機械学習とAIエージェントを組み合わせたハイブリッド型フィッシングサイト検出システムの開発と評価

阿曽村 一郎[1,a)]　戸田 宇亮[1,2]　飯島 涼[3,1]　森 達哉[1,2,4]

概要：
フィッシング被害が年間 86.9 億円に達する中，従来の機械学習（ML）による検出手法は高速処理が可能な一方で，判定が困難な境界領域では見逃しが発生するという本質的な課題を抱えていた．本研究では，この課題に対して，ML と大規模言語モデル（LLM）をベースとした AI エージェントの相補的な強みを活かしたハイブリッド型検出システムを提案する．提案システムは 2 段階の処理で構成される．第一段階では ML モデルが全ドメインを高速にスクリーニングし，フィッシングサイトであるか否かを判定する．第二段階では，ML モデルの判定確信度が低いドメインに対し，LLM ベースの AI エージェントが詳細分析を行い，ML モデルの識別境界を超える潜在パターンを検出する．この二段階設計により，ML の処理効率を維持しつつ，LLM の高度な分析能力を必要な領域に限定的に適用し，実用性と精度の両立を実現した．640,356 件のデータセットを用いた評価実験では，XGBoost モデル（精度 95.70%）によるフィッシング判定において，確信度スコアが低かった 4,215 件に対して AI エージェントを適用した結果，84.8%のフィッシングサイトを正しく検出し，システム全体の偽陰性率を 6.58%から 1.04%へと大幅に削減した．

キーワード：フィッシング検出, AI エージェント, 偽陰性分析, 特徴抽出, LangGraph

Development and Evaluation of Hybrid Phishing Detection System Combining Machine Learning and AI Agents

Ichiro Asomura[1,a)]　Toda Takaaki[1,2]　Ryo Iijima[3,1]　Mori Tatsuya[1,2,4]

Abstract:
While phishing damages have reached 8.69 billion yen annually, conventional machine learning (ML)-based detection methods, although capable of high-speed processing, suffer from an inherent issue: missed detections in ambiguous boundary regions. To address this problem, this study proposes a hybrid detection system that leverages the complementary strengths of ML and AI agents based on large language models (LLMs). The proposed system consists of a two-stage process.
In the first stage, the ML model rapidly screens all domains and outputs the predicted probability of each domain being a phishing site. In the second stage, the LLM-based AI agent performs detailed analysis only on domains that the ML model missed as false negatives, detecting patterns that ML alone cannot capture. This design maintains the processing efficiency of ML while selectively applying the advanced analytical capabilities of LLMs only where needed, achieving both practicality and accuracy.
In evaluation experiments using a dataset of 640,356 samples, we applied the AI agent to 4,215 cases that the XGBoost model (95.70%) classified with low confidence. As a result, the agent successfully detected an additional 84.8%, significantly reducing the system's overall false negative rate from 6.58% to 1.04%.

Keywords: phishing detection, AI agent, false negative analysis, feature extraction, LangGraph

1. はじめに

日本ではフィッシング攻撃による被害額が年間 86.9 億円に達し，2024 年には報告件数が 1,718,036 件に上った．一方，世界全体では SSL/TLS の普及に伴いフィッシングサイトの HTTPS 化が 83%以上に達しており [1],「鍵マークがあれば安全」という従来の判断基準はもはや有効ではない．この状況を踏まえ，フィッシングサイト検出技術は依然として重要な研究課題である．

従来研究では機械学習（ML）モデルの改良や大規模言語モデル（LLM）の活用により高精度な検出手法が提案されてきた．機械学習では特徴量エンジニアリング，アンサンブル学習，深層学習などの手法により偽陰性と偽陽性の削減が図られ，LLM を用いた手法 [2] では高度な文脈理解による複雑なパターン認識が可能となった．しかし，個々の誤判定が生じる要因の体系的分析は不十分であり，特に機械学習モデルが高確信度で正常と判定したドメインに潜むフィッシングサイトの検出には課題が残る．また，LLM 単体での処理は推論時間（平均 20 秒/ページ）や API コストが実装上の制約となる．

本研究では，軽量で高精度な性能を得られる ML として，XGBoost による高速スクリーニングと LLM ベースの AI エージェントによる精密分析を組み合わせた二段階ハイブリッドアーキテクチャを提案する．第一段階では ML モデルが全ドメインを高速にスクリーニングし，フィッシングサイトであるかを判定する．第二段階では，ML モデルによるフィッシング判定の確信度スコアが低いドメインに対し，AI エージェントがブランド偽装検出，証明書分析，短いドメイン名分析，文脈的リスク評価の四つの専門ツールで詳細分析を行う．特に ML による判定確信度が低い領域に潜むフィッシングサイトの検出に焦点を当てる．

640,356 件のデータセットを用いた評価実験では，XGBoost モデル（精度 95.70%，ROC-AUC 0.9894）の判定において確信度が低い 4,215 件（$p < 0.5$）に対して AI エージェントを適用した結果，3,576 件（84.8%）のフィッシングサイトを正しく検出し，システム全体の偽陰性率を 6.58%から 1.04%へ低減した．また，特にフィッシングサイトとしての判定の確信度が低い ($p < 0.2$)，すなわちどちらかというと正常サイトである可能性が高いと判定された 2,196 件においてフィッシングサイトの検出率 100%を達成した．これらの結果は，ML によって正常判定の可能性が高いと判定サイトにも攻撃が含まれ得ることを示唆するものである．さらに，AI エージェントは各判定に対して証明書の組織情報の欠如やブランド名と TLD の不整合などの証拠を提示することで説明可能性を確保できること，および平均処理時間 2.52 秒/ドメインであり，大規模データへの適用が可能であることを明らかにした．

本論文の構成は以下のとおりである．第 2 章では関連研究として既存のフィッシング検出手法を整理する．第 3 章では提案手法の詳細を示し，誤判定要因の分析と判定補完の仕組みを説明する．第 4 章では実験設定と評価結果を示し，偽陰性の改善効果を定量的に評価する．第 5 章では考察として抽出された特徴パターンの意味と将来のモデル改善への応用を議論する．最後に第 6 章でまとめと今後の課題を述べる．

2. 関連研究

2.1 機械学習によるフィッシング検出

フィッシング検出における機械学習の応用は過去 10 年間で急速に発展したが，多くは精度向上に焦点を当て，特定サイトが見逃される理由の分析は不十分であった．

Sahingoz ら [3] は，Random Forest，SVM，ニューラルネットワークなど 7 種のアルゴリズムを比較し，Random Forest が 97.98%の精度を達成したと報告したが，偽陰性の詳細な分析はない．

近年は深層学習の活用も進み，Alshingiti[9] は LSTM で 96.8%の精度を達成したが，偽陰性パターンの抽出は行われなかった．

これらに共通する課題はモデルの「ブラックボックス」性であり，高精度でも誤判定の原因が不明なままである．本研究はこの点に対し，偽陰性の原因分析を試みる．

2.2 大規模言語モデルを活用したフィッシング検出

大規模言語モデル（LLM）の登場により，フィッシング検出は新たな展開を迎えた．LLM は文脈理解に優れ，従来手法では捉えにくい複雑なパターンを認識できる．

Koide ら [2] は GPT-4V を用い，98.7%の精度と 1.3%の偽陰性率を達成したが，20 秒/ページの処理時間が実用化の障壁であり，偽陰性の特徴分析も限定的だった．

Liu ら [6] は LLaMA-2 をファインチューニングし，URL と HTML を用いて 99.53%の精度を達成したが，全ドメインに LLM を適用する設計でコストが高く，スケーラビリティや説明性に課題が残る．

PhishLang ら [7] は複数の LLM をアンサンブルし精度を向上させたが，計算コストが増大し実用性を欠いた．総じて LLM は高精度を示す一方，偽陰性の原因分析は不足している．

2.3 ハイブリッド手法

機械学習と LLM を統合するハイブリッド手法は有望だが，多くは単純な組み合わせに留まり，偽陰性の分析視点は欠けている．

[1] 早稲田大学/Waseda University
[2] 理化学研究所 革新的知能統合研究センター/RIKEN AIP
[3] 産業技術総合研究所/AIST
[4] 情報通信研究機構/NICT
[a)] asomura@nsl.cs.waseda.ac.jp

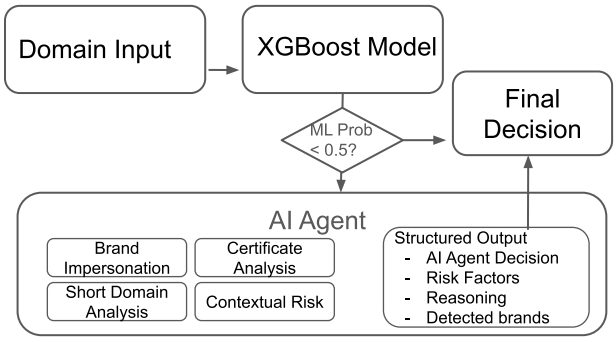

図 1 提案する二段階ハイブリッド型フィッシング検出システムのアーキテクチャ．第一層の XGBoost による高速スクリーニングと第二層の LLM エージェントによる精密分析の役割分担を示す．

Connolly ら [8] はランダムフォレストと BERT を組み合わせ 99.53%の精度を達成し，初期フィルタ後に境界事例のみ BERT を適用したが，目的は精度向上に限定されていた．

2.4 本研究の位置づけ

既存研究は精度向上という「結果」に主眼を置いてきた．これに対し，本研究は機械学習モデルが高い判定確信度で正常と判定したドメインに着目し，その誤判定の「原因」を解明することを狙いとする．本研究の貢献は，第一に，各判定に対する根拠を明示して説明可能性を高め，将来のモデル改善に資する知見を提供する点，第二に，偽陰性率を 6.58%から 1.04%に低減し，とくに判定確信度 < 0.2 の領域で検出率 100%を達成する点にある．本研究は，偽陰性率の削減にとどまらず，低い判定確信度領域に潜む巧妙な攻撃パターンの分析を通じて継続的改善を促す枠組みを提示し，フィッシング検出研究における焦点を「精度向上の追求」から「機械学習モデルの盲点の理解と補完」へと転換することを目指す．

3. 設計と実装

3.1 全体的な設計

従来のフィッシングサイト検出研究は，主に ML モデルの精度向上に焦点を当ててきた．しかし，多くの研究では「偽陰性にどのようなケースが含まれているのか」に対する詳細な分析が不足しており，ML モデルの判定確信度が低いにもかかわらず，実際にはフィッシングであるケースが見逃される課題が残されている．

本研究ではこの課題に対処するため，二段階構成のアーキテクチャを設計した（**図 1**）．

第一層では XGBoost により大量のドメインを高速にスクリーニングし，第二層では LLM エージェントが「判定確信度は低いが実際には危険である可能性が高いドメイン」を重点的に精密分析する．これにより，従来見逃されがちであった偽陰性を補完的に評価できる．

特に本研究が注目するのは「判定確信度が低い領域」である．これは ML モデルの確率出力が 0.2 未満となる領域であり，XGBoost による偽陰性サンプルの分析から，この領域にフィッシングサイトが集中する傾向が観察された．本論文ではこの課題を低確信度領域問題（Low-Confidence Region Problem）と呼び，第二層の重点的な解析対象とする．

本研究における設計の方針は，次の二点を目的としている．第一にスケーラビリティの実現を狙い，膨大なドメインを効率的に処理できる高速性を持たせる．第二に精密性の確保を目指し，低確信度領域に潜むフィッシングサイトを取りこぼさない精度を実現する．

3.2 第一層：XGBoost による高速スクリーニング

第一層は大量のドメインを効率的にスクリーニングし，疑わしい候補を抽出する役割を担う．本研究では勾配ブースティング決定木アルゴリズムである XGBoost[10] を用い，設計した特徴量群を入力として高速かつ高精度な初期分類を実現している．

ここで利用する特徴量は大きく 3 種類に分かれる：

- ドメイン構造に基づく特徴量（15 個）
- TLS 証明書に基づく特徴量（5 個）
- ブランドに基づく特徴量（動的生成による 1 個）

ドメイン構造に関する 15 個の特徴量は，サブドメイン数，ドット数，ドメイン長，最長部分文字列長，シャノンエントロピー，数字比率，母音比率，最大連続子音長などで構成される．これらは Feature Importance の分析により主要な判別要因であることが確認されている．

以下では証明書特徴量とブランド特徴量の設計について詳述する．

3.2.1 TLS 証明書に基づく特徴量

証明書は正規サイトとフィッシングサイトを区別する上で重要な補助情報である．本研究では以下の 5 つの特徴量を設計した：

- 証明書有効日数：証明書の有効期間．短期間の証明書利用はフィッシングに多い傾向がある．
- 発行者名の長さ：正規認証局と無料 CA の区別に寄与する．
- ワイルドカード証明書の有無
- SAN 数：Subject Alternative Names の数
- 自己署名証明書の有無

これらによりドメイン構造だけでは捉えにくい証明書の特徴を数値化し，分類に活用する．

3.2.2 ブランドに基づく特徴量

フィッシング攻撃の標的ブランドは時間とともに変動する．金融機関が集中的に狙われる時期もあれば，EC サイトや通信サービスが主な標的となる時期もある．従来の固

定的なブランドリストではこの変化に対応できない．

そこで本研究では動的ブランドキーワード生成機構を導入した．処理の流れは以下の通りである：

(1) **データ収集**：JPCERT/CC のレポートおよび PhishTank データベースからブランド名を抽出する．

(2) **正規化**：GPT-4o-mini を用い表記ゆれを統一する（例：「Bank of America Corporation」→「bofa」，「楽天」→「rakuten」）．

(3) **選別**：出現頻度が一定以上のブランドを選定し，最終的に 64 個のブランドキーワードを生成する．

得られたキーワードはブランド名含有という特徴量に変換され，ドメイン名に標的ブランドが含まれるか否かを二値で表す．これによりブランド偽装型フィッシングの検出力が向上する．

3.3 第二層：精密分析層の内部構成

第二層は第一層で XGBoost による判定の確信度が 0.5 未満であったドメインのうち，偽陰性の可能性が高いものを精密に再評価する．単一モデルでは扱いきれない判定確信度が低い領域を多角的な分析で補完することが設計思想である．LangGraph と LangChain を組み合わせ，LLM エージェントが複数の専門ツールを統合的に活用することで透明性と拡張性を備えた検出プロセスを実現している．

本研究ではこの LLM エージェントを PhishingDetectionAgent と呼ぶ．入力として与えられるドメイン情報をもとに適切なツールを選択・実行し，結果を統合する．

制御フロー：PhishingDetectionAgent は以下の流れで処理を行う：

(1) **ツール選択**：判定確信度に応じて使用ツールを決定（特に 0.2 未満では全ツールを強制実行）．

(2) **ツール実行**：選択されたツールを並列に実行．

(3) **結果統合**：各ツールの出力を集約し最終判定を生成．

状態管理：状態は「ドメイン」「判定確信度」「ツール結果」の 3 要素で管理される．各ツールの出力は独立して保持され，後段で動的に参照可能である．この仕組みによりツールの追加や削除にも柔軟に対応できる．

並列ツール実行：LangChain の Tool Calling 機能により 4 つの専門ツールを並列に実行し，異なる視点から同一ドメインを多面的に評価する：

- **ブランド偽装チェック**：ドメイン文字列と生成されたブランドキーワードを照合しブランド偽装の有無を判定する．1 文字置換・削除・追加といったタイポスクワッティングも検出対象とする．
- **証明書解析**：TLS 証明書の発行者，組織名，有効期間を解析し，正規認証局と無料 CA の差異や自己署名証明書の乱用を検出する．
- **短いドメイン名解析**：ドメインの長さ，シャノンエントロピー，TLD（Top-Level Domain）を指標に異常なパターンを検出する．
- **状況リスク評価**：上記 3 ツールの結果を統合し，単独では弱いシグナルを組み合わせて判定確信度が低い領域のリスクを顕在化させる複合リスク評価を行う．

Structured Output による統合：各ツールの出力は以下の統一フォーマットに変換され，判定理由の透明性と再現性を確保している：

- `is_phishing`（boolean）
- `confidence`（0–1）
- `risk_factors`（dict）
- `reasoning`（string）

第一層との役割分担：第二層は LangGraph と LangChain を活用した LLM エージェントにより多角的で柔軟な分析を実現する．第一層が「高速スクリーニング」により大量データを処理するのに対し，第二層は「精密分析」により判定確信度が低い領域を重点的に精査する．この明確な役割分担によりスケーラビリティと精密性の両立が可能となっている．

図 2 XGBoost モデルの学習曲線

4. 評価

4.1 第一層：XGBoost モデルの性能評価

4.1.1 学習条件と訓練過程

XGBoost モデルの訓練には以下のハイパーパラメータを使用した：

- **最大深さ** 8
- **学習率** 0.1
- **早期停止** 20 イテレーション
- **早期停止**: 20 ラウンドの改善なしで停止, 296 イテレーションで収束

訓練は 512,284 件のデータに対して実施された．Early Stopping により過学習を防ぎつつ, 最適な複雑さのモデルを獲得した．

5-fold 交差検証による ROC-AUC は **0.9889 (± 0.0001)** となり, モデルの安定性と汎化性能が確認された．

図 2 に, 学習過程における Learning Curve を示す．訓練誤差と検証誤差が収束しており, 過学習が生じていないことが確認できる．

4.1.2 全体的な性能指標

128,072 件のテストデータに対する評価結果を**表 1** に

表 1　XGBoost モデルの性能指標

指標	値	説明
精度（Accuracy）	95.70%	全体的な正解率
適合率（Precision）	97.89%	フィッシング判定の正確性
再現率（Recall）	93.42%	フィッシングサイトの検出率
F1 スコア	0.9560	適合率と再現率の調和平均
ROC-AUC	0.9894	識別能力の総合指標
偽陽性率（FPR）	2.02%	正常サイトを誤検出した割合
偽陰性率（FNR）	6.58%	フィッシングを見逃した割合

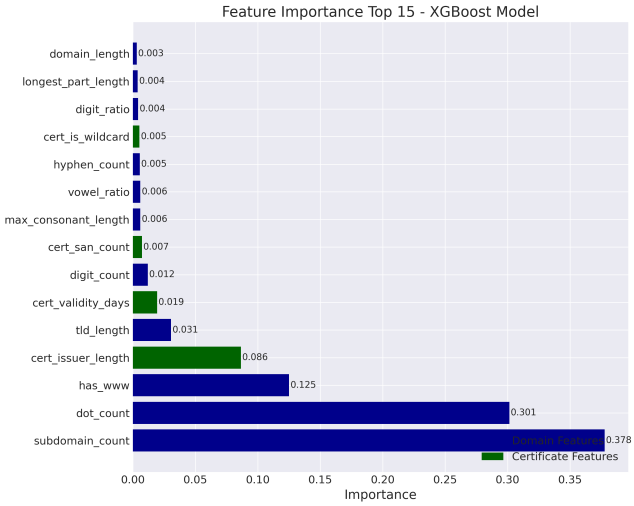

図 3　XGBoost モデルの特徴量重要度

示す.

本モデルは精度 95.70%と高い性能を達成したが，偽陰性率が 6.58%に達し，件数にして 4,215 件のフィッシングサイトを見逃している．この偽陰性は実運用上，無視できない課題となる.

4.1.3　特徴量の重要度分析

XGBoost が学習した特徴量の重要度を分析した結果, ドメイン構造に関する特徴量が上位を占めた（**図 3**）.

上位 5 つの重要特徴量：

(1) subdomain_count (0.378)：サブドメイン数が最も重要
(2) dot_count（0.301）：ドット数による階層構造の複雑さ
(3) has_www（0.125）：www プレフィックスの有無
(4) cert_issuer_length（0.086）：証明書発行者名の長さ
(5) tld_length（0.031）：TLD の文字列長

ドメイン特徴量の重要度合計は 0.882 に達し, 証明書特徴量は 0.118 に留まった．これは, 攻撃者が正規の証明書を容易に取得できる現状を反映している可能性がある.

4.1.4　偽陰性の詳細分析

検出に失敗した 4,215 件の偽陰性について, ML 判定確信度の分布を分析した結果を**表 2** に示す.

特に注目すべきは, **52.1%（2,196 件）の偽陰性が ML 判定確信度 0.2 未満**に集中している点である．これらは XGBoost が高い確信度で「正常」と誤判定したケースであり, 以下の特徴を持つ：

表 2　ML 判定確信度別の偽陰性分布

予測確率範囲	件数	割合	累積割合
0.00–0.05	617	14.6%	14.6%
0.05–0.10	528	12.5%	27.1%
0.10–0.20	1,051	24.9%	52.1%
0.20–0.30	736	17.5%	69.6%
0.30–0.40	640	15.2%	84.8%
0.40–0.50	643	15.2%	100.0%

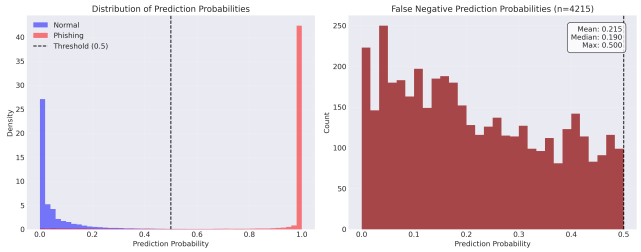

図 4　偽陰性の予測確率分布

- ドメイン長: 平均 10.9 文字（正常サイトに類似）
- TLD 分布: .com（59.3%）, .cn（5.6%）, .top（2.5%）が上位
- データソース別: PhishTank 由来で 24.1%と高い偽陰性率
- 証明書: 多くが Let's Encrypt 等の無料 CA を使用

図 4 に, 偽陰性の予測分布を示す．低確率領域に偏っていることが確認できる.

4.1.5　性能限界と改善の必要性

XGBoost モデルは高い精度を示したものの，次のような限界が明らかになった．とくに，判定確信度 < 0.2 の低確率領域で 2,196 件の見逃しが発生した．また，正規サイトに酷似した構造を持つフィッシングに対して脆弱で，誤判定が生じた．さらに，無料 CA の利用率が高い環境では証明書情報の寄与が限定的で，識別性能の向上に結び付かなかった.

これらは，統計的パターンに依存する機械学習モデルの構造的制約を示す．攻撃者が意図的に正常サイトを模倣する場合，XGBoost 単体での検出は困難である.

一方，過学習の検証では AUC のギャップは 0.0009 と小さく，モデルの汎化性能は良好であった．したがって，4,215 件の偽陰性の存在は，ML モデルの識別境界を超える潜在パターンが存在することを示唆する.

結論として，XGBoost の高速性と基本的検出能力を第一層として活用し，第二層では AI エージェントによる二次判定で補完する構成が，実運用における偽陰性リスクの低減に有効である.

4.2　第二層: 精密分析層の評価

4.2.1　エージェントの解析プロセスの事例

本研究では，偽陰性 4,215 件の中から代表的な 3 事例を抽

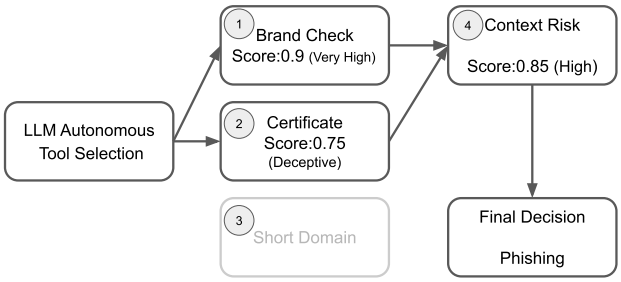

図 5 paypal-secure-login.info に対する AI エージェントの解析トレース．LLM が動的な選択により 3 つのツールを選択的に実行し, 処理時間を 25%削減しながら高リスクと判定.

出した．選定はランダムではなく, 判定確信度 (0.01–0.26), 特徴パターン（ブランド偽装・証明書異常・地域偽装), AI エージェント信頼度（0.70–0.85）を総合的に考慮した．選定の視点は，偽陰性に共通する主要パターンを代表する典型性，異なる攻撃手法や対象を網羅する多様性，容易なケースから巧妙なケースまで段階的に配置する難易度階層のいずれか，および／または複数を満たすこととした．

これにより，3 事例は偽陰性の特性を示すものとなっている．

事例 1: paypal-secure-login.info (金融ブランド偽装型)

- 判定確信度: 0.342 / AI 信頼度: 0.825 / リスクレベル: High

このケースは, PayPal を模倣した典型的な金融系フィッシングサイトである．ドメイン名には「paypal」「secure」「login」という正規サービスを連想させるキーワードを組み合わせているが, 実際の PayPal とは無関係である．予測確率 0.342 は中間的な値であり, 機械学習モデルは判定に迷いを示した.

LLM 動的選択プロセス (判定確信度 $>$ 0.2):

エージェントは予測確率が 0.2 を超えるため, 選択的ツール実行モードを採用した．4 つの利用可能ツールから 3 つを選択した (**図 5**):

LLM 動的選択プロセス（予測確率 $>$ 0.2)：

エージェントは予測確率が 0.2 を超えるため, 選択的ツール実行モードを採用した．4 つの利用可能ツールから 3 つを選択した（図 5)：

✓ ブランドチェック (実行)
✓ 証明書解析（実行）
× 短いドメイン解析（スキップ - 20 文字で正常範囲）
✓ 状況リスク評価（実行）

この結果, 処理時間を 25%削減しながら検出精度を維持し, 中間確率帯における動的ツール選択の有効性を確認できた.

事例 2: my-ledger-secure.com (暗号資産ウォレット偽装型)

- 予測確率: 0.016 / AI 信頼度: 0.85 / リスクレベル: High

Ledger を偽装したケースで, モデルは 0.016 と極めて低い予測確率を出力し, 正規サイトと誤判定した．エージェントは以下を根拠に高リスクと判定した:

- 証明書に組織情報が欠如 (正規 Ledger 社では必ず存在)
- 「secure」という高リスクキーワードを検出
- 0.02 以下の判定確信度と実際のリスクの矛盾を認識
- Ledger のドメイン命名規則からの逸脱

特筆すべきは, 異常に低い予測確率そのものを「偽陰性の兆候」と検知できた点である.

事例 3: mercari.buzz (EC サイト偽装・地域ターゲット型)

- 予測確率: 0.258 / AI 信頼度: 0.65 / リスクレベル: Medium-High

メルカリを偽装した日本市場向け攻撃である．モデルは 0.258 という境界的な判定確信度を出力し, 陰性と判定した．エージェントは以下を評価した:

- .buzz TLD の異常性 (正規は .com / .jp)
- ブランド名「mercari」の一致と, 正規ドメイン構造との不一致
- 証明書の組織情報欠如, 信頼性スコア 0.35
- 地域的文脈の不自然さ (日本市場向けで不自然な TLD 利用)

これにより, モデルの境界的判定を文脈的理解で補完し, Medium-High リスクと判定した.

4.2.2 精密分析層による偽陰性削減効果

精密分析層の性能を定量的に評価するため，第一層の XGBoost モデルが見逃した 4,215 件の偽陰性データ全件に対して LLM エージェントによる再検証を実施した.

4.2.2.1 偽陰性削減の定量的評価

4,215 件の偽陰性データに対して LLM エージェントを適用した結果，3,576 件（84.8%）をフィッシングサイトとして正しく検出することに成功した．これにより，システム全体の偽陰性率は 6.58%から 1.00%へと 5.58 ポイント削減され，偽陰性の 85%を排除できた特に注目すべきは，XGBoost モデルが 0.2 未満の極めて低い判定確信度を出力した 2,196 件について，エージェントが 100%（2,196 件全て）を正しくフィッシングと判定したことである.

4.2.2.2 判定確信度別の検出性能

エージェントの検出率は判定確信度の範囲によって異なる特性を示した．判定確信度 0.0-0.1 の範囲では 100%（1,145 件/1,145 件)，0.1-0.2 の範囲でも 100%（1,051 件/1,051 件）の完璧な検出率を達成した．判定確信度 0.2-0.3 では 82.1%（604 件/736 件)，0.3-0.4 では 52.5%（336 件/640 件)，0.4-

0.5 では 68.4%（440 件/643 件）と，判定確信度が中間領域に近づくにつれて検出率が変動する傾向が見られた．この結果は，エージェントが特に低判定確信度領域（0.2 未満）において機械学習モデルとは異なる判断基準を持ち，相補的に機能することを示している．

4.2.2.3 信頼度スコアの分布

エージェントが出力する信頼度スコアの平均は 68.8%（標準偏差 15.4%）であった．フィッシングと判定したケースの信頼度は平均 73.3%，非フィッシングと判定したケースは平均 43.2%となり，正しい判定ほど高い信頼度を示す傾向が確認された．信頼度スコアは，high（2,146 件），medium（1,521 件），medium-high（482 件），critical（56 件），low（10 件）の 5 段階のリスクレベルに分類され，より詳細なリスク評価を可能にした．

4.2.2.4 システム全体の性能向上

第一層と第二層を組み合わせたハイブリッドシステム全体の性能は，精度 99.00%，適合率 97.91%，再現率 99.00%，F1 スコア 0.9845 となった．これは第一層単体と比較して，精度が 3.30 ポイント，再現率が 5.58 ポイント向上したことを意味する．特に再現率の大幅な改善により，実運用において重大なリスクとなる見逃しを最小限に抑えることができた．

4.2.3 残された課題と今後の展望

本研究により, 第一層の機械学習モデルと第二層の LLM エージェントを組み合わせることで, 偽陰性率を大幅に低減できることを示した．しかし, 依然としていくつかの課題が残されている．

(1) 境界的ケースの扱い toogpraat.com のように, 機械学習モデルでは「リスク指標が検出されず安全」と判定されつつ, 実際にはオンラインカジノ等のグレーゾーンに位置するサイトが存在する．これらのケースは明確な悪性指標に欠けるため, 現行の二層システムでは扱いが難しく, 追加的なコンテキスト理解や外部データ参照が求められる．

(2) 過去の悪用履歴を持つドメイン一部のドメインは, 過去にはフィッシングやマルウェア配布に利用されながら, 現在は正常な企業サイトとして運用されている場合がある．このようなケースでは, 「常に悪性」と判定すれば誤検知を増加させ, 「常に正常」と判定すれば将来的な再悪用を見逃すリスクがある．したがって, **時間的コンテキストを考慮したリスクスコアリング**や, **過去履歴を統合的に参照する仕組み**の導入が不可欠である．

(3) 履歴参照機能の必要性現行のエージェントは単一時点での分析を前提としているが, 今後は WHOIS や証明書, URLScan 履歴などの外部データベースを統合し, 過去から現在に至る変遷を踏まえた判断を可能にする必要がある．このような「時間軸に沿った脅威認識」は, 動的に変化するフィッシング攻撃の検知精度をさらに高めると考えられる．

5. 議論

5.1 制約事項

本研究の提案手法にはいくつかの制約がある．第一に，計算資源の要求が挙げられる．本研究で採用した LLM モデルの実行には GPU（推奨：VRAM 24GB 以上）が必要であり，小規模組織での導入にはハードルとなる可能性がある．ただし，FP8 量子化によりメモリ要求量の削減や知識蒸留による軽量モデルへの移植も可能である．そのような軽量化したケースにおける性能評価は今後の課題である．

第二に，処理時間の制約がある．AI エージェントが達成した平均 2.52 秒/ドメインの処理時間は，大規模なフィッシングサイト情報を処理する環境においては，リアルタイムに実行することが困難である．現在の実装は，XGBoost の初期判定後に適用する二段階処理を前提としているが，ML による一時フィルタ後のドメイン到着率が上記の速度を超える場合はさらなる性能向上が必要である．

第三に，言語依存性の問題がある．現在のブランドキーワード辞書は日英のみ対応しており，他言語のフィッシングサイトへの適用には拡張が必要である．ただし，Qwen3 モデル自体は多言語対応のため，辞書の拡張により対応可能である．

5.2 今後の課題

本研究のアプローチと実験を通じ，以下の研究課題が明らかになった．

(1) 適応的特徴量学習：本研究で明らかになった, XGBoost が低いフィッシング確率を出力した領域に巧妙な攻撃が集中する傾向は，攻撃者が ML モデルの特性を分析し，意図的に検出を回避している可能性を示唆している．この知見に基づき，AI エージェントが抽出した特徴パターン（証明書の組織情報欠如，ブランド-TLD 不整合，情報量削減）を自動的に ML モデルの再訓練に反映させるパイプラインの構築が必要である．これにより，システムは攻撃者の戦略変化に継続的に適応できる．

(2) 実用化に向けた技術的課題：提案手法の実用化には，処理性能の改善が不可欠である．知識蒸留によるモデルの軽量化や抽出済みパターンのキャッシング機構と，未知パターンのみの LLM 分析を組み合わせることで，応答時間を短縮することで，リアルタイムフィルタリングへの適用可能性を探ることは今後の課題である．

(3) 汎用フレームワークとしての発展：本研究で提案した ML と LLM のハイブリッド型アプローチの応用は，フィッシング検出に限定されない．マルウェア検知，スパムフィルタリング，不正取引検出など，他のセキュリティ分野においても，ML モデルの境界領域を分析することで新たな知見が得られる可能性がある．各分野特有の特徴抽出ツー

ルの開発により，ML と LLM の相補的統合の汎用フレームワークへの発展が期待される．

(4) 説明可能性の深化：現在の特徴抽出は記述的レベルに留まっているが，より実用的な知見を提供するには深い分析が必要である．反事実的説明（counterfactual explanation）技術の統合により，「なぜ見逃されたか」だけでなく「どう改変すれば検出可能か」まで提示できるシステムの実現が期待できる．これにより，セキュリティアナリストへのより実践的な支援が可能となる．

6. まとめ

本研究では，ML モデルと LLM ベースの AI エージェントを組み合わせるハイブリッド型フィッシング検出システムを提案した．鍵となるアイディアは，ML モデルの予測信頼度が低いドメインに対してのみ，AI エージェントを用いた深い分析を実施するアプローチを採用することにある．640,356 件のデータセットを用いた実験では，XGBoost モデルが低いフィッシング判定確信度（0.5 未満）を出力した 4,215 件のドメインに対し，AI エージェントによる詳細分析を適用したところ，そのうちの 3,576 件 (84.8%) をフィッシングとして正しく判定し，システム全体の偽陰性率を 6.58%から 1.04%まで削減することに成功した．特に，XGBoost によるフィッシング判定確信度が低かった (0.2 未満)2,196 件のフィッシングサイトに対し，AI エージェントによる検出率は 100%を達成した．これらの結果は，ML によるフィッシング判定確信度が低いサンプルに巧妙な攻撃が潜むことを示唆するものであり，AI エージェントが既存の ML 方式の不足を補うことを明確に示すものである．提案方式の実用化に向けた AI エージェントのさらなる軽量化，および分析で得られた特徴量の自動学習は今後の課題である．

謝辞　本研究の実施にあたり，フィッシングサイトデータの提供にご協力いただいた JPCERT/CC ならびに PhishTank プロジェクトに深く感謝いたします．

参考文献

[1] Anti-Phishing Working Group: Phishing Activity Trends Report, 1st Quarter 2021, https://docs.apwg.org/reports/apwg_trends_report_q1_2021.pdf .

[2] Koide, T., Fukushi, N., Nakano, H., and Chiba, D.: Detecting Phishing Sites Using ChatGPT, arXiv preprint arXiv:2306.05816 (2023).

[3] Sahingoz, O.K., Buber, E., Demir, O., and Diri, B.: Machine learning based phishing detection from URLs, Expert Systems with Applications, Vol.117, pp.345–357 (2019).

[4] Utami, M.R.T.: Enhancing Phishing Detection: Integrating XGBoost with Feature Selection Techniques, SSRN preprint SSRN:5087049 (2024).

[5] Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q.E.U., Saleem, K., and Faheem, M.H.: A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN, Electronics, Vol.12, No.1, p.232 (2023).

[6] Trad, F. and Chehab, A.: Prompt Engineering or Fine-Tuning? A Case Study on Phishing Detection with Large Language Models, Machine Learning and Knowledge Extraction, Vol.6, No.1, pp.367–384 (2024).

[7] Chen, F., et al.: Adapting to Cyber Threats: A Phishing Evolution Network (PEN) for Phishing Generation and Pattern Analysis using Large Language Models, arXiv preprint arXiv:2411.11389 (2024).

[8] Achary, R., Bugath, S.N., Chakrapani, G., and Venkatesh, M.: Enhanced Phishing Detection Using LSTM, CNN, and SVM Techniques, Proc. Intelligent Strategies for ICT (ICTCS 2024), LNNS Vol.1320, pp.185–204 (2025).

[9] AlSabah, M., Nabeel, M., Choo, E., and Boshmaf, Y.: Content-Agnostic Detection of Phishing Domains using Certificate Transparency and Passive DNS, Proc. 25th International Symposium on Research in Attacks, Intrusions and Defenses (RAID'22), pp.446–459 (2022). https://dl.acm.org/doi/10.1145/3545948.3545958

[10] Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.785–794 (2016).