**匿名通信トラフィックにおける悪性挙動のモデル化とXAIによる解釈**

1. **研究背景と目的**

**1.1. 技術的課題：暗号化による「観測不能性」**

TLSに代表される通信の暗号化は、現代のインターネットにおいて不可欠な技術です。しかし、この「観測不能性」は、攻撃者がマルウェアのC2（Command & Control）通信などを隠蔽する格好の手段となっています。特にTorのような多層暗号化と匿名化ルーティングを用いるネットワークでは、パケットのペイロード（中身）を解析する従来型のIDS/IPS（侵入検知/防御システム）は完全に無力化されます。

この結果、我々は通信の**メタデータ**（パケット長、通信タイミング、方向など）のみを観測可能な「ブラックボックス」と向き合わなければなりません。

**1.2. 本研究の目的：検知から「挙動のモデル化」へ**

本研究の目的は、単に暗号化された悪性トラフィックを分類・検知することに留まりません。最終的なゴールは、**「特定のマルウェアに依存しない、悪性挙動の"普遍的"な統計的特徴とは何か？」**という問いに答えることです。

そのために、機械学習モデルとXAI（説明可能AI）を組み合わせ、以下の2点を達成します。

1. 悪性挙動を高精度で分類するモデルを構築する。
2. XAIを用いてモデルの判断根拠を解明し、人間が理解できる**「悪性挙動のシグネチャ」**を定量的に定義する。

**2. アプローチの新規性と独自性 💡**

本研究の新規性は、既存技術の単純な組み合わせではなく、**研究のパラダイムそのもの**にあります。AIを単なる答えを出す機械としてではなく、未知のパターンを発見するための対話的なパートナーとして活用する点が、他の多くの研究とは一線を画します。

| **観点** | **従来のアプローチ** | **本研究のアプローチ** |
| --- | --- | --- |
| **目的** | **分類 (Classification):** 正常か異常かの二値分類問題として扱う。精度（Accuracy）が主たる評価指標。 | **モデル化 (Modeling):** なぜ異常なのか、その挙動の本質的な特徴を定義する。精度に加え、**解釈可能性 (Interpretability)**を重視。 |
| **AIの役割** | **予測器 (Predictor):** 高精度なブラックボックスとして利用。判断理由は問わない。 | **分析ツール (Analyzer):** AIを一種のセンサーとして用い、その内部状態をXAIで観測することで、データに潜む本質的なパターンを抽出する。 |
| **プロセス** | **一方向：** データ → 学習 → 予測 | **フィードバックループ：** データ → 学習 → **XAIで解釈 → 知見を得る → 特徴量を改良** → 再学習 |

**3. 実験設計と技術的プロセス**

**3.1. データセットの可視化とベースラインモデル**

倫理的配慮から、正常通信（Webブラウジング）と悪性通信（C2ビーコン）を模倣したシミュレータをPythonで開発し、pcap形式のデータセットを構築しました。まず、両者の振る舞いの違いを可視化しました。

グラフ, 散布図

AI 生成コンテンツは誤りを含む可能性があります。図1: 正常なWebブラウジングのトラフィックパターン。通信がバースト状に発生し、受信（青）が送信（赤）を圧倒する非対称な構造が見られる。

**グラフ, 散布図

AI 生成コンテンツは誤りを含む可能性があります。**

**図2: C2通信のトラフィックパターン。通信が等間隔（周期的）に発生し、送受信のパケットサイズが近い対称な構造が見られる。**

この視覚的な違いに基づき、個々のパケットが持つ基本的な特徴量（relative\_time, length, is\_outgoing）のみを用いてベースラインとなるランダムフォレストモデルを訓練した結果、**Accuracy: 96%** という高い精度を達成しました。

**3.2. XAIによるブラックボックスの解明**

しかし、この時点では「なぜ96%の精度が出るのか」は不明です。そこで、モデルの判断根拠を可視化するため**SHAP (SHapley Additive exPlanations)** を適用しました。

**グラフ, 箱ひげ図

AI 生成コンテンツは誤りを含む可能性があります。**

**図3: SHAPによる特徴量の重要度分析。relative\_time（時間）やlength（サイズ）が判断に大きく影響していることがわかる。**

SHAP値の分析から、モデルが以下の3つの要素を強く重視していることが定量的に明らかになりました。

1. **時間的周期性:** パケットの到着間隔の規則性。
2. **パケット長の均一性:** 大きなサイズのパケットが存在せず、全体的にサイズが均一であること。
3. **通信の対称性:** 送信と受信のデータ量やパケット数が近いこと。

これは、C2通信が「定期的に」「軽量なビーコンを」「送受信する」という本質的な振る舞いと完全に一致します。

**3.3. XAIの知見を活用した特徴量エンジニアリング**

XAIが示してくれた上記3つの「答え」を基に、より本質的な特徴量を設計しました。分析の単位を個々のパケットから、**10秒間のタイムウィンドウ（フロー）**へと移行し、以下の特徴量を新たに算出しました。

* **周期性の指標:** パケット到着間隔の標準偏差 (std\_iat)
* **均一性の指標:** パケット長の平均値 (avg\_packet\_size) および標準偏差 (std\_packet\_size)
* **対称性の指標:** 送受信パケット数の比率 (direction\_ratio)

このプロセスは、AIによる発見的知識を人間のドメイン知識で洗練させ、再度AIに与えるという、解釈可能性を前提とした開発ループそのものです。

**4. 結果と考察 📈**

高度な特徴量で最終モデルを再訓練した結果、性能は劇的に向上しました。

| **評価指標** | **初期モデル (パケット単位)** | **最終モデル (フロー単位)** |
| --- | --- | --- |
| **Recall (検知率)** | 72% | **83%** |
| **Precision (適合率)** | 75% | **100%** |
| **False Positive Rate** | 高 | **0%** |

**表1: 初期モデルと最終モデルの性能比較。特にPrecisionとFalse Positive Rateが大幅に改善した。**

グラフ, ツリーマップ図

AI 生成コンテンツは誤りを含む可能性があります。

図4: 初期モデルの混同行列。正常通信を悪性と誤検知するFalse Positive（右上）が55件発生している。

**グラフ が含まれている画像

AI 生成コンテンツは誤りを含む可能性があります。**

**図5: 最終モデルの混同行列。False Positive（右上）が0件になり、誤報が完全に撲滅された。**

特筆すべきは、**Precisionが100%**、すなわち**誤報がゼロ**になった点です。これは、XAIの示唆した特徴量が、ノイズ（正常通信）とシグナル（悪性通信）を分離する上で極めて有効であったことを示しています。

**考察：高度な攻撃への耐性について**

「攻撃者は通信間隔をランダム化するのでは？」という当然の疑問が浮かびます。しかし、本アプローチの強みは、複数の特徴を総合的に評価する点にあります。攻撃者が一つの特徴を偽装しようとすると、別の特徴に矛盾が生じる**「いたちごっこ（Attacker's Dilemma）」**の状況を作り出すことができます。

* 攻撃者が**周期性**を隠すために通信間隔をランダム化しても、C2通信の本質である**均一性**や**対称性**は残ります。
* 攻撃者が**パケット長**を偽装するために巨大なデータを送信すれば、それは正常なWeb閲覧（**送信データ量が受信データ量より著しく少ない状態**）とは真逆の異常な対称性（**送信データ量が受信データ量より著しく多い状態**）を生み出します。

我々のモデルは、この「偽装しようとする不自然な努力」そのものを異常として捉えることが可能です。

**5. 結論と技術的貢献**

本研究は、単に高精度な分類器を開発しただけではありません。その技術的貢献は以下の3点に集約されます。

1. **方法論の確立:** 「AIによる分類 → XAIによる解釈 → 人間による特徴量エンジニアリング」というフィードバックループを実証し、ブラックボックス問題に対する具体的な解決策を提示しました。
2. **普遍的特徴の定義:** 匿名通信下の悪性挙動を特徴づける3つの統計的指標（周期性、均一性、対称性）を、データに基づき定量的に定義しました。
3. **実用性の高いモデル:** 誤報率ゼロを達成した最終モデルは、セキュリティ監視の現場における「アラート疲れ」を軽減しうる、実用性の高いアプローチであることを示唆しています。

この研究は、AIを「答えを出す機械」から**「人間と協働して未知の洞察を発見するパートナー」**へと昇華させる、次世代のAI応用の一つの形を示した事例と言えます。