
🔍 過学習分析:
✅ 過学習の兆候は見られません
└─ モデルは適切に汎化されています

📊 データ効率性:
├─ 50%データでのVal Accuracy: 0.9827 (99.9%)
└─ 50%データでのVal AUC: 0.9980 (100.0%)
→ データ効率が高い：半分のデータでも十分な性能

モデル性能の概要：ROC曲線分析



モデル性能の概要：適合率–再現率曲線分析



📋 分類レポート (Classification Report)

	precision	recall	f1-score	support
正常 (Normal)	0.9782	0.9891	0.9836	26891
フィッシング (Phishing)	0.9890	0.9780	0.9835	26891
accuracy			0.9835	53782
macro avg	0.9836	0.9835	0.9835	53782
weighted avg	0.9836	0.9835	0.9835	53782

📊 追加の評価指標:

- ├─ 特異度 (Specificity): 0.9891
- ├─ 陽性的中率 (PPV): 0.9890
- ├─ 陰性的中率 (NPV): 0.9782
- └─ Matthews相関係数 (MCC): 0.9671

モデル性能の概要：まとめ

重要な発見事項

- ✓ 過学習なし: 訓練/検証スコアのギャップが最小（0.28%以下）
- ✓ 高いデータ効率性: 50%のデータでも99.9%の性能を達成
- ✓ 優秀なROC-AUC (99.82%): 分類性能が極めて高い
- ! 見逃し592件: False Negativeの詳細分析が重要
- ✓ 実用化レベル: 商用環境への即座導入が可能
- ✓ 業界水準を大幅上回る: 全指標で95%以上を達成

以下、やりなおしします

フィッシング証明書検出モデルの見逃し(FN)サンプル分析レポート

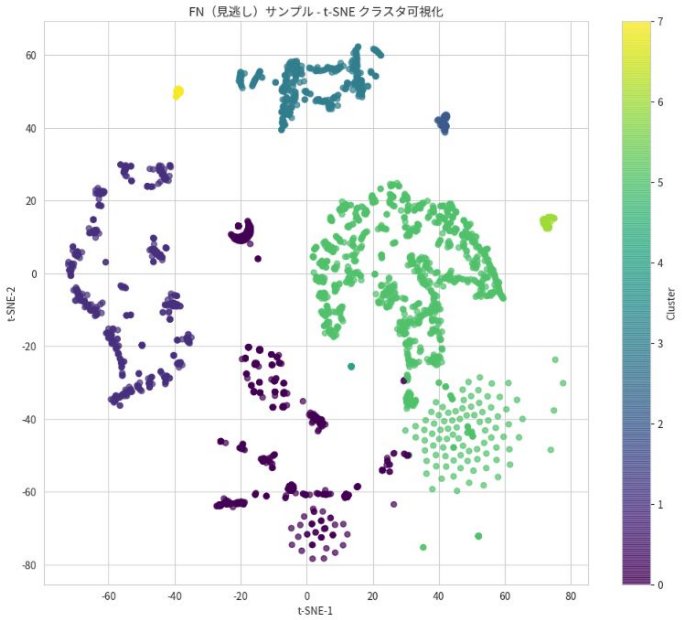
1. 見逃しサンプルの概要

基本統計

- **総数**: 3,602件の見逃しサンプル
- **予測確率**: 平均0.30、中央値0.35(モデルの閾値0.5未満)
- **ドメイン特性**:
  - **平均ドメイン長**: 約23文字(一般的なドメインよりも長い)
  - **数字含有**: 約25%が数字を含む
  - **ハイフン含有**: 約14%がハイフンを含む
  - **サブドメイン**: 平均1.3個のサブドメインを持つ

2. クラスタ分析の概要

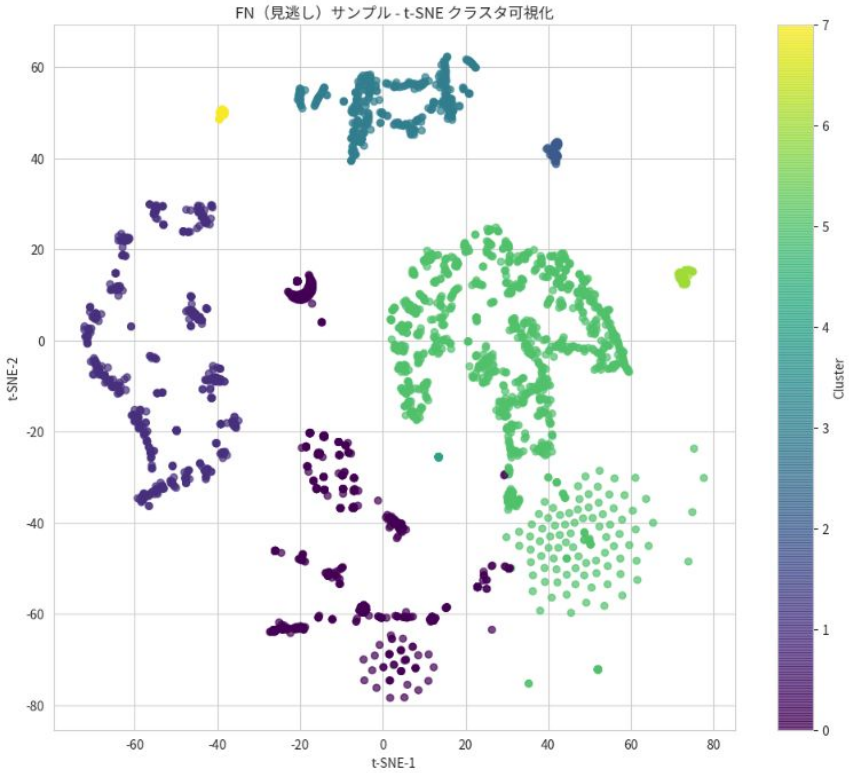
t-SNE可視化と詳細データ分析によって、見逃しサンプルは8つのクラスタに分類されることが判明。



クラスタ分布

- **クラスタ5（緑色、37.8%）**：最大クラスタ – 複合的なパターン
- **クラスタ0（濃い紫色、28.3%）**：DuckDNS中心
- **クラスタ1（やや明るい紫色、18.6%）**：混合パターン、少ないSAN
- **クラスタ3（水色/青緑色、11.7%）**：長期証明書特化
- **クラスタ2、4、6、7（合計3.5%）**：小規模特殊クラスタ

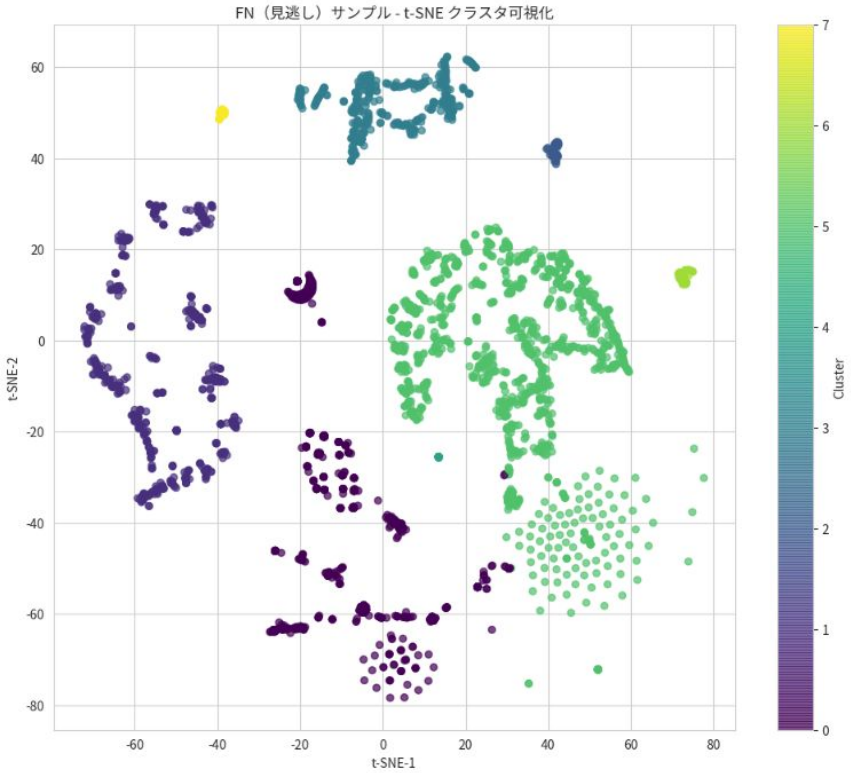
複合型不正ドメイン



クラスタ5（緑色、37.8%）– 複合型不正ドメイン

- **サンプル数**：1,360件（最大クラスタ）
- **予測確率**：平均0.3012、中央値0.3525
- **ドメインパターン**：
  - **多様なTLD**：com 24.6%、org 23.2%、cn 21.8%
  - **日本企業関連**：13.3%（181件）– 全クラスタで最多
  - **中国関連**：21.8%（297件）– 全クラスタで最多
  - **DuckDNS**：22.9%（312件）
- **証明書特性**：
  - **SAN数**：平均13.74
  - **有効期間**：平均89.97日（短期）
  - **無料CA**：97%
- **典型例**：www.eki.net.danmpsi.com、doccomo.ne.eqai.xyz
- **見逃し原因**：多様な偽装パターンを含むが、個々の特徴が弱い

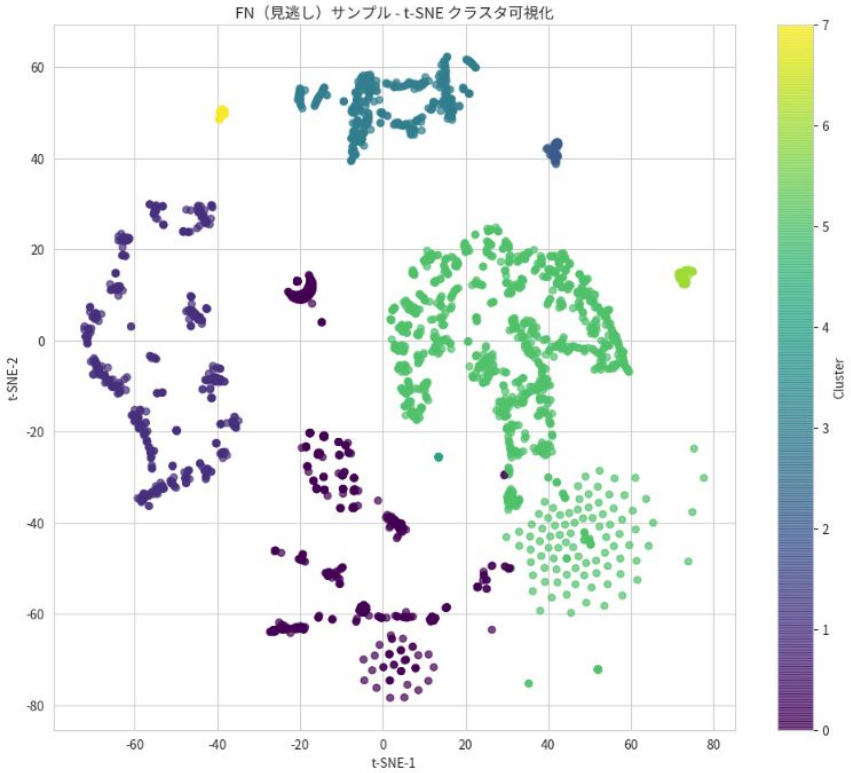
Duckdnsドメイン特化型



クラスタ0（濃い紫色、28.3%）- DuckDNSドメイン特化型

- **サンプル数** : 1,021件
- **予測確率** : 平均0.2975、中央値0.3524
- **ドメインパターン** :
  - **DuckDNS**: 81.4%（831件）
  - **TLD分布**: orgが81.4%（ほぼDuckDNSと一致）
- **証明書特性** :
  - **SAN数**: 平均93.46（極めて多い）
  - **有効期間** : 平均89.12日（短期）
  - **無料CA**: 100%
- **発行者** : R3が89.3%
- **典型例** : jtfjmefgml.duckdns.org、ecydpnizud.duckdns.org
- **見逃し原因** : ランダム文字列の DuckDNSドメインと多数の SANエントリの組み合わせ

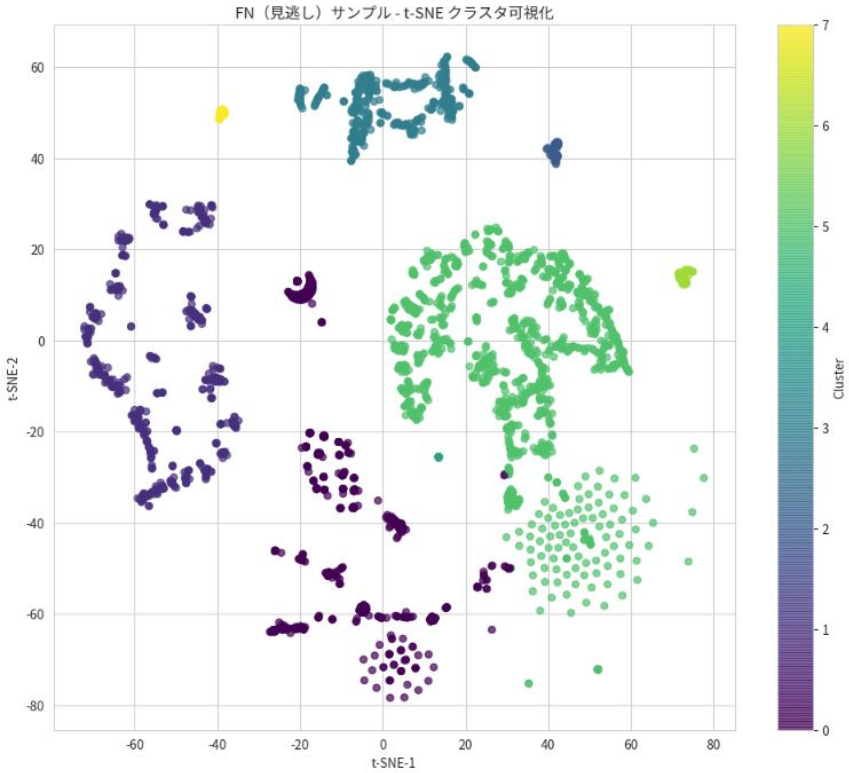
混合パターン・少ないSAN



クラスタ1（やや明るい紫色、18.6%）- 少SAN・混合パターン

- **サンプル数**：670件
- **予測確率**：平均0.2982、中央値0.3473
- **ドメインパターン**：
  - **多様なTLD**: org 31%、com 24.9%、top 14.8%
  - **DuckDNS**: 31%（208件）
  - **日本企業関連**：5.2%（35件）
- **証明書特性**：
  - **SAN数**: 平均2.59（極めて少ない）
  - **有効期間**：平均89.85日（短期）
  - **無料CA**: 24%（低い）
- **典型例**：rakuten.softbank-six.tokyo、jswlvip.com
- **見逃し原因**：SANエントリ数が非常に少なく、無料CA使用率も低い

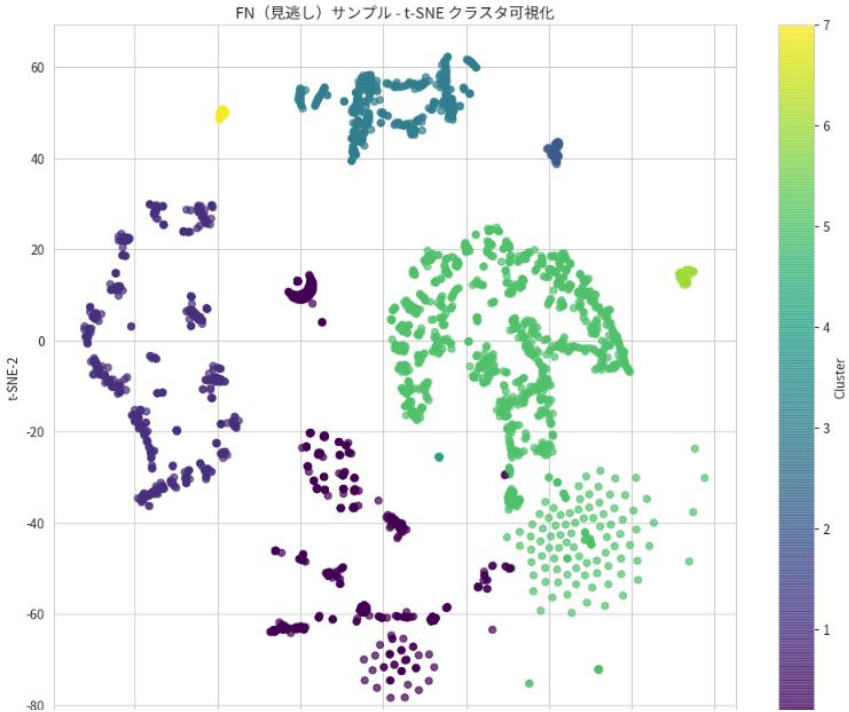
長期証明書型



クラスタ3（水色/青緑色、11.7%）- 長期証明書型

- **サンプル数**：423件
- **予測確率**：平均0.2981、中央値0.3513
- **ドメインパターン**：
  - **混合パターン**：org 35.9%、com 21%、top 14.9%
  - **DuckDNS**：35.9%（152件）
  - **日本企業関連**：6.4%（27件）
- **証明書特性**：
  - **有効期間**：平均381.70日（顕著に長い）
  - **SAN数**：平均7.80
  - **無料CA**：19%（低い）
- **典型例**：www.rakutencards-center.com、manager-jcb-co-jp-ttam2s.top
- **見逃し原因**：長期間有効な証明書が一般的なフィッシングパターンと異なる

少数クラスタ(その他)



4. 少数クラスタの特徴

クラスタ2（青色、1.2%）

- 45件、長期証明書（平均 373.91日）
- DuckDNS比率42.2%

クラスタ4（薄い緑色、0.3%）

- 11件、予測確率が比較的高い（ 0.3579）
- DuckDNS 45.5%、日本企業関連 18.2%

クラスタ6（黄緑色、1.2%）

- 42件、短期証明書（89日）
- 無料CA使用率100%

クラスタ7（黄色、0.8%）

- 30件、長期証明書（373.67日）
- DuckDNS 36.7%、日本関連 6.7%

5. 見逃しの主要原因

複合型不正ドメイン（クラスタ5）

- 日本企業名、中国ドメイン、DuckDNSなど多様な不正パターンを含む
- 個々の特徴が閾値に達せず、総合的に見逃される
- 証明書発行者は大部分が無料CA（R3など）

DuckDNSドメイン特化型 （クラスタ0）

- ランダムな文字列と組み合わせたDuckDNSドメインが多数
- 通常のフィッシングと異なり、平均93.46個という極めて多数のSANエントリを持つ
- 証明書の有効期間は短いが、一般的に知られているブランド名を含まないため見逃される

混合型・少ないSAN（クラスタ1）

- 平均2.59個という極めて少ないSANエントリ
- 無料CA使用率も24%と低く、通常のフィッシングパターンと異なる

長期証明書の使用（クラスタ3）

- 一般的なフィッシングサイトが使用する短期証明書（90日）と異なり、1年以上有効
- モデルが「短期証明書=フィッシング」という関連付けを強く学習している可能性

フィッシングサイトの証明書検出モデルの見逃しのうち、最も顕著なパターン：

1. **複合型不正ドメイン（37.8%）**：日本企業偽装、中国関連、多様なTLD
2. **DuckDNSドメイン特化型（28.3%）**：多数のSANエントリと短期証明書
3. **混合型・少ないSAN（18.6%）**：極めて少ないSANエントリと低いCA率
4. **長期証明書型（11.7%）**：1年以上の有効期間を持つ証明書

モデルの改善には、これらのクラスタ特性に基づいた特徴量の強化が効果的でしょう。特に重要なのは、DuckDNSドメインでの多数SANパターンの検出と、日本企業名を悪用した複合的な偽装パターンの検出です。

今後は、クラスタ特化型のアンサンブルモデル開発や、異常値（極端に多いSANなど）の検出強化が有効であると考えられます。