訓練データへの実行可能なノイズ付与による IoT マルウェア画像分類手法の堅牢性向上の検証

川田 隼大[1,a)] 稲村 浩[2] 石田 繁巳[2]

概要：マルウェアをグレースケール画像に変換し，CNN を用いて分類する手法が注目されている．このような分類手法は軽量かつ高い精度を示す一方，画像へノイズを加えることで誤分類を引き起こす脆弱性を持つ．本研究では，画像へのノイズ追加に耐性を持たせるため，分類精度と誤分類検体の予測確信度を目的関数とした訓練データの構成比最適化手法を提案する．提案手法の有効性を検証するため，ベースライン画像分類器，単一ノイズ種で構成された画像分類器，最適構成の画像分類器の三種を用いて比較実験を行った．その結果，訓練データへのノイズ追加によって分類精度が改善され，さらに構成比を最適化することで，誤分類検体における予測確信度が抑制され，堅牢性向上に有効なことが示された．

1. はじめに

従来，マルウェアが主に感染対象としていた PC は Windows OS を搭載したものであった．しかし Windows OS 自体のセキュリティ向上やアンチウイルスソフトの普及により，Windows OS を狙った攻撃の難易度は高い．その結果，ここ数年ではインターネットに繋がった監視カメラや Wi-Fi ルータといった IoT 機器が，新たな主要な攻撃対象として注目されている．Zscalar が実施した 2023 年度の調査によれば，IoT 機器を対象としたマルウェア攻撃の件数は，2022 年度と比較して 400%以上増加している [1]．IoT 機器のセキュリティ対策に関しては管理者の目が届きづらい，初期設定のままでの使用等の理由から，攻撃者にとって格好の攻撃対象となっている．

IoT マルウェアの急速な増加の背景には，ソースコードの一部を改変して作成される亜種マルウェアの存在が挙げられる．例えば IoT マルウェアである Mirai ではソースコードがインターネット上に公開されたことにより，多数の亜種が作成され，甚大な被害をもたらしたとされている [2]．このような状況を踏まえると，流通しているマルウェアの種類や特徴といった動向を把握することが，IoT マルウェアへの効果的な対策を確立する上で重要である．なぜならば，動向を的確に捉えることで，現実の脅威に即した防御策や検知技術の設計が可能となるためである．

IoT マルウェアの動向を把握する手段として，マルウェアの分類が有用である．マルウェアの分類は，オリジナルのマルウェアとオリジナルの一部を改変して作成されたマルウェアを，1 つのマルウェアファミリーとしてグループ化し，解析対象のマルウェアがどのグループに属するかを分類するものである．深層学習を用いてマルウェア分類を行うことにより，高精度かつ人的リソースを消費しない高速な分類が実現されている [3]．

深層学習を用いたマルウェア分類手法の 1 つとして，マルウェアを画像化し CNN（Convolutional Neural Network）をベースとしたモデルで分類する手法が提案されている [3]．しかし，CNN ベースの画像分類手法は，画像にノイズを追加することで誤分類を引き起こす脆弱性を持っており，堅牢性の観点から課題が残されている．堅牢性を向上させる基本的な対策として，ノイズを加えた画像を訓練データに取り入れ耐性を獲得する方法が考えられるが，ノイズを加えた画像を訓練データに取り入れると，訓練データの総数が増加する．訓練データ総数の増加は，計算資源や学習時間といった学習コストの増大につながる．

学習コストが増大する課題を解決するためには，訓練データの総数を固定した上で，各ノイズ種別ごとの訓練データにおける構成比を適切に決定する必要がある．本研究における構成比とは，各種ノイズ付与手法を適用した訓練用サンプル群の，データセット内における比率を指す．

そこで本研究では，訓練データの総数固定という制約下での堅牢性向上を実現することを目的に，訓練データの更新における構成比算出手法を提案し，その有効性を検討す

[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
Grad. Sch. Systems Information Science, Future Univ. Hakodate
[2] 公立はこだて未来大学システム情報科学部
Sch. Systems Information Science, Future Univ. Hakodate
[a)] g2125026@fun.ac.jp

Algorithm 1 マルウェアバイナリの画像化

```
function CREATE_IMAGE(binary)
    s ← √sizeof(binary)
    image[s][s] ← 0
    for y = 0 to s − 1 do
        for x = 0 to s − 1 do
            image[y][x] ← binary[y × s + x]
        end for
    end for
    return image
end function
```

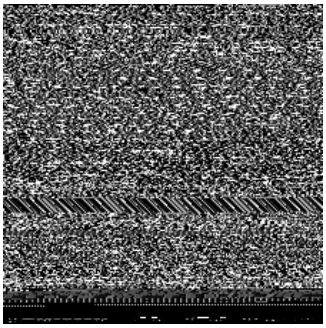


図 1: マルウェアのグレースケール画像

る．本研究の貢献は，訓練データの総数を増加させずに堅牢性向上を実現することにより，計算資源や学習時間といった学習コストを維持した手法を示した点である．

本稿の構成は以下の通りである．まず，第 2 節にて本稿を理解する上で必要となるマルウェア画像分類手法の原理について説明する．第 3 節ではマルウェア画像分類器の堅牢性向上に関する関連研究を示し，第 4 節で訓練データ更新における構成比最適化の提案を説明する．第 5 節で実験的評価を行い，最後に第 6 節にてまとめとする．

2. マルウェア画像分類手法

本節では，本稿の理解を助けるためのマルウェア画像分類手法について述べる．

2.1 マルウェアバイナリの画像化

マルウェアバイナリの画像化アルゴリズムを **Algorithm 1** に示す．マルウェア画像分類手法では，マルウェア実行形式バイナリの 1 バイトを 1 ピクセルとしたグレースケール画像に変換し，ファイルサイズに応じて一辺の長さを決定した正方形画像とする．この処理により，バイナリの構造を視覚的なパターンとして表現することが可能になる．IoT マルウェア実行形式バイナリをグレースケール画像に変換した例を**図 1** に示す．ここで **Algorithm 1** では，マルウェアごとに異なるサイズの画像が生成されるが，CNN は入力サイズを統一する必要があるため，リサイズを行う．

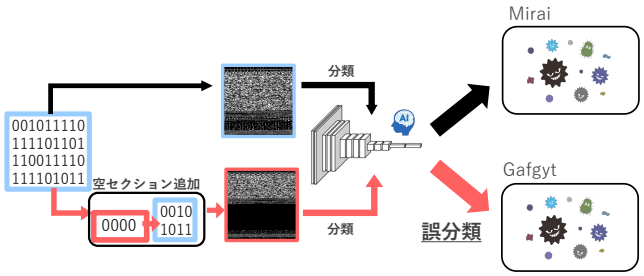


図 2: 空セクションを用いたノイズの追加の概要

2.2 特徴抽出と分類モデル

変換された画像に対しては，VGG19 などといった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間に見られるセクション構造やコード分布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 [3,4] においても有効性が報告されている．

さらに画像分類モデルに事前学習済みの重みを用いたファインチューニングを行うことで，モデルをマルウェア画像分類タスクに適応させることが可能である．このため，限られた学習データにおいても，転移学習の恩恵により精度の向上が期待できる．

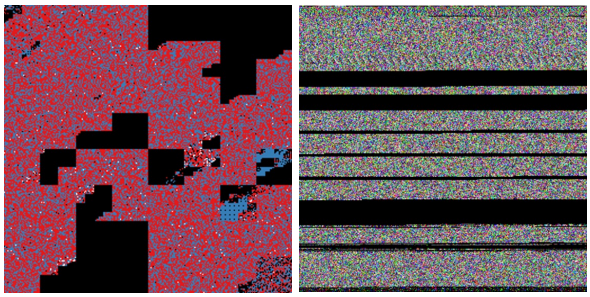
2.3 マルウェア画像分類手法の課題

マルウェア画像分類手法には課題も存在する．課題の 1 つとして，軽微な変形や敵対的なノイズの追加によって分類性能が低下するリスクが挙げられる．

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ノイズの追加によって誤分類が発生し，ノイズ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 [5]．空セクション追加を用いたノイズ付与の概要を**図 2** に示す．空セクションとは，全てのデータが 0 であるセクションのことであり，**Algorithm 1** を用いてグレースケール画像に変換すると，黒の領域が増加する．黒の領域の増加が分類におけるノイズとして働き，誤分類が発生していると考えられる．

小久保らは Windows マルウェアに対して敵対的なパッチを挿入することによる攻撃を行い，誤分類が発生することを明らかにした [6]．この手法は，実行ファイルの機能に影響を与えない範囲で敵対的なノイズを追加するものであり，機械学習モデルを用いた Windows マルウェア分類の脆弱性を明らかにした．

Gu らは Android マルウェアに対してヘッダー部分を除くセクションを対象に，任意の 1 ピクセルを変更する One-Pixel Attack [7] を適用することで，Android マルウェア画像分類器で誤分類が発生することを示した [8]．これにより極めて小さなノイズの追加であっても，マルウェア



(a) 空間充填曲線 z-order　　(b) バイトプロット

図 3: Reilly 手法で生成された敵対的サンプル（文献 [10] より引用）

画像分類器の分類精度に大きな影響を与えることが示された．

文献 [5,6,8] より，軽微な変形や敵対的なノイズの追加への対策を明らかにすることは，マルウェア画像分類手法の堅牢性向上における重要な研究課題である．

3. 関連研究

3.1 データセット更新による堅牢性向上

Marastoni らはマルウェア画像分類手法の分類精度向上および難読化されたマルウェアへの耐性を持たせるため，データ拡張と転移学習を活用する手法を提案している [9]．この提案では，マルウェアのバイナリコードの一部に難読化を適用することでデータ数を増やしている．さらに，拡張したデータセットで CNN や Bi-LSTM などの深層学習モデルを訓練し，大規模マルウェアデータセットに対して転移学習させることで，約 98.5% の高い精度を達成した．

Reilly らは GAN（Generative Adversarial Networks）を用いて，敵対的サンプルを生成し，訓練データに追加することによる堅牢性の向上を提案した [10]．この研究では，空間充填曲線 z-order やバイトプロットによって変換されたマルウェア画像を用いた分類器を対象としている．この手法によって生成された敵対的サンプルを**図 3** に示す．元のモデルでは PGD 攻撃によるモデルの正解率が約 95% から約 4.5% に低下するのに対し，敵対的サンプルを訓練データに加えたモデルでは，攻撃後の正解率が大幅に回復し，およそ 70%以上を維持することを示した．

文献 [9,10] の研究では，マルウェア画像分類手法における堅牢性向上を目的として，難読化済みデータや敵対的サンプルを訓練データに追加することでモデルの堅牢性を向上させている．しかし，これらの研究ではノイズを含むデータを訓練データに追加することのの有効性を示しているものであり，ノイズを含むデータをどの程度訓練に用いるべきかという課題に対しては，議論が行われていない．これに対し本研究は，訓練データにおける最適な構成比を明らかにすることで課題を解決する．従来の研究では固定的に扱われていた訓練データの構成比を本研究では制御可能とし，モデルの汎化性能と攻撃耐性のバランスを柔軟に調整することを実現する．

3.2 ノイズ検知・除去による堅牢性向上

Li らは，マルウェア分類における敵対的サンプルへの耐性を強化するために，ハッシュ変換と DAE（Denoising Autoencoder）を組み合わせた HashTran-DNN フレームワークを提案した [11]．この手法では，マルウェアサンプルにハッシュ関数を適用し，その後 DAE を用いてノイズを除去することで，分類モデルの堅牢性を向上させている．実験結果から，提案手法が複数の敵対的攻撃手法に対して有効であることが確認された．

この研究では，特定のノイズや攻撃手法に対する検知・除去機構を導入することで，マルウェア分類モデルの堅牢性を高めている．しかし，ノイズに対応する専用の検知・除去機構を追加することは，システムの複雑化や学習コストの増大が懸念される．

これに対し，本研究では，訓練データにノイズを付与した画像を加えることで，分類モデルの堅牢性を向上させる手法を提案する．特に，ノイズを含むデータと通常のデータの最適な訓練データ構成比を明らかにすることで，学習コストの増大を伴わない堅牢性向上を実現する．

3.3 モデル構造変更による堅牢性向上

Ravi らは，マルウェア分類の堅牢性を向上させるために，注意機構を組み込んだ CNN モデルを提案している [12]．この手法では，マルウェア画像中の重要な領域に焦点を当てることで，分類精度と堅牢性の向上を図っている．実験により，提案手法が従来の CNN モデルと比較して，敵対的攻撃に対する耐性が向上することが示された．

Shao らは，マルウェア画像分類モデルの堅牢性を向上させるために，深層残差ネットワークとハイブリッド注意機構を組み合わせた手法を提案している [13]．この手法では，チャネル注意機構と空間注意機構を組み合わせることで，マルウェア画像中の重要な特徴に焦点を当て，不要な情報を抑制することを目的としている．実験により，提案手法が従来の CNN モデルと比較して，分類精度と堅牢性の両面で優れていることが示された．

これらの研究では，マルウェア画像中の重要な領域に焦点を当てるために注意機構を導入しており，モデル構造が複雑化している．これにより，分類精度や堅牢性の向上が報告されている一方で，計算グラフが深くなり，パラメータ数や演算量が増加するため，学習コストも増大するという課題がある．

これに対し，本研究ではモデル構造を変更せず，訓練データの構成比を最適化するという手法を採用している．このため，注意機構などを追加することなく，学習コストの増大を伴わない堅牢性向上を実現している点で，モデル

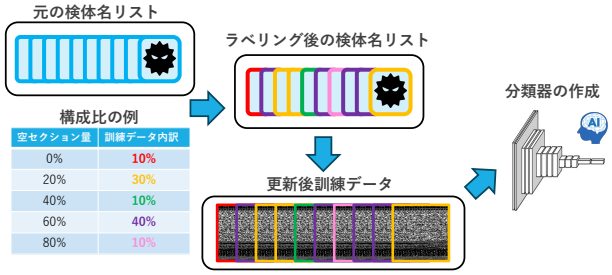


図 4: 構成比に基づく訓練データ生成の概要

構造の変更とは異なるアプローチである．

4. 訓練データ更新における構成比最適化の提案

本研究では，分類精度を最大化しつつ，誤分類検体の予測確信度を最小化するという，2 つの目的を満たす構成比を多目的最適化によって探索しそれに基づく訓練データで分類器を作成する手法を提案する．

4.1 構成比の定義

本研究では訓練データの構成比を最適化し，分類器の堅牢性向上を実現する．本節では構成比の定式化について説明する．本研究における構成比は以下の n 次元の実数ベクトルとして定義する．

$$\boldsymbol{x} = (x_1, x_2, ..., x_n), \quad \begin{cases} \sum_{i=1}^{n} x_i = 1 \\ 0 \leq x_i \leq 1 \end{cases} \tag{1}$$

n は対象とするノイズ種の数であり，マルウェア画像分類器に耐性を持たせるノイズの種類に応じて任意に決定する．$\boldsymbol{x}$ の各要素は対応するノイズ種が訓練データ全体に占める割合を示す．

4.2 構成比に基づく訓練データ生成

構成比に基づく訓練データ生成の概要を**図 4** に示す．構成比に基づく訓練データ生成では，まず従来の訓練データセットに含まれる各検体に対し，構成比に応じたラベリングを行う．このラベルは，データセット生成時に付与するノイズの種類を示す．ラベリングの方法としては，構成比に従って訓練データを n 個のグループにランダム分割する．例えば，耐性を持たせるノイズ種として $noiseA$，$noiseB$ がある場合，構成比は $\boldsymbol{x} = (original, noiseA, noiseB)$ という式で表現される．ここで構成比が $\boldsymbol{x} = (0.1, 0.2, 0.7)$ である場合，訓練データの 10%にはノイズ付与なし（$Original$）をラベリング，訓練データの 20%には $noiseA$ をラベリング，訓練データの 70%には $noiseB$ をラベリングする．

その後ラベリング結果にしたがって，各検体に対応するノイズを加えた画像を生成し，それを新たな訓練データとすることで，訓練データの生成を行う．

4.3 NSGA-II による構成比の最適化

本研究では，分類精度の最大化と誤分類検体の予測確信度の最小化という 2 つの目的を満たす構成比を算出することから，多目的最適化問題として解決を図る．多目的最適化問題を解決する手法として，NSGA-II [14] を使用し，構成比を進化的に最適化させる．

NSGA-II とは，遺伝的アルゴリズムを単一目的の最適化問題から多目的の最適化問題へ拡張したものである．このアルゴリズムではまず，個体群に対して複数の目的関数の値を評価し，非優劣ソートという手法で個体をランク付けする．非優劣ソートでは，複数の目的において他の個体に劣っていない個体を，同じランクのグループとしてまとめる．このようにして，個体は目的の達成度に応じて複数のランクに分類される．次に，同じランクに属する個体同士の分布のばらつきを示す混雑度を算出する．混雑度が大きい個体は，他の個体と比べて多様性が高い解であり，探索範囲の拡大に寄与する．その後，非優劣ソートによるランク付けの結果と混雑度をもとに，混雑度トーナメント選択を行い，次世代の個体を生成するための親個体を選択する．この選択では，ランクが高い個体が優先され，同じランクの場合は混雑度が大きい個体が選ばれる．本研究では，構成比を示す n 次元の実数ベクトルの各要素を NSGA-II における遺伝子，ベクトルを NSGA-II における個体とする．

構成比の最適化に必要となる分類精度の取得は，各構成比に基づいて構成された訓練データでモデルを学習した後，検証データに対して予測を行い，正しく分類された検体の割合として算出する．誤分類検体の予測確信度は，検証データの予測において誤って分類された各検体に対してモデルが予測したクラスの最終出力層の値を集計し，その平均値として算出する．

4.4 最優秀個体の選定

最優秀個体の選定にあたっては，まず NSGA-II によって得られた各試行のパレート最適解を収集する．次に収集したパレート最適解に対し，重みに基づいたスコアを算出する．その後スコアが最も高かった個体を最優秀個体として採用する．スコアの算出方法を以下に示す．

$$\text{Score} = \alpha \cdot \text{Accuracy} - \beta \cdot \text{Confidence} \tag{2}$$

α，β は分類精度の最大化を重視するか，誤分類検体の予測確信度の最小化を重視するかによって調整可能だが，本研究では両方の目的関数を同等に重視する方針とし，それぞれ 0.5 に設定する．

5. 評価

本研究では訓練データの構成比を最適化する提案手法によって作成された最適構成分類器の有効性を評価する．評

表 1: ラベリングの内訳

ファミリ名	検体数	割合（%）
benign	354	3.16
Mirai	5444	48.55
Gafgyt	3779	33.70
Tsunami	266	2.37
Rootkit	279	2.49
Generica	770	6.87
Flooder	118	1.05
Dofloo	204	1.82

表 2: 抽出したデータセットの内訳

データタイプ	検体数	ファミリあたりの検体数
訓練用	805	161
検証用	115	23
テスト用	230	46

価ではまず，ノイズ追加による攻撃に耐性のないベースライン画像分類器と単一ノイズ種で構成された分類器をそれぞれ作成し比較する．その後ベースライン画像分類器と最適構成分類器を比較する．これにより，訓練データへのノイズ追加による性能の向上と，訓練データへの最適化された複数種のノイズ追加によって構成される分類器の有効性を検証する．

評価では，事前に耐性を持たせるノイズ追加手法を定める必要がある．このため，評価では空セクションの追加 [5] を使用する．具体的には，GNU Binutils の objcopy コマンドを用いて，ELF ファイルに新たなセクションを付加する．セクション追加時には，noload および readonly フラグを設定しており，実行時にロードされず，書き込みも行われないことを保証している．これにより，実行可能性を維持したノイズ追加手法を適用している．

5.1 使用するデータセット

本研究では，Olsen らによって提供されたラベル付き IoT マルウェアデータセット [15] を使用した．このデータセットには複数アーキテクチャの IoT マルウェアが含まれている．アーキテクチャごとに画像特徴が大きく異なるため，マルウェア画像分類器の作成においては，単一のアーキテクチャを使用する必要がある．そこで評価においては，Intel 80386 アーキテクチャの検体を対象とした．Intel80386 アーキテクチャにおけるマルウェアファミリごとの検体数の内訳を**表 1** に示す．分類対象とするマルウェアファミリは，検体数が十分に存在する Mirai，Gafgyt，Tsunami，Rootkit，Generica の 5 種とし，各ファミリからランダムに 230 検体を抽出した．これにより，各ファミリの検体数が均等となるよう調整した実験用のデータセットを新たに作成した．実験用のデータセットにおける検体数の内訳を**表 2** に示す．抽出後のデータは，訓練用 70%，

表 3: ベースライン画像分類器の分類性能

Null Size	Accuracy	Precision	Recall	F1-score	Confidence
0%	0.852	0.864	0.852	0.852	0.810
20%	0.647	0.747	0.647	0.620	0.905
40%	0.578	0.639	0.578	0.565	0.790
60%	0.530	0.612	0.530	0.545	0.838
80%	0.395	0.573	0.395	0.386	0.888
Average	0.600	0.687	0.600	0.594	0.846

テスト用 20%，検証用 10%の割合で分割した．訓練用はモデルの学習に，テスト用は作成した分類器の評価に，検証用は NSGA-Ⅱ による最適化に必要となる分類精度と誤分類検体の予測確信度の算出に用いた．

5.2 ベースライン画像分類器の作成

マルウェア画像分類器の作成には，Google Colab 環境上で TensorFlow および Keras を用いた．ベースモデルとして VGG19 を採用し，ファインチューニングすることで作成した．モデルの訓練では，損失関数として categorical_crossentropy を，最適化手法として Adam を採用し，エポック数は 30，バッチサイズは 8 とした．ベースライン分類器の訓練には，空セクションが追加されていない画像のみを使用し，訓練データへのノイズ追加がされていない従来データセットでの分類器作成を再現した．

作成したベースライン画像分類器をテスト用データを用いて評価した．分類性能を**表 3** に示す．表における Accuracy は分類精度，Precision は適合率，Recall は再現率，F1-Score は適合率と再現率の調和平均を表す．各指標は式 (3)〜(6) で算出した．

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

表 3 で，Confidence は誤分類検体における予測確信度の平均を表している．本研究における分類モデルでは最終出力層に活性化関数として softmax 関数を適用しており，各クラスに属する確率を出力する．誤分類検体における予測確信度は，モデルが誤って分類したクラスに対応する softmax 出力の値を取得する．

表を見ると，先行研究 [5] と同様に，追加する空セクション量の増加によって分類精度は低下しており．F1-score は 0%の 0.852 から 80%の 0.395 まで低下した．さらに追加す

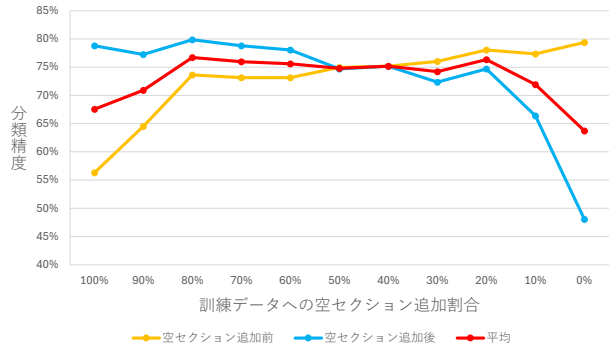


図 5: ノイズ追加画像の割合に対する分類精度の変化

表 4: 比較対象となる分類器の分類性能

Null Size	Accuracy	Precision	Recall	F1-score	Confidence
0%	0.873	0.876	0.873	0.874	0.901
20%	0.773	0.808	0.773	0.766	0.944
40%	0.843	0.850	0.843	0.842	0.902
60%	0.830	0.849	0.830	0.829	0.883
80%	0.921	0.926	0.921	0.922	0.800
Average	0.848	0.862	0.848	0.847	0.886

る空セクション量の増加によって，Confidence は増加しており，0%では 0.810 だったものが，80%では 0.888 となっている．これは，モデルがノイズが追加された画像に対しても高い確信をもって誤った予測をしていることを示しており，堅牢性が欠如していることがわかる．平均値を見ると，Accuracy は 0.600，F1-score は 0.594，Confidence は 0.846 という性能にとどまっている．提案手法の評価では，この分類精度を基準に構成比の最適化の有効性を評価する．特に，ノイズを含む検体に対しても安定した分類精度を維持しつつ，誤分類検体における予測確信度を低く抑えることが可能かどうかを検証することで，提案手法の有効性を明らかにする．

5.3 単一ノイズ種で構成された画像分類器の作成

構成比の最適化アプローチは堅牢性向上が見込める一方，複数のノイズ種を用いることによる学習コストや設計の複雑さが課題となる．もし，単一のノイズ種で構成された画像分類器で十分な堅牢性を実現できれば，その方が堅牢性向上において効率的である．

そこで本研究では，ベースライン画像分類器とは異なる分類器として，単一ノイズ種を訓練データへ追加した画像分類器を作成した．作成においては，空セクション追加によって生成したノイズ追加画像の割合を，0%，10%，20%，30%，…，100%と変化させた分類器の分類精度を比較し．最大となった分類器を選択した．

作成の際には，ベースライン画像分類器と同様のモデル構造，損失関数，最適化手法，ハイパーパラメータを用いた．本評価においては，単一ノイズ種とする空セクション量は，元の IoT マルウェアのファイルサイズの 80%に設定した．

図 5 に評価結果を示す．グラフの横軸は訓練データにおけるノイズ追加画像の割合，縦軸は分類精度である．評価結果で示されているように，ノイズ追加画像が訓練データの 80%を占める構成比において，ノイズ追加前および追加後の検体に対する分類精度の平均 76.73%となり，最も高い分類精度を示した．

以上の理由から，本研究では構成比を（0.200, 0.000, 0.000, 0.000, 0.800）とした分類器を，単一ノイズ種で構成された画像分類器として採用する．先述の構成比で作成した分類器の分類性能を**表 4** に示す．

表を見ると，ベースライン画像分類器と比較して空セクション追加の影響を受ける状況下においても高い分類精度を維持できている．特に空セクション量が 80%の検体に対しては Accuracy が 0.921，F1-score が 0.922 とベースライン画像分類器からの改善が見られる．平均値においても，Accuracy が 0.848，F1-score は 0.847 といずれもベースライン画像分類器を上回る結果となった．

一方，空セクション量が 20%〜60%といった比較的少量の空セクション追加に対する分類精度はベースライン画像分類器と比べれば向上しているものの，その改善幅は 80%に比べれば限定的であり，精度向上が 80%のノイズの検体に偏っている傾向が見られる．さらに，Confidence は 20%で 0.944，40%で 0.902 と非常に高く平均値においても，ベースライン画像分類器からの増加が見られた．これは分類器が軽度な空セクション追加に対して過剰な自信を持ち，堅牢性がベースライン画像分類器よりも低い可能性を示している．

まとめると，単一ノイズ種での訓練データ構成は画像分類器の分類精度を高めるものであるが，画像分類器の堅牢性を低下させる要因になると考えられる．このため，複数ノイズ種の構成比を最適化し，ベースライン分類器からの堅牢性の低下を避けつつ，分類精度を向上させることが望まれる．

5.4 構成比の最適化

本研究では NSGA-Ⅱ を使用して構成比の最適化を行う．NSGA-Ⅱ による最適化の際には DEAP ライブラリを使用した実装を行った．設定した NSGA-Ⅱ のパラメータは，個体数 32，世代数 5，交叉確率 1.0，突然変異率 0.2 とした．遺伝子は構成比の 5 次元実数ベクトルとした．これは空セクションの追加量が異なる 4 種のノイズ追加画像と空セクションを追加していない画像の構成比を最適化するためである．

評価における構成比の最適化では，空セクションの追加量を変化させた 4 種類のノイズに対する評価を行う．具体

表 5: 上位 10 個体の構成比

x_1	x_2	x_3	x_4	x_5	Accuracy	Confidence
0.12	0.16	0.25	0.30	0.17	0.849	0.776
0.18	0.16	0.23	0.24	0.19	0.816	0.792
0.15	0.14	0.32	0.24	0.15	0.800	0.777
0.33	0.10	0.26	0.13	0.19	0.793	0.775
0.34	0.13	0.05	0.36	0.12	0.805	0.787
0.18	0.11	0.23	0.30	0.18	0.790	0.785
0.35	0.00	0.25	0.36	0.04	0.809	0.805
0.21	0.54	0.07	0.14	0.03	0.812	0.809
0.06	0.15	0.33	0.10	0.35	0.730	0.731
0.30	0.32	0.29	0.08	0.02	0.734	0.736

的には，元の IoT マルウェアバイナリのファイルサイズに対して空セクションが占める割合を 20%，40%，60%，80%と設定し，それぞれ 4 種類のノイズ強度を定義する．これに加え，空セクションを追加しない元のファイルも含め，構成比の最適化を行い，その有効性を評価する．

以上のパラメータおよび遺伝子の設定とし 6 回の試行でパレート解を収集した．収集したパレート解には重みづけランキングを実施し，上位 10 個体を選定した．収集したパレート解の分布および上位 10 個体のパレート解の分布を**図 6** に示す．さらに，上位 10 個体のパレート解の構成比と，それぞれの Accuracy および Confidence を降順に上から並べた結果を**表 5** に示す．

図を見ると，Accuracy と Confidence の間には正の相関関係が見られた．これは，分類精度を向上させることと誤分類検体の予測確信度を抑制することはトレードオフの関係にあることを示している．さらに，重みづけランキングを行った結果，上位の個体は Accuracy が 0.8 程度，Confidence が 0.75 から 0.8 の領域に集中していることが見られた．

表を見ると，最も高い分類精度を示した個体は，各要素の値域が 0.1 から 0.3 となっており，いずれかに極端に偏ることない構成であった．このような傾向は他の上位個体にも共通しており，すべての空セクション追加割合を適度に含めた構成が，分類性能向上に寄与している可能性を示している．さらに，空セクション量が 40%および 60%の比率がわずかに高めに設定された個体が上位に多く見られた．これにより，中程度のノイズに対する学習が分類器の堅牢性向上に有効であることが考えられる．一方で，空セクション量が 20%および 80%の構成比率が高い構成は，Accuracy および Confidence がともに低下しており，極端な構成比が性能劣化を引き起こす可能性を示している．

以上より，構成比の最適化によって，空セクション追加の程度が異なる多様な検体を同程度の比率で訓練データに取り入れることが，堅牢性の向上に有効であることが明らかとなった．

表 6: 最適構成の画像分類器の分類性能

Null Size	Accuracy	Precision	Recall	F1-score	Confidence
0%	0.862	0.864	0.862	0.852	0.811
20%	0.867	0.871	0.867	0.854	0.848
40%	0.870	0.880	0.870	0.867	0.839
60%	0.887	0.897	0.887	0.886	0.803
80%	0.870	0.881	0.870	0.868	0.812
Average	0.871	0.879	0.867	0.866	0.823

5.5 最優秀構成比での分類器作成

最終的に選定された最優秀構成比を用いて新たに最適構成の画像分類器を作成した．最適構成の画像分類器の分類性能を**表 6** に示す．表を見ると，最適構成の画像分類器はベースライン画像分類器および単一ノイズ種で構成された画像分類器と比較して，分類精度の改善が見られた．これは全ての空セクション追加量で安定して高い性能を示しており，特に空セクション量が 60%で 0.887，40%で 0.870 となっている．これにより単一ノイズ種で構成された画像分類器に比べて，中間的な空セクション量への対応力が向上しており，構成比の最適化により訓練データのバランスが改善された結果と考えられる．

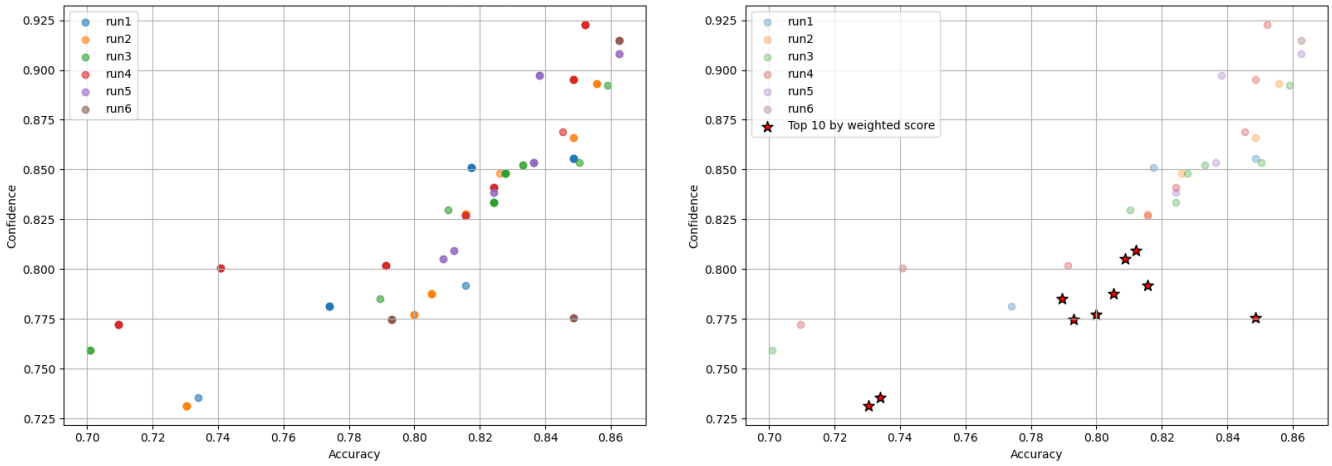
誤分類検体の予測確信度に着目すると，最適構成分類器は平均 Confidence が 0.823 と最も低く，0%で 0.811，60%で 0.803，80%で 0.812 など，幅広い空セクション量で Confidence を抑制しており，単一ノイズ種で構成された画像分類器と比較して誤分類時に過度な確信を示すことなく予測の不確かさを適切に反映している．このように，最適構成分類器は精度だけでなく予測確信度の観点からも最も信頼性の高い分類器であり，構成比最適化による堅牢性向上の有効性を示している．

6. おわりに

本稿では，従来の訓練データでは誤分類が発生する画像へのノイズ追加に対して，分類精度と誤分類検体の予測確信度の最適化による構成比算出手法を提案し，その有効性を検討した．

提案手法の有効性を検証するために，ベースライン分類器，単一ノイズ種で構成された分類器，最適構成の画像分類器の三種類の分類器を作成し，それぞれの分類性能を比較した．

評価の結果，ノイズ追加画像の誤分類による分類性能の低下に対して，訓練データへノイズを追加することにより，分類精度を改善することができた．さらに，最適化した構成比によって生成された訓練データを用いて作成された分類器では，誤分類検体における予測確信度を単一ノイズ種で構成された分類器と比較して抑制することができた．このため，訓練データの構成比の最適化が画像へのノイズ追加に対する堅牢性向上に貢献することが明らかとなった．



(a) 全試行　　(b) 上位 10 個体

図 6: 収集したパレート最適解の分布

今後は，本手法がより広範なノイズ追加手法に対しても有効であるかを評価する必要がある．さらに，より多数の検体やマルウェアファミリを持つデータセットに対しても本手法が有効であるかを評価する必要がある．

参考文献

[1] ThreatLabz. 2023 年版 zscaler threatlabz エンタープライズ iot および ot の脅威レポート— zscaler. `https://bit.ly/3UV576b`, 2023. (Accessed on 12/16/2024).

[2] Lei Chao, Zhang Zhibin, and Hu Cecilia. Old wine in the new bottle: Mirai variant targets multiple iot devices. `https://unit42.paloaltonetworks.com/mirai-variant-iz1h9/`, May 2023. (Accessed on 07/18/2024).

[3] Jiawei Su, Danilo Vasconcellos Vargas, Sanjiva Prasad, Daniele Sgandurra, Yaokai Feng, and Kouichi Sakurai. Lightweight Classification of IoT Malware based on Image Recognition, February 2018.

[4] Danish Vasan, Mamoun Alazab, Sobia Wassan, Babak Safaei, and Qin Zheng. Image-Based malware classification using ensemble of CNN architectures (IMCEC). *Computers & Security*, Vol. 92, p. 101748, 2020.

[5] 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. March 2024.

[6] 小久保恵弘, 大山恵弘. マルウェア検知器に対する敵対的パッチ攻撃におけるパッチ配置位置についての考察. 暗号と情報セキュリティシンポジウム 2023 予稿集, 2023.

[7] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, Vol. 23, No. 5, pp. 828–841, 2019.

[8] Shangyu Gu, Shaoyin Cheng, and Weiming Zhang. From Image to Code: Executable Adversarial Examples of Android Applications. In *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, pp. 261–268, Tianjin China, April 2020.

[9] Niccolò Marastoni, Roberto Giacobazzi, and Mila Dalla Preda. Data augmentation and transfer learning to classify malware images in a deep learning context. *Journal of Computer Virology and Hacking Techniques*, Vol. 17, No. 4, pp. 279–297, December 2021.

[10] Ciaran Reilly, Stephen O Shaughnessy, and Christina Thorpe. Robustness of Image-Based Malware Classification Models trained with Generative Adversarial Networks. In *Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference*, EICC '23, pp. 92–99, New York, NY, USA, June 2023.

[11] Deqiang Li, Ramesh Baral, Tao Li, Han Wang, Qianmu Li, and Shouhuai Xu. HashTran-DNN: A Framework for Enhancing Robustness of Deep Neural Networks against Adversarial Malware Samples, September 2018.

[12] Vinayakumar Ravi and Mamoun Alazab. Attention-based convolutional neural network deep learning approach for robust malware classification. *Computational Intelligence*, Vol. 39, No. 1, pp. 145–168, February 2023.

[13] Yanli Shao, Yang Lu, Dan Wei, Jinglong Fang, Feiwei Qin, and Bin Chen. Malicious Code Classification Method Based on Deep Residual Network and Hybrid Attention Mechanism for Edge Security. *Wireless Communications and Mobile Computing*, Vol. 2022, No. 1, p. 3301718, 2022.

[14] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, Vol. 6, No. 2, pp. 182–197, April 2002.

[15] Stian Hagbø Olsen and Tj OConnor. Toward a Labeled Dataset of IoT Malware Features. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 924–933, Torino, Italy, June 2023.