異種マルウェア混入による IoT マルウェア画像分類への標的型攻撃の検討

川田 隼大[1]　稲村 浩[2]　石田 繁巳[2]

概要：IoT マルウェアの急増に伴い，IoT マルウェア画像分類手法を解析対象となる IoT マルウェアのトリアージに使用し，効率的に解析することは攻撃の動向把握において有効なアプローチである．しかしながら，攻撃者が分類結果を恣意的に操作可能になった場合，誤ったトリアージの判断による攻撃の動向把握に遅れが生じる可能性がある．そこで本研究では，IoT マルウェア画像分類に対する標的型攻撃の実現可能性の検証を目的とし，異種マルウェア混入手法を提案した．従来手法と提案手法の ASR を比較した結果，提案手法によって ASR の向上が見られ，提案手法による標的型攻撃の有効性を示した．さらに，提案手法の成功可否と検体サイズ比率の関係を分析した結果，提案手法の安定した成功には，検体サイズ比率が 0.75 を超える大量の混入が必要であることを示した．

1. はじめに

IoT（Internet of Things）デバイスの普及に伴い，それらを標的とした IoT マルウェアの脅威が深刻化している [1]．IoT マルウェアは，公開されたソースコードを基にした亜種が大量に作成されており [2]，限られた解析者のリソースで全ての検体を詳細に解析することは現実的ではない．増大し続ける IoT マルウェアを効率的に解析することが，攻撃の動向把握には重要である．

解析対象となる IoT マルウェアをマルウェアファミリー分類の結果でトリアージすることは，効率的な解析を実現する上で有用である．Su らは IoT マルウェアのバイナリをグレースケール画像として可視化し，畳み込みニューラルネットワーク（CNN）を用いて自動分類する手法を提案した [3]．この手法は人手による特徴抽出なしで高精度な分類を実現している．さらに，Singh らは，未知のパッカーで圧縮されたマルウェアの分類で 60.50%の精度を達成し，パッキングに対する一定の堅牢性を示した [4]．以上のことから IoT マルウェア画像分類手法を解析対象となる IoT マルウェアのトリアージに使用し，効率的に解析することは攻撃の動向把握において有効なアプローチである．

一方で，解析者がマルウェアファミリー分類の結果を信頼するようになると，攻撃者は分類結果を恣意的に操作しようとする動機が生じる．具体的には，新機能を追加した「脅威の高いファミリーのマルウェア」を，「脅威の低いファミリーのマルウェア」と誤分類させることで，新機能の詳細な解析を回避するという動機が考えられる．マルウェアファミリー分類の結果に基づいたトリアージの判断を誤らせる攻撃は，攻撃の動向把握が遅れ，最悪の場合，脅威の永続化へつながる危険性がある．

攻撃者が現実の IoT マルウェア解析基盤に対して，分類結果の恣意的な操作を実現するためには，ターゲットへの誘導を図りつつ，実行可能性を維持したバイナリ改変を行う必要がある．本研究は，IoT マルウェア画像分類による解析検体のトリアージを導入した解析基盤を標的とする．検体がハニーポット等の収集経路を通過するためには，実際に被害を及ぼす実体が必要であり，実行不能なファイルは収集対象から除外される可能性が高い．したがって，元の機能を損なわないバイナリ改変は，IoT マルウェア画像分類に対する攻撃を成立させるための重要要件となる．

従来のマルウェアバイナリに対する攻撃手法は，マルウェアとしての実行可能性を維持できるものの，非標的型攻撃にとどまるという限界がある．具体的には，悪性ファイルを良性ファイルに偽装する手法や，もとのファミリーとは異なるファミリーに誤分類させることを目的とした手法が広く議論されてきた [5]．これらの手法は，マルウェアとしての機能を損なわずにバイナリを改変可能であるものの，攻撃者が意図した特定のファミリーを狙って分類結果を誘導する標的型攻撃を実現できない．

これに対し，標的型攻撃が可能なアプローチとして勾配ベースの手法が存在するが，こちらはマルウェアの実行可能性を維持できないという課題がある [6]．自然画像分類

[1] 公立はこだて未来大学大学院システム情報科学研究科
[2] 公立はこだて未来大学システム情報科学部

Algorithm 1 マルウェアバイナリの画像化

```
function CREATE_IMAGE(binary)
    L ← sizeof(binary)
    s ← ⌈√L⌉
    padded_binary ← zeros(s × s)
    for i = 0 to L − 1 do
        padded_binary[i] ← binary[i]
    end for
    image ← array(s, s)
    for y = 0 to s − 1 do
        for x = 0 to s − 1 do
            image[y][x] ← padded_binary[y × s + x]
        end for
    end for
    return image
end function
```

の分野から応用された勾配ベースの手法をマルウェアファミリー分類に適用した場合，標的型攻撃自体は達成できるものの，ピクセル値が変更された画像をバイナリへ対応づける際，バイナリの構造が破壊され，実行可能性を維持できない．したがって，現実の IoT マルウェア解析基盤に対する攻撃手法を想定する際は，標的型攻撃の成立と実行可能性の維持を両立したアプローチの検討が極めて重要である．

しかしながら，これら双方の要件を同時に満たす攻撃手法や，それがもたらす現実的な脅威については，これまで十分に議論されていない．そこで本研究では，IoT マルウェア画像分類手法に対し，実行可能性を維持した標的型攻撃が実現可能であるかを検証することを目的とする．

本研究では，ファミリーを隠したいマルウェア検体に対して，異なるファミリーの検体を新規セグメントとして実行可能性を維持して混入する，異種マルウェア混入手法を提案する．本稿の評価では，異種マルウェア混入手法が与える分類結果への影響を評価し，IoT マルウェア画像分類に対する標的型攻撃の有効性および攻撃成立に必要なファイルサイズの条件を明らかにする．

本稿の構成は以下のとおりである．第 2 節にて，関連研究について説明し，第 3 節にて，提案手法の詳細を説明する．第 4 節で，提案手法の評価を行い，最後に第 5 節にてまとめとする．

2. 関連研究

2.1 IoT マルウェア画像分類手法

IoT マルウェアのファミリー分類を自動で行う手法として，Su らによって提案された，IoT マルウェア画像分類手法が挙げられる [3]．この手法では，マルウェア実行形式バイナリをグレースケール画像（**図 1**）に変換し，マルウェアファミリー分類を行う．この手法における画像変換アルゴリズムを **Algorithm 1** に示す．ここで，入力 *binary* はマルウェア実行形式バイナリを，出力 *image* は生成される

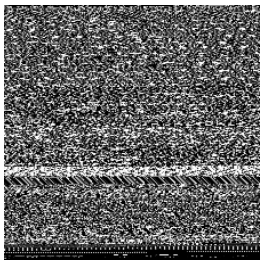


図 1: IoT マルウェア（Mirai）のグレースケール画像

グレースケール画像の 2 次元配列を表す．IoT マルウェアの画像化プロセスでは，マルウェア実行形式バイナリの 1 バイトを 1 ピクセルの輝度値として，2 次元の行列にマッピングしている．画像サイズはファイルサイズに応じて決定され，ファイルサイズの平方根を切り上げることで正方画像を生成している．バイナリデータが画像の全ピクセルを満たさない場合，余剰となるピクセルには輝度値 0 が設定される．

生成された画像に対して，VGG19 などの既存の画像分類アーキテクチャを適用し，CNN を用いて特徴抽出と分類を行う．CNN は画像の局所的なパターンや構造を自動的に学習できる特性を持ち，マルウェアのファミリー間に見られるコード構造やデータ構造の違いにより生じる画像上の特徴を捉えるのに適している．このアプローチは，既存研究 [3, 7] においても有効性が報告されている．

IoT マルウェア画像分類手法の利点として，パッキングに対して高い堅牢性を有していることが挙げられる．マルウェア分類における課題として，パッカーやクリプターによる実行ファイルの難読化が挙げられる．これは，従来のシグネチャベースの手法や，特定の文字列・命令列のパターンマッチングに基づいた分類を無力化してしまう．しかし Alkhateeb らは，CNN が学習過程において，パッキングツール自体がバイナリに付与する特有の構造的パターンを特徴として自動的に学習することを示している [8]．さらに，未知のパッカーで処理された実行ファイルに対しても，グレースケール画像と深層学習の組み合わせは，高い精度とバランスの取れた F1-Score を達成していることを示している．このことから，パッキングされたマルウェアの早期分類において非常に有用なアプローチであることが IoT マルウェア画像分類手法の利点であるといえる．

異なる利点として，転移学習アプローチの採用により，少数の学習データでもマルウェアの特徴を効果的に学習できる点が挙げられる．通常，深層学習モデルをゼロから訓練するには膨大な量のラベル付きデータと計算リソースが必要となる．しかし，IoT マルウェアデータセットにおいては，ファミリー間の検体数の不均衡性が課題であり，特に新種のマルウェアファミリーに対して十分な量の学習データを迅速に収集することは困難である．しかし Prima らは，MALIMG などの不均衡データセットを用いた検証

実験で，事前学習済みの VGG16 や ResNet-50 を転移学習させたモデルが，大規模なデータ拡張を必要とせずに、平均して 98%から 99.97%という分類精度を達成できることを示した [9]．このことから，少数の学習データしか利用できない現状において，非常に有用なアプローチであることが IoT マルウェア画像分類手法の利点であるといえる．

2.2 従来の攻撃手法

マルウェア自動分類システムに対する従来の攻撃手法として，実行ファイルに影響を与えない末尾領域などへの Random Noise の付加や Zero Padding が挙げられる．Park らは，マルウェアのセクション末尾にランダムなバイト列を追加する攻撃を提案し，実行可能性を維持しつつも誤分類が誘発されることを示した [10]．筆者らは，マルウェア解析におけるマルウェア収集経路を用いた攻撃として，空セクションを用いた Zero Padding による攻撃を提案し，実行可能性を維持しつつもゼロデータ量の増加によって誤分類率が増加することを明らかにした [11]．しかし，これらの手法は誤分類を誘発するに過ぎず，攻撃者が意図した特定のクラスへと分類器を誘導する標的型攻撃は不可能であり，非標的型攻撃にとどまるという課題がある．

一方，自然画像分類の分野では勾配ベースの標的型攻撃が多数提案されている．Alzaidy らは，勾配ベースの標的型攻撃である JSMA と C&W を Windows マルウェア画像分類に適用する事で，CNN モデルに対して 91%の回避率を達成することを示した [12]．このことから，自然画像分類の分野で提案されている勾配ベースの標的型攻撃手法を IoT マルウェア画像分類に適用することで，極めて高い精度で標的型攻撃が実現できることが考えられる．しかしながら，ピクセル値の微小な変更が検体のコードや ELF ヘッダーの構造を破壊してしまうため，作成されたマルウェアはプログラムとしての実行可能性を損ない，現実の脅威として機能しないという課題がある．

3. 提案手法

3.1 攻撃の前提条件

本研究は，攻撃者が分類器の内部情報にアクセスできないブラックボックス環境における攻撃シナリオを想定する．このシナリオにおける攻撃者は，内部パラメータや訓練データを取得不可能であるものの，出力としての予測確信度は取得可能である．この設定は，攻撃者が公開 API や商用解析基盤を利用する際の制約に一致する．したがって，現実的な脅威モデルの検討には，ブラックボックス環境下で分類器の判断を誘導する手法の設計が不可欠である．

3.2 キーアイディア

本研究のキーアイディアは，異種マルウェアのバイナリ全体を元の検体に混入することである．異種マルウェアのバイナリ全体を混入させることで，異種マルウェアが持つファミリー固有のテクスチャを顕在化させる．

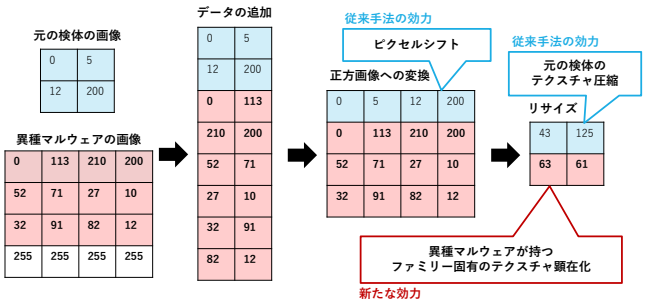


図 2: キーアイディアの概要

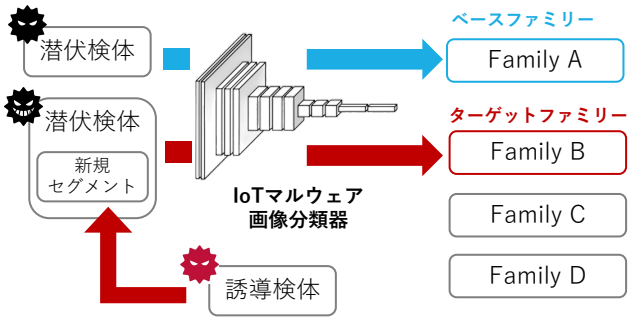


図 3: 提案手法の概要図

本研究のキーアイディアの概要を**図 2** に示す．キーアイディアにおいて画像上に発生するテクスチャの変化は大きく 3 つに分類される．具体的な変化は，ピクセルシフト，元の検体のテクスチャ圧縮，異種マルウェアが持つファミリー固有のテクスチャ顕在化である．このうち，ピクセルシフトと元の検体のテクスチャ圧縮は，元の検体のテクスチャの破壊による分類精度の低下を目的としており，従来手法も有する効力である．一方，異種マルウェアのテクスチャ顕在化は本提案によって追加される新たな効力である．

画像上に顕在化したテクスチャは，分類器が画像から抽出する特徴量を異種マルウェアが属するファミリーの特徴量に近づけ，誘導に寄与する．画像上に異種マルウェアが持つファミリー固有のテクスチャが顕在化することで，CNN が画像から抽出する特徴量と CNN が学習した異種マルウェアが属するファミリーの特徴量との類似度が上昇する．この類似度上昇により，分類器の分類結果を異種マルウェアが属するファミリーへと誘導する．

以降では，キーアイディアを踏まえた本研究の提案手法である，異種マルウェア混入手法について詳述する．

3.3 提案手法の概要

図 3 に IoT マルウェア画像分類に対する標的型攻撃手法を示す．本稿では，議論を明確にするために以下の用語を定義する．

- ターゲットファミリー：攻撃者が分類を誘導したい目

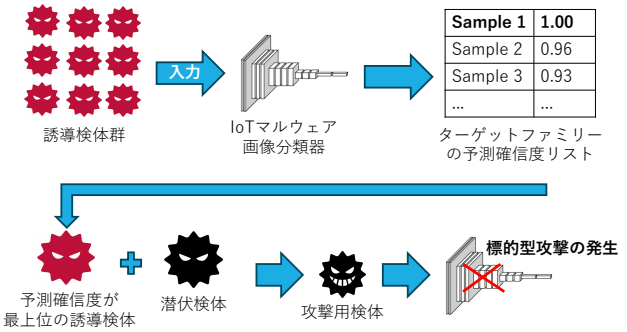


図 4: 誘導検体の選定

標のファミリー

- 潜伏検体：異種マルウェア混入対象となる，ファミリーを隠したいマルウェア検体
- 誘導検体：誤分類を誘発するために利用する，ターゲットファミリーに属するマルウェア検体

提案手法では，誘導検体のバイナリ全体を潜伏検体に混入することで，分類結果をベースファミリーからターゲットファミリーへ誘導する．

提案手法は，予測確信度に基づいて誘導検体の選定を行った後，実行可能性を維持して誘導検体のバイナリ全体を混入するという 2 つのプロセスで構成される．以降では，各プロセスについて詳述する．

3.4 誘導検体の選定

誘導検体の選定プロセスを**図 4** に示す．攻撃者は公開された検体や独自収集した検体を利用して，誘導検体群を用意する．用意した誘導検体群を IoT マルウェア画像分類器へ入力し，各検体の予測確信度を得る．その後，予測確信度が最も高い検体を誘導検体として選定する．

選定指標として予測確信度を採用する妥当性は，CNN における特徴量の重要度評価，およびソフトマックス関数の単調増加性から説明される．Zhou らは，CNN の最終層の出力が画像内の各空間位置における重要度の総和に等しいと述べており [13]，本研究の対象である IoT マルウェア画像において，この重要度はバイナリ由来の局所的なテクスチャ特徴に対応する．予測確信度はこの出力をソフトマックス関数で正規化したものであり単調増加性を有するため，予測確信度の高い検体の選定は，ターゲットファミリー固有のテクスチャを高純度で保持した検体の選定に等しい．したがって，本プロセスによって，最適な誘導検体の選定が可能になる．

3.5 実行可能性を維持した誘導検体バイナリの混入

実行可能性を維持した誘導検体の混入は，潜伏検体に新規の実行可能セグメントを作成し，セグメントのコンテンツを誘導検体のバイナリに設定することで実現する．誘導検体のバイナリ混入プロセスを**図 5** に示す．バイナリ混入プロセスではまず，潜伏検体に新たなセグメント（PT_LOAD）を追加する．追加するセグメントには，実行可能かつ読み取り可能な属性（R+X）を設定する．次に，セグメントのコンテンツとして誘導検体のバイナリを設定する．最後に，セグメントの追加に伴う ELF ヘッダーやプログラムヘッダーの整合性を保つ．この際，ヘッダーの再配置に伴って新たにセグメントが追加される場合もある．この手法は，潜伏検体のエントリーポイントや制御フローに干渉しないため，ターゲットファミリー固有のテクスチャを導入しつつ実行可能性を維持できる．

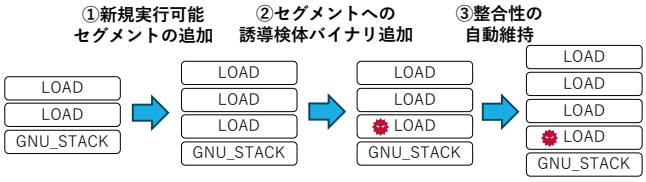


図 5: 誘導検体バイナリの混入プロセス

本手法の特徴は，外部ライブラリの利用によって実行可能性の維持を容易に実装できる点にある．例えば LIEF ライブラリ [14] を使用すれば，セグメント追加時に ELF ヘッダやプログラムヘッダテーブルの整合性が自動的に保たれる．攻撃者は高度な編集技術を要さずに標的型攻撃を実現可能であるため，早急な防御策が必要である．

4. 評価

提案手法の有効性を評価するため，2 つの Research Question を立て，実験を行った．

- **RQ1: 異種マルウェア混入手法における誘導効力の追加は，標的型攻撃の成功に寄与するか?**
  異種マルウェア混入手法における誘導効力の追加が，標的型攻撃の成功に寄与するかを検証した．従来手法は潜伏検体のファミリー固有のテクスチャを破壊し，正確な分類を妨害していた．異種マルウェア混入手法では従来手法の効力に加え，ターゲットファミリー固有のテクスチャを顕在化させた．したがって，従来手法と異種マルウェア混入手法を比較することで，誘導効力追加の有効性を明らかにした．

- **RQ2: 提案手法が有効となる誘導検体のファイルサイズはどれくらいか?**
  提案手法の限界を評価するため，攻撃成功に必要な誘導検体のファイルサイズを明らかにした．ファイルサイズの極端な肥大化は解析者に分類結果への不審感を抱かせ，追加の解析プロセスに組み込まれるリスクを高くする．このため，検体が解析から逃れるためには，必要最小限のバイナリ混入での標的型攻撃の成功が望ましい．したがって，潜伏検体と誘導検体の検体サイズ比率を段階的に変化させ，分類器の判断が反転する

表 1: ラベリングの内訳

ファミリー名	検体数	割合（%）
Mirai	5444	39.43
Gafgyt	3779	27.37
Generica	770	5.58
Rootkit	279	2.02
Tsunami	266	1.93
Dofloo	204	1.48
Flooder	118	0.85
Jiagu	46	0.33
Mrblack	8	0.06
Vpnfilter	6	0.04
Other	2595	18.79

閾値を特定することで，現時点での提案手法の限界を明らかにした．

4.1 評価準備

4.1.1 データセットの準備

本研究の評価では，Olsen らによって提供されているラベル付き IoT マルウェア検体データセット [15] を使用した．Olsen らは，各検体のラベリングに AVClass [16] を使用し，複数のアーキテクチャにわたる検体を収集することでデータセットを構築した．IoT マルウェア画像は同一のマルウェアファミリーであっても，アーキテクチャが異なると見た目も大きく異なる．このため，評価では Intel 80386 アーキテクチャの検体のみを使用した．Intel 80386 アーキテクチャにおけるマルウェアファミリーごとの検体数内訳を**表 1** に示す．本研究で対象としたマルウェアファミリーは，検体数が十分に多い Gafgyt，Mirai，Tsunami，Rootkit，Generica，Dofloo の 6 ファミリーである．

対象ファミリーの各検体は **Algorithm 1** に則って，グレースケールの正方画像に変換した．本研究では，特徴抽出器として VGG19 を使用したため，生成されたグレースケールの正方画像は入力形式である 224 × 224 にリサイズした．

4.1.2 IoT マルウェア画像分類器の作成

評価に使用する IoT マルウェア画像分類器の作成を行った．作成においてはデータセットを分割し，訓練用には 7520 検体，検証用には 1074 検体，テスト用には 2153 検体を使用した．テストデータは作成した画像分類器の性能評価に使用し，提案手法の評価に妥当な分類器となっているかを確認した．

IoT マルウェア画像分類器は Pytorch を用いて実装し，以下の設定で学習した．

- 特徴抽出器: VGG19（ImageNet1K_V1 で事前学習済み）
- 最適化アルゴリズム: Adam
- 学習率: 0.0001

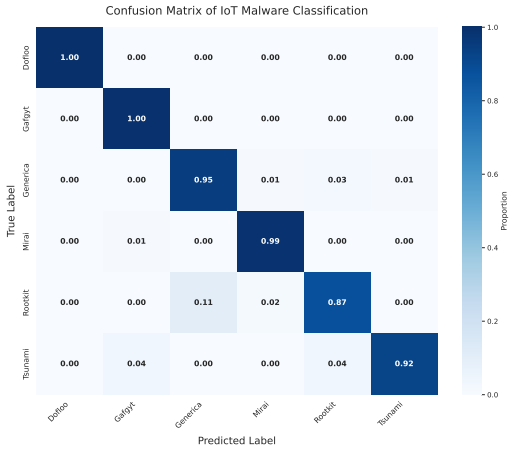


図 6: IoT マルウェア画像分類器の分類性能

- バッチサイズ: 32
- エポック数: 20

テストデータを用いた分類結果を**図 6** に示す．図を見ると，全てのファミリーに対して 8 割以上の F1-Score を達成していることが確認できる．このことから，作成した IoT マルウェア画像分類器は各ファミリー固有のテクスチャを十分に学習しており，評価における攻撃対象として十分な堅牢性を有しているモデルであることを確認した．

4.1.3 評価用検体の選定

評価に使用する潜伏検体と誘導検体の選定を行った．潜伏検体は，データセットにおける Gafgyt，Mirai から，ファイルサイズに多様性を持たせて 25 検体ずつ選定した．具体的には，35KB 程度の軽量な検体から 2000KB 程度の大規模な検体までを網羅的に抽出した．誘導検体は，データセットにおける Gafgyt，Mirai，Rootkit，Dofloo，Tsunami，Generica からファイルサイズがおおよそ同程度かつターゲットファミリーの予測確信度が 1.0 である検体を 3 検体ずつ選定した．

4.2 RQ1

RQ1 は，「異種マルウェア混入手法における誘導効力の追加は，標的型攻撃の成功に寄与するか?」である．RQ1 の評価では提案手法と従来手法（Random Noise, Zero Padding）をそれぞれ適用した検体を作成し，Attack Success Rate（ASR）を比較した．このとき，従来手法適用時の Random データおよび Zero データは，提案手法における誘導検体のファイルサイズと同一にした．ASR の算出方法を (1) に示す．

$$\mathrm{ASR} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(f(x_i') = y_t) \qquad (1)$$

ASR とは，手法を適用した全検体のうち標的型攻撃が成功した検体の割合を指す．標的型攻撃の成功とは，ターゲットファミリーと手法適用後の検体の予測ラベルが一致することを指す．(1) における N は評価に用いた全検体数，x_i'

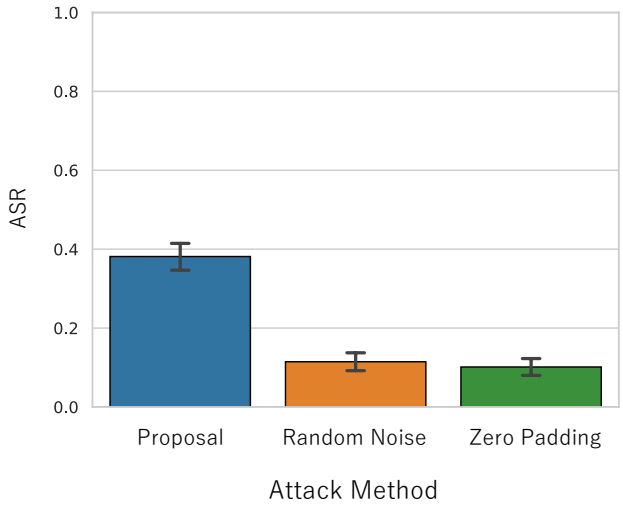


図 7: 提案手法と従来手法の ASR 比較

は手法適用後の検体から生成されるグレースケール画像，$f(x_i')$ は IoT マルウェア画像分類による予測ファミリー，y_t はターゲットファミリーを表す．また，$\mathbb{I}(\cdot)$ は指示関数であり，カッコ内の条件が真である場合に 1，偽である場合に 0 を出力する．

提案手法と従来手法の ASR を比較した結果を**図 7** に示す．図から，従来手法の ASR が 0.11, 0.10 と低かったのに対し，提案手法の ASR は 0.38 と高いことが確認できる．この結果から，従来手法の効力のみでは標的型攻撃が成功しなかった検体が，提案手法における誘導効力の追加によって標的型攻撃が成功したことがわかる．

各ターゲットファミリーの ASR を**図 8** に示す．図から，全ての潜伏検体のファミリーと誘導検体のファミリーの組み合わせで，提案手法が従来手法を上回る ASR を記録したことが確認できる．特に，Mirai や Gafgyt に比べ，相対的に脅威が低く設定される Rootkit や Dofloo [17] をターゲットファミリーとした場合においても ASR の向上が確認できる．この結果は，新機能の詳細な解析を回避するという攻撃者の動機を直接実現するものであり，提案手法の脅威の高さを示している．

これらの評価を通じて，従来手法の効力のみでは標的型攻撃の成功は困難であり，異種マルウェア混入手法における誘導効力の追加が，標的型攻撃の成功に大きく寄与したと本 RQ では結論づける．

4.3 RQ2

RQ2 は，「提案手法が有効となる誘導検体のファイルサイズはどれくらいか?」である．RQ2 の評価では，ターゲットファミリーごとで横軸に検体サイズ比率（R_{mix}），縦軸に成功可否（0 or 1）をとった散布図を作成し，ロジスティック回帰で近似曲線を算出した．算出したモデルから，近似曲線の値が 0.5 となる臨界検体サイズ比率を算出し，攻撃

表 2: 臨界検体サイズ比率

潜伏検体のファミリー	ターゲットファミリー	臨界検体サイズ比率
Gafgyt	Dofloo	0.69
Gafgyt	Generica	0.33
Gafgyt	Mirai	0.69
Gafgyt	Rootkit	0.67
Gafgyt	Tsunami	0.64
Mirai	Dofloo	0.75
Mirai	Generica	0.36
Mirai	Gafgyt	0.61
Mirai	Rootkit	0.70
Mirai	Tsunami	0.70

成立に必要な誘導検体のファイルサイズを明らかにした．検体サイズ比率の算出方法を (2) に示す．

$$\mathrm{R_{mix}} = \frac{S_G}{S_G + S_H} \tag{2}$$

検体サイズ比率（R_{mix}）とは，手法適用後の検体のファイルサイズである誘導検体（S_G）と潜伏検体（S_H）の和のうち，誘導検体のファイルサイズが占める割合を指す．

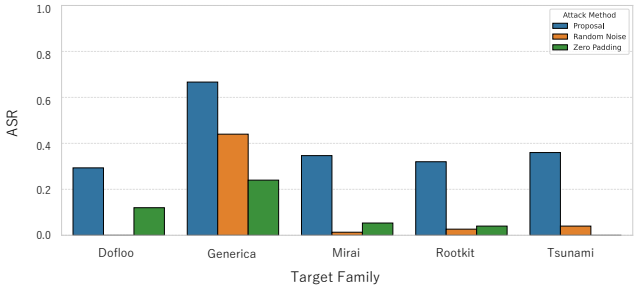
提案手法の成功可否と検体サイズ比率の関係を**図 9** に示す．これらの図から，Mirai から Tsunami への標的型攻撃を除いて，多くの組み合わせにおいて，近似曲線は緩やかな曲線を描いていることが確認できる．それぞれの組み合わせにおける臨界検体サイズ比率を**表 2** に示す．表から，攻撃成功に必要な検体サイズ比率の閾値は Generica ファミリーを除いておおよそ 0.6 から 0.75 の範囲に収まることが確認できる．

これらの結果から，提案手法が有効となる誘導検体のファイルサイズはおおよそ検体サイズ比率が 0.75 である必要がある．しかしながら，安定した成功を実現するには 0.75 よりも大きい検体サイズ比率といった量の誘導検体の混入が必要であると結論づける．

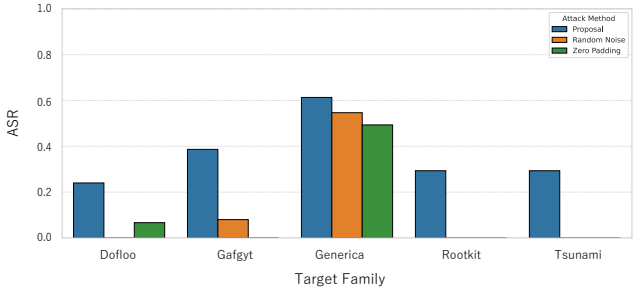
5. おわりに

本研究では，IoT マルウェア画像分類手法に対する実行可能性を維持した標的型攻撃が実現可能であるかを検証することを目的とし，異種マルウェア混入手法を提案した．提案手法では，予測確信度に基づいて選定した誘導検体のバイナリ全体を潜伏検体の新規実行可能セグメント部分に混入した．これにより，実行可能性を維持しながらターゲットファミリー固有のテクスチャを顕在化させ，従来手法への誘導効力の追加を実現した．

従来手法である Random Noise および Zero Padding と提案手法の ASR を比較した結果，従来手法では ASR が 0.11，0.10 だったのに対し，提案手法では ASR が 0.38 であった．さらに各ファミリーの組み合わせで ASR を比較した結果，全てのファミリーの組み合わせで提案手法が従

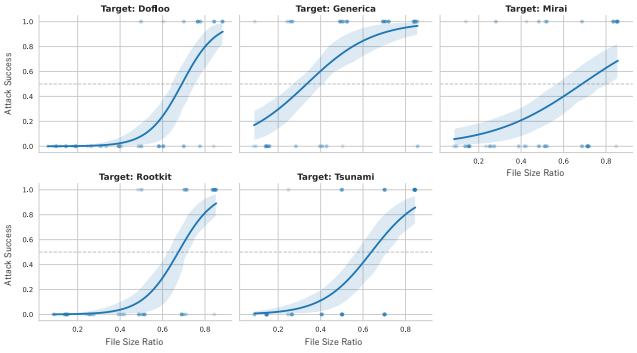


(a) 潜伏検体ファミリー: Gafgyt

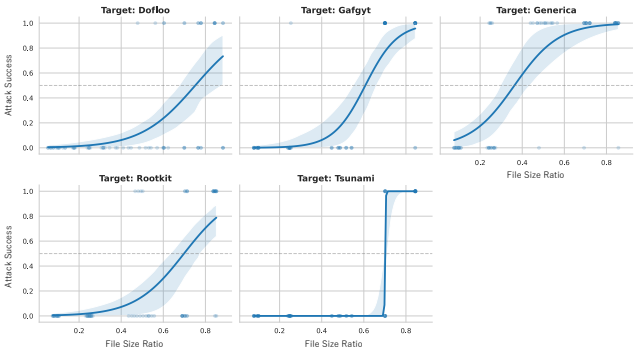


(b) 潜伏検体ファミリー: Mirai

図 8: 各ファミリーの ASR 比較



(a) 潜伏検体ファミリー: Gafgyt



(b) 潜伏検体ファミリー: Mirai

図 9: 提案手法の成功可否と検体サイズ比率の関係

来手法の ASR を上回った．このことから，異種マルウェア混入手法における誘導効力の追加が標的型攻撃の成功に寄与することを示した．

提案手法の成功可否と検体サイズ比率の関係を分析した結果，提案手法の成功には，検体サイズ比率がおおよそ 0.75 必要であることを示した．さらに，安定した成功を実現するには 0.75 より大きい検体サイズ比率といった大量の誘導検体の混入が必要であることを示した．

これらの評価を通じて，IoT マルウェア画像分類に対する実行可能性を維持した標的型攻撃は実現可能であり，現実的な脅威となり得ることを示した．今後は本研究の提案手法に対応した，IoT マルウェア解析基盤の堅牢性をより向上させていくための設計指針を考案する必要がある．

参考文献

[1] 国立研究開発法人情報通信研究機構 (NICT): NICTER 観測レポート 2025, https://www.nict.go.jp/press/2026/02/05-1.html, (2026), (Accessed on 05/15/2026).

[2] Chao, L., Zhibin, Z. and Cecilia, H.: Old Wine in the New Bottle: Mirai Variant Targets Multiple IoT Devices, https://unit42.paloaltonetworks.com/mirai-variant-iz1h9/, (2023), (Accessed on 05/15/2026).

[3] Su, J., Vasconcellos, D. V., Prasad, S., Sgandurra, D., Feng, Y. and Sakurai, K.: Lightweight classification of IoT malware based on image recognition, *2018 IEEE 42Nd annual computer software and applications conference (COMPSAC)*, Vol. 2, IEEE, pp. 664–669 (2018).

[4] Singh, A., Handa, A., Kumar, N. and Shukla, S. K.: Malware Classification Using Image Representation, *Cyber Security Cryptography and Machine Learning* (Dolev, S., Hendler, D., Lodha, S. and Yung, M., eds.), Cham, pp. 75–92 (2019).

[5] Li, D. and Li, Q.: Adversarial Deep Ensemble: Evasion Attacks and Defenses for Malware Detection, *IEEE Transactions on Information Forensics and Security*, Vol. 15, pp. 3886–3900 (2020).

[6] Khamaiseh, S. Y., Bagagem, D., Al-Alaj, A., Mancino, M. and Alomari, H. W.: Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification, *IEEE Access*, Vol. 10, pp. 102266–102291 (2022).

[7] Vasan, D., Alazab, M., Wassan, S., Safaei, B. and Zheng, Q.: Image-Based Malware Classification Using Ensemble of CNN Architectures (IMCEC), *Computers & Security*, Vol. 92, p. 101748 (2020).

[8] Alkhateeb, E., Ghorbani, A. and Lashkari, A. H.: Packed malware detection using grayscale binary-to-image representations, *arXiv preprint arXiv:2512.15414* (2025).

[9] Prima, B. and Bouhorma, M.: Using transfer learning for malware classification, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 44, pp. 343–349 (2020).

[10] Park, D., Khan, H. and Yener, B.: Generation & evaluation of adversarial examples for malware obfuscation, *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, IEEE, pp. 1283–1290 (2019).

[11] 川田 隼大，稲村 浩，石田 繁巳: IoT マルウェア画像分類

手法に対する実行可能なノイズ付与による攻撃手法の検討，情報処理学会第 87 回全国大会講演論文集 (2025).
[12] Alzaidy, S. and Binsalleeh, H.: Adversarial Attacks with Defense Mechanisms on Convolutional Neural Networks and Recurrent Neural Networks for Malware Classification, *Applied Sciences*, Vol. 14, No. 4, p. 1673 (online), DOI: 10.3390/app14041673 (2024).
[13] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A.: Learning deep features for discriminative localization, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929 (2016).
[14] Thomas, R.: LIEF - Library to Instrument Executable Formats, https://lief.quarkslab.com/ (2017), (Accessed on 05/15/2026).
[15] Olsen, S. H. and OConnor, T.: Toward a Labeled Dataset of IoT Malware Features, *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, pp. 924–933 (2023).
[16] Sebastián, M., Rivera, R., Kotzias, P. and Caballero, J.: Avclass: A tool for massive malware labeling, *International symposium on research in attacks, intrusions, and defenses*, Springer, pp. 230–253 (2016).
[17] Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J. A., Invernizzi, L. and Kallitsis, M.: Understanding the Mirai Botnet, *26th USENIX Security Symposium (USENIX Security 17)*, pp. 1093–1110 (2017).