修士論文

XAI を活用したマルチモーダルマルウェア検出の説明可能性に関する研究

YU ZHOU

主指導教員　BEURAN RAZVAN

北陸先端科学技術大学院大学

先端科学技術研究科

（情報科学）

令和 8 年 3 月

概要

昨今、マルウェア検知の分野においてマルチモーダル深層学習は目覚ましい成果を上げている。しかしながら、モデルの高度な複雑化に伴う判断プロセスのブラックボックス化は、高い信頼性が要求されるセキュリティ運用において、無視できない重大な懸念材料となっている。とりわけ、全体の平均正解率が高いことによって、特定のクラスで発生している検知漏れ（False Negative）が統計データの中に埋没してしまう現象は、極めて重大な課題である。本研究では、マルチモーダル検出フレームワーク『MIDALF』の中核をなす画像処理モジュール『MALSSL』に焦点を当て、その挙動を精査した。その結果、ベンチマークテストでは優秀な成績を収めているにもかかわらず、混同行列を用いた詳細解析においては、複雑なテクスチャを持つ特定のマルウェアファミリを一切識別できない（正解率 0%）という、致命的な『性能の乖離（パラドックス）』が浮き彫りとなった。

この構造的な弱点の根本原因を突き止めるべく、本研究では第一段階として、モデルの内部挙動を診断する上で最も適した説明可能 AI（XAI）アルゴリズムの選定に取り組んだ。Grad-CAM、Grad-CAM++、XGrad-CAM、LayerCAM、SHAP の 5 つの主要な手法に対し、**忠実度（Fidelity）、複雑性（Complexity）、堅牢性（Robustness）**という 3 つの指標に基づく多面的な検証を行った。CNN 深層における微細な特徴情報を最も正確に保持できる手法として LayerCAM を特定し、これを診断の核として採用した。

LayerCAM による解析の結果、ベースラインとして採用した ResNet-18 バックボーンでは、特徴抽出能力が不十分であることが明らかになった。ResNet-34、50、101、および 152 の各バリアントを対象に比較検証を行い、計算効率と抽出能力のバランスが最適な

ResNet-50 を選定することで、対象ファミリを含む全体精度を 98.7% まで向上させることに成功した。

さらに、視覚的な定性評価に加え、客観的な妥当性を確保するために**「LayerCAM ＋シャノンエントロピー」**による新たな統合診断評価フレームワークを新たに策定し、実証実験を実施した。定量的な分析を行ったところ、ResNet-18 は注視領域において 6.82 bits という高いエントロピー値を記録し、暗号化ペイロード領域を的確に捕捉できていることが実証された。しかし、それにもかかわらず分類に失敗（正解率 0%）したという事実は、ResNet-18 の浅い構造では、エントロピーが高い領域に隠された複雑なパターンを、識別処理に必要なレベルの特徴表現として捉えきれていなかったことを意味する。対照的に、ResNet-50 はさらに高い 7.04 bits を達成した。このエントロピーの向上は、より深い層を持つ ResNet-50 が、ペイロード内の微細な悪性特徴に対してより強く、かつ確実な焦点を結んでいることを数学的に証明している。「本研究の貢献は、特定のマルウェアファミリにおける認識不全を解消した点にとどまらない。特筆すべきは、XAI と情報理論を統合的アプローチとして用いることで、モデルの『注視領域』の特定に加え、その『理解の深度』までも定量的に計測可能な、新しい評価の枠組み（パラダイム）を提示したことにある。

キーワード：マルウェア検出、マルチモーダル学習、ResNet-50、説明可能 AI（XAI）、エントロピー

Abstract

In recent years, multimodal deep learning has demonstrated remarkable success in the field of malware detection. However, the ”black box” nature of decision-making processes, a byproduct of increasing model complexity, poses a significant and undeniable risk in security operations where reliability is paramount. In particular, the phenomenon where false negatives in specific classes are obscured by high overall average accuracy metrics represents a critical challenge.

This study focuses on ”MALSSL,” the core image processing module of the multimodal detection framework ”MIDALF,” conducting a comprehensive analysis of its behavioral characteristics. Results indicate that while the model performs exceptionally well in benchmark tests, detailed confusion matrix analysis reveals a critical flaw: a complete inability (0% accuracy) to identify specific malware families exhibiting similar textures.

To investigate the root cause of this structural vulnerability, the first phase of this study involved selecting the most appropriate Explainable AI (XAI) algorithm for diagnosing the model's internal behavior. Through an analysis of five primary methods—Grad-CAM, Grad-CAM++, XGrad-CAM, LayerCAM, and SHAP—the results demonstrated that the ResNet-18 backbone, employed as the baseline, possesses insufficient feature extraction capabilities. Consequently, comparative verification across

ResNet variants (34, 50, 101, and 152) led to the selection of ResNet-50 for its optimal balance between computational efficiency and feature extraction, successfully elevating the overall accuracy, including the target families, to 98.7

Furthermore, validation based on three metrics—Fidelity, Complexity, and Robustness—identified LayerCAM as the method most capable of accurately retaining subtle feature information within deep CNN layers, establishing it as the core diagnostic tool.

Additionally, to ensure objective validity beyond visual qualitative assessment, this study developed a novel integrated diagnostic framework combining "LayerCAM + Shannon Entropy" and conducted empirical experiments. Quantitative analysis revealed that ResNet-18 recorded a high entropy value of 6.82 bits in the attention regions, confirming its accurate capture of encrypted payload areas. However, the persistence of classification failure (0% accuracy) implies that the shallower structure of ResNet-18 failed to translate the complex patterns hidden within these high-entropy regions into the feature representations necessary for effective identification. In contrast, ResNet-50 achieved a higher entropy value of 7.04 bits. Mathematically, this increase demonstrates that the deeper architecture of ResNet-50 establishes a stronger and more definitive focus on subtle malicious features within the payload.

Keywords: Malware Detection, Multimodal Learning, ResNet-50, Explainable AI (XAI), Entropy

目次

図目次

表目次

第 1 章

はじめに

1.1　背景

現代社会におけるデジタルエコシステムの相互接続性が深化するにつれ、マルウェアは世界のネットワークインフラに対する最大の脅威の一つとして定着している。サイバー空間における攻撃手法と防御技術の対立は、終わりの見えない軍拡競争へと激化しているのが現状である。McAfee の調査報告 [1] が示す統計データによれば、新たなマルウェアの発生数は驚異的なペースで増加しており、世界規模で年間数十億件もの攻撃事例が確認されている。このような脅威の急増は、2017 年に猛威を振るった WannaCry [3] のような甚大な経済的損失を引き起こすだけでなく、Stuxnet [2] の事例が示唆するように、産業制御システムや国家の重要インフラに対する物理的な破壊工作のリスクも孕んでいる。こうした危機的な状況に対し、従来型の防御パラダイムでは対応しきれない限界が顕在化している。

1.1.1 従来型検知技術の限界と受動的な性質

これまで、アンチウイルス製品の多くは、シグネチャマッチングやヒューリスティック分析といった静的な検知手法 [4] を主軸としてきた。しかし、これらの手法は本質的に「既知の脅威」に基づく事後対応（Reactive）のアプローチであり、攻撃に対して常に後手に回らざるを得ないという受動的な性質を持っている。セキュリティ専門家による手作業での特徴抽出プロセスは膨大な時間と労力を要するため、指数関数的に増殖する亜種への対応能力には限界がある。加えて、難読化（Obfuscation）、パッキング、ポリモーフィズム（Polymorphism）といった高度な回避技術を実装した現代のマルウェアに対しては、検知精度が著しく低下するという構造的なボトルネックが課題となっている。

1.1.2 深層学習の導入と「ブラックボックス問題」

上述した限界を打破するため、学術界および産業界ではデータ駆動型のアプローチへの転換が進み、機械学習（Machine Learning）や深層学習（Deep Learning）[5] の導入が加速している。とりわけ、画像データやバイナリシーケンスなど、異種情報を統合するマルチモーダル学習技術は、マルウェア解析の自動化プロセスに革新をもたらした。

しかしながら、モデルの構造が深層化し複雑さを増すにつれて、その内部での意思決定プロセスは高度に非線形かつ不透明なものとなり、いわゆる「ブラックボックス化」が進行している。誤検知が許されない（ゼロトレランスな）サイバーセキュリティの現場において、「なぜその判定を下したのか」という根拠を提示できないシステムは、運用上の深

刻なリスクとなる。モデルの出力が、マルウェアの本質的な特徴を捉えた結果なのか、あるいは単にデータセット内の背景ノイズに過学習した結果なのかを区別することが困難であるためである。

1.2 本研究の目的

本研究では、自己教師あり学習（SSL）を活用した最先端のフレームワークである「MIDALF（特に画像処理モジュールである MALSSL）」を主要な調査対象とする。当該モデルは、一般的なベンチマークテストにおいて極めて高い正解率を達成しているが、筆者が実施した詳細な再現実験および混同行列（Confusion Matrix）による分析の結果、特定の複雑なテクスチャを有するマルウェアファミリに対してのみ、識別能力が完全に欠如し、正解率が 0% になるという不可解な現象（高性能モデルのパラドックス）が確認された。

この事実は、統計的には高性能に見えるモデルであっても、その内部の特徴抽出器（ResNet-18）には重大な構造的死角が潜んでいることを示唆している。全体的な高スコアの陰に隠蔽されたこの致命的な脆弱性を解明し、根本的な修正を施すことが本研究の喫緊の課題である。

この「特徴抽出能力の欠如」という課題を解決するため、本研究では単なる推論に頼るのではなく、説明可能 AI（XAI）を診断ツールとして採用し、以下のプロセスで研究を展開した。

まず、Grad-CAM 等の可視化技術を用いた診断により、認識失敗の主たる原因が

ResNet-18 の受容野（Receptive Field）の狭さにあることを特定した。当該モデルは高周波なテクスチャの詳細を捉えきれず、判断の根拠が背景ノイズへと拡散していたのである。この構造的な欠陥に対処するため、本研究ではバックボーンネットワークをより深い層を持つ ResNet-50 へと拡張する手法を提案した。層と残差ブロックを増強することで、微細なテクスチャ特徴に対する捕捉能力の向上を図る。さらに、視覚的な定性評価だけでなく、数学的な客観性を確保するために、「Layer-CAM による活性化マップ」と「シャノンエントロピー（Shannon Entropy）」を融合させた新たな定量的評価指標を構築し、モデルが着目している領域の確実性を数値化した。

実験の結果、改良されたアーキテクチャはエントロピースコアにおいて 7.04 という高い数値を記録し、モデル内部の判断基準の混乱が解消されたことが実証された。本研究における主な貢献は以下の 3 点に集約される：

1. 死角の発見と実証：高性能とされるモデル MALSSL において、特定の条件下で発生する認識不全（正解率 0%）を発見し、現行の主流アーキテクチャに潜む構造的欠陥を明らかにした。
2. XAI による診断と修復：XAI を用いた診断プロセスを確立し、複雑なマルウェアテクスチャの処理における ResNet-50 の優位性を証明することで、ベースモデルの弱点を克服した。
3. 定量的評価指標の確立：解釈性の分析に対し、客観的な数学的尺度を提供する「Layer-CAM とシャノンエントロピー」を組み合わせたハイブリッド評価指標を

提案・検証した。

1.3 本論文の構成

本論文の構成は以下の全 6 章から成る。

第 1 章では、マルウェア検知を取り巻く現状の課題を概観し、本研究の目的および貢献について述べた。

第 2 章では、関連研究として、従来のマルウェア検知技術、深層学習を用いたアプローチ、および近年注目を集める説明可能 AI（XAI）の動向について整理する。

第 3 章では、調査対象である MIDALF フレームワークの構造を解説するとともに、予備実験によって発覚した「高性能モデルにおける局所的な完全な失敗（パラドックス）」について詳述し、解決すべき問題点を明確化する。

第 4 章では、提案手法について論じる。XAI を活用した診断アプローチ、バックボーンネットワークの ResNet-50 への換装、および「Layer-CAM とシャノンエントロピー」を用いた新規の定量的評価指標の設計について説明する。

第 5 章では、提案手法の有効性を検証する実験結果を示す。混同行列を用いた精度の再評価、XAI による視覚的確認、およびエントロピー指標を用いた定量的分析の結果を提示し、考察を加える。

第 6 章では、本研究の総括を行うとともに、将来の展望について言及する。

第 2 章

関連研究

2.1　マルウェア検知技術の進化：単一モダリティからマルチモーダルへ

近年、マルウェアの対抗技術が高度化するにつれ、従来の静的シグネチャマッチングや動的サンドボックス解析は大きな課題に直面している。膨大かつ複雑な悪性サンプルに対応するため、研究領域は単一モダリティ（グレースケール画像やオペコード列のみ等）への依存から、マルチモーダル融合検知へのパラダイムシフトを遂げている。本節では、この技術的進化の道筋を概観し、特に自己教師あり学習の応用とその潜在的な構造的欠陥について論じる。

2.1.1 グレースケール画像に基づく可視化検知手法

マルウェアのバイナリファイルをグレースケール画像へマッピングする手法は、可視化検知における先駆的な研究である。Nataraj ら [6] は、バイナリファイルのバイト値（0-255）を画像の画素値として直接マッピングする手法を初めて提案した。このマッピングは悪性コードの元のレイアウト構造を保持するため、同一ファミリのマルウェアはテクスチャやレイアウトにおいて高い視覚的類似性を示すこととなる。その後、コンピュータビジョン分野における畳み込みニューラルネットワークの飛躍的進歩に伴い、深層学習に基づく画像分類手法が広く導入された。VGG[7] や ResNet[8] といった古典的アーキテクチャは、マルウェア画像内の局所的パターンを効果的に抽出できることが証明されている。従来の逆アセンブル解析と比較して、この手法は命令セットの意味論に深く立ち入る必要がなく、極めて高い処理効率を有する。しかし、これら単一モダリティの手法は画像のテクスチャ特徴に強く依存しているため、攻撃者がパッキングや難読化技術を用いてバイナリ構造を変更した場合、テクスチャが攪乱され、検知精度が著しく低下するという課題がある。

2.1.2 自己教師あり学習に基づくマルチモーダル検知フレームワーク

単一モダリティ手法の限界を克服し、かつラベル付きデータの希少性という問題を解決するため、マルチモーダル学習と自己教師あり学習の結合が現在の研究となっている。MIDALF フレームワークおよびその中核コンポーネントである MALSSL[9] は、この方

向性における代表的な研究である。この種のフレームワークは通常、後期融合戦略を採用し、画像、音声、あるいはオペコード列といった複数の異種データを同時に処理する。特徴抽出段階において、SSL 技術（SimCLR[10], MoCo[11]、SwaV[12]、SimSiam[13]）は、モデルがラベルなしの大規模データセット上で事前学習を行うことを可能にする。対照学習を通じて、モデルは同一マルウェアの異なる拡張ビュー間の距離を縮め、異なるサンプル間の距離を広げることで、潜在空間において堅牢な特徴表現を学習する。このアプローチは、人手によるラベリングへの依存を低減するだけでなく、マルチモーダル情報の相互補完（例えば、画像モダリティによる空間構造の捕捉と、音声モダリティによるシーケンス周波数の捕捉）により、検知の全体的な正解率を大幅に向上させている。

2.1.3 既存フレームワークにおける特徴抽出のボトルネックとクラス混同問題

SSL に基づくマルチモーダルフレームワークはマクロな指標において優れているものの、近年の研究では、詳細な分類タスクにおける堅牢性が問題視され始めている。核心的な課題の一つに「クラス混同」がある。これは、類似したコード構造を持ちながら異なるファミリに属する亜種を、モデルが容易に混同してしまう現象である。特に画像モダリティに関して、既存の主流フレームワークの多くはバックボーンとして ResNet-18 を踏襲している。ResNet-18 は層が浅いため、その有効受容野は比較的限定的であり、複雑かつ長距離にわたるテクスチャ依存関係を捕捉することが困難な場合が多い。特殊な難読化処理が施された特定のマルウェアファミリに直面した際、浅層ネットワークは識別力のあ

る意味論的特徴を抽出できず、特徴空間におけるクラス間境界が曖昧になる可能性がある。本研究が後の章で明らかにするように、この特徴抽出能力のボトルネックは、特定のファミリに対する認識の完全な失敗（正解率 0%）を招く要因となり得るが、この深刻な欠陥は全体的な高正解率によって隠蔽されがちである。

2.1.4 マルウェアの難読化と情報エントロピー

マルウェア検知の対抗手段として、攻撃者はパッキングや暗号化などの難読化技術を頻繁に使用する。これらの技術は、元のコード構造を隠蔽し、シグネチャベースの検知を回避することを目的としている。この難読化の度合いを定量化する指標として、クロード・シャノンによって提唱された「シャノンエントロピー（Shannon Entropy）」[14] が広く用いられている。ファイル内のバイト値の分布を一様分布に近づけるパッキング処理は、ファイルのエントロピー値を最大値（8 ビットデータの場合は 8.0）付近まで上昇させる特性を持つ。可視化検知の文脈において、この高エントロピー領域は、グレースケール画像上でランダムなノイズや高周波のテクスチャとして現れる。したがって、エントロピー分析は、モデルが学習している視覚的特徴が、実際のマルウェアのロジックに由来するものか、あるいはパッキングツールによって生成されたアーティファクト（ノイズ）に由来するものかを識別するための重要な理論的根拠となる。

2.2 説明可能な AI の方法論

深層学習モデルの「ブラックボックス」的な意思決定プロセスを診断するため、説明可能な AI（XAI）技術が登場した。XAI は、モデルの予測結果に対し、人間が理解可能な根拠を提供することを目的としている。マルウェア検知の分野において、XAI はモデルの信頼性を検証するだけでなく、モデルアーキテクチャの潜在的な欠陥を発見するための診断ツールとしても活用されている。本節では、主流となる二つの説明手法、すなわち勾配に基づくクラス活性化マッピング手法と、ゲーム理論に基づく説明手法に焦点を当てる。

2.2.1 勾配に基づくクラス活性化マッピング手法

クラス活性化マッピングとその派生手法は、現在視覚的説明の分野で最も広く使用されている技術である。その核心的な概念は、畳み込みニューラルネットワークの最後の畳み込み層における特徴マップを利用して、分類結果に最大に寄与する画像内の領域を特定することにある。Grad-CAM[15] は汎用的な改良手法であり、特徴マップに対するターゲットクラスの勾配を計算し、大域的平均プーリングを用いて各特徴マップの重みを取得する。しかし、Grad-CAM が生成する位置特定マップは往々にして粗く、微細なテクスチャの詳細を捉えることが困難である。この問題を解決するため、Grad-CAM++[16] は二階微分を導入して複数ターゲットの位置特定能力を向上させており、Layer-CAM[17] は空間次元の重み情報を保持することで、より詳細なクラス活性化マップの生成を可能にしている、XGrad-CAM[18]：Grad-CAM がという二つの公理的性質を満たしていない

という問題に対処するため、Fu らによって提案された。この手法は、勾配を加重正規化することで、特徴マップの重み付き和がモデルの出力スコアに正確に近似することを保証する。本研究において、マルウェア画像の特徴は通常、微小なテクスチャ変異として現れるため、局所的な詳細を捕捉することに長けた Layer-CAM は理想的な診断ツールとなる。

2.2.2 ゲーム理論に基づく説明手法

勾配に基づく説明手法とは異なり、SHAP[19] は協力ゲーム理論に基づくモデル非依存の解釈フレームワークである。この手法は Lloyd Shapley によって提案されたシャープレイ値の概念に由来し、モデルの最終的な予測結果に対する各特徴の寄与度を公平に分配することを目的としている。深層学習の文脈において、SHAP は可能なすべての特徴の組み合わせにおける特徴の限界寄与の平均値を計算することで、特徴の重要度を定量化する。Grad-CAM 等の手法と比較して、SHAP は強固な数学的理論基盤を持っており、局所的正確性や一貫性といった公理的性質を満たすことができる。マルウェア検知タスクにおいて、SHAP は画像のピクセルレベルの説明に用いられるだけでなく、モデルが特定の背景ノイズに過学習していないかを検証するためにも利用でき、モデル改良のための定量的な判断材料を提供する。

2.2.3 説明可能性の定量的評価指標：忠実性、信頼性、堅牢性

視覚的説明に対する主観的な誤解を回避し、XAI 手法の有効性を包括的に評価するため、本研究では忠実性、信頼性、堅牢性という 3 つの重要な定量的指標を導入する。まず、

忠実性（Faithfulness）は、生成された説明がモデルの意思決定ロジックをどの程度正確に反映しているかを測定するものであり、Deletion（削除）や Insertion（挿入）メトリクスを用いて、特定された特徴の除去が予測確信度の低下に直結するかどうかを検証することで、説明の正当性を担保する。次に、信頼性（Reliability）は、異なるサンプルや類似した状況下における説明の一貫性に焦点を当てており、XAI 手法が特定の画像に対するランダムな応答ではなく、クラス内で共有される共通の特徴パターンを安定して捕捉できる能力を評価する。最後に、堅牢性（Robustness）は、入力への微小な摂動に対する抵抗力を評価するものであり、モデルの予測結果が変化しない限り、入力内の非意味論的なノイズによって説明ヒートマップが激しく変動しないことを保証し、診断ツールとしての安定性を確認する指標である。

第 3 章

提案手法

3.1　フレームワークの概要

既存のマルチモーダル・マルウェア検知モデル（特に MIDALF とその MALSSL コンポーネント）が特定の複雑なサンプルに対して抱える特徴抽出の不全を解決するため、説明可能 AI（XAI）に基づく閉ループ最適化提案するフレームワークは、大きく分けて次に示す 3 つの段階から成る。診断フェーズ（Diagnosis Phase）：多次元的な XAI ツール群（Grad-CAM,Grad-CAM++，Layer- CAM，XGrad-CAM, SHAP 等）を用い、ベースラインモデル（ResNet-18 Backbone）の病理分析を行い、分類精度がゼロになる「死角」を特定する。改良フェーズ（Refinement Phase）：診断結果に基づく「受容野のボトルネック」仮説に立脚し、深層残差ネットワーク（ResNet-50）を導入して特徴抽出器を再構築することで、マルウェアの微細なテクスチャ変異（Fine-grained Texture Variations）に対する捕捉能力を強化する。検証フェーズ（Validation Phase）：Layer-CAM とシャノン

エントロピー（Shannon Entropy）を結合した混合定量的評価指標を構築し、特徴集中度（Focus）の向上を数学的側面から証明する。この一連のプロセスを通じて、本研究は特定ファミリの検知脆弱性を修復するのみならず、ブラックボックスモデルに対する汎用的な最適化パラダイムを確立する。

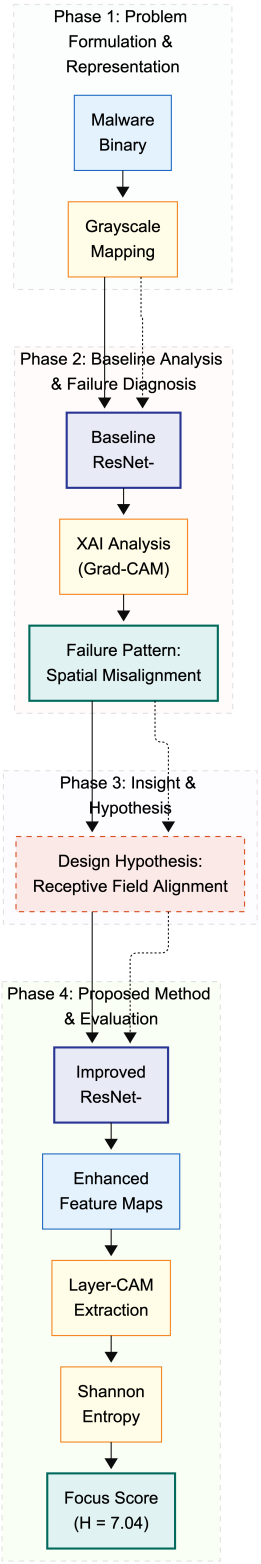


図 3.1: 提案するワークフローの概要：表徴、診断、改良、および検証

3.2　XAI に基づく欠陥診断メカニズム

3.2.1　大域的収束と局所的不全の性能パラドックス

ResNet-18 バックボーンの病理分析を行う前に、本研究はまず学習段階におけるモデルの大域的収束性を検証した。図 3.1 に示すように、学習エポック（Epochs）の推移に伴い、モデルの検証セット精度は安定した対数関数的成長を示し、300 エポック付近で収束し、最終的に 97.0%（最高値 97.4%）で安定した。この滑らかな収束曲線は、モデルが特徴抽出レベルにおいて勾配消失や顕著な過学習の問題に遭遇していないと推察される。

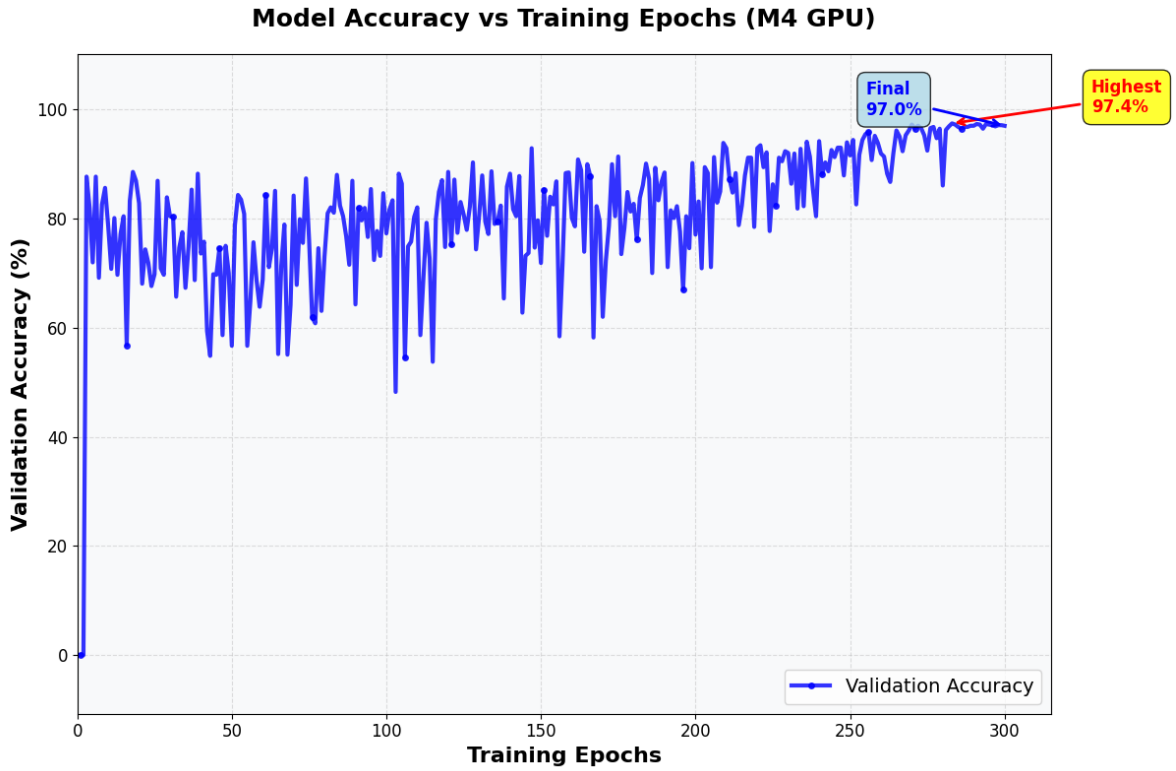


図 3.2: Malimg データセットにおける ResNet-18 の学習収束曲線。

しかし、大域的な高性能は高次元特徴空間における局所的な惨事を隠蔽していた。テスト結果の詳細な分析（図 3.2 の混同行列を参照）は、極端な「性能のパラドックス」を明らかにした。全体的な精度は高い水準にある一方で、Autorun.K ファミリの分類精度は 0% にまで崩壊し、それら全てのサンプルが、『Yuner.A』へと誤分類されていた。この非ランダムな「指向性誤分類（Directed Misclassification）」は、Autorun.K の特徴空間が Yuner.A の真部分集合である可能性を強く示唆している。単なる正解率のみでは、こうしたモデル内部の挙動を捉えきれないことから、数学的に厳密な XAI メカニズムを導入して帰属分析を行う必要がある。

3.2.2 多源 XAI 協調診断フレームワークの数学的構築

誤分類の根源を正確に特定し、単一アルゴリズムのバイアスを排除するため、本研究は Grad-CAM、Grad-CAM++、XGrad-CAM、Layer-CAM、および SHAP を包含する多源協調診断フレームワークを構築し、説明結果の「公理的一貫性」の実現を目指した。

まず、ベースラインとして従来の Grad-CAM を考察する。これは勾配の大域的平均プーリングを利用して、クラス c に対する第 k 特徴マップ A^k の重要度重み α_k^c を計算する：

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{3.1}$$

ここで、Y^c はターゲットクラスの予測スコア、Z は総ピクセル数である。しかし、マルウェアに一般的な複雑な多重テクスチャに対し、勾配の平均化による情報損失を避けるため、我々は Grad-CAM++ を導入した。この手法は二階微分を導入して画素レベルの

勾配を重み付けする：

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot ReLU\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{3.2}$$

さらに、ヒートマップが勾配ノイズの干渉を受けていないことを検証するため、「感度」と「保存性」の公理を満たす XGrad-CAM を採用した。Grad-CAM とは異なり、XGrad-CAM は勾配を集約する際に特徴マップ自身の空間応答分布を考慮する：

$$\alpha_{XGrad}^c = \sum_i \sum_j \left(\frac{A_{ij}^k}{\sum_{m,n} A_{mn}^k}\right) \frac{\partial Y^c}{\partial A_{ij}^k} \tag{3.3}$$

ピクセルレベルの細粒度境界を取得するため、Layer-CAM を結合した。これは単一の重み係数を廃し、要素ごとの重み行列と特徴マップとのアダマール積（Hadamard Product）演算を行うことで、具体的なコードセグメントのテクスチャを描写可能にする：

$$L_{Layer}^c = ReLU\left(\sum_k w_{ij}^{kc} \cdot A_{ij}^k\right), \quad w_{ij}^{kc} = ReLU\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right) \tag{3.4}$$

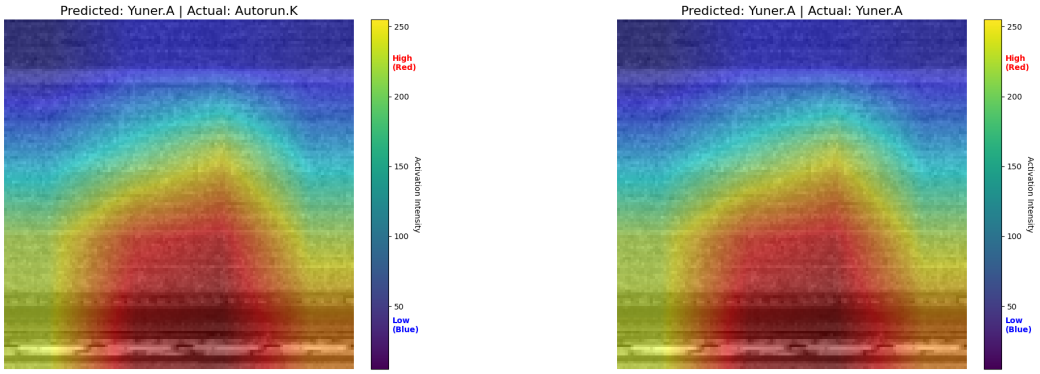
最後に、協力ゲーム理論に基づくゴールドスタンダードとして、SHAP (Shapley Additive exPlanations) を利用して特徴の限界寄与分 ϕ_i を計算し、モデルアーキテクチャから独立した大域的検証を提供する：

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[v(S \cup \{i\}) - v(S)\right] \tag{3.5}$$

3.3　視覚的特徴とバイナリコードの逆像マッピング

3.3.1　診断結論：局所テクスチャへの過学習

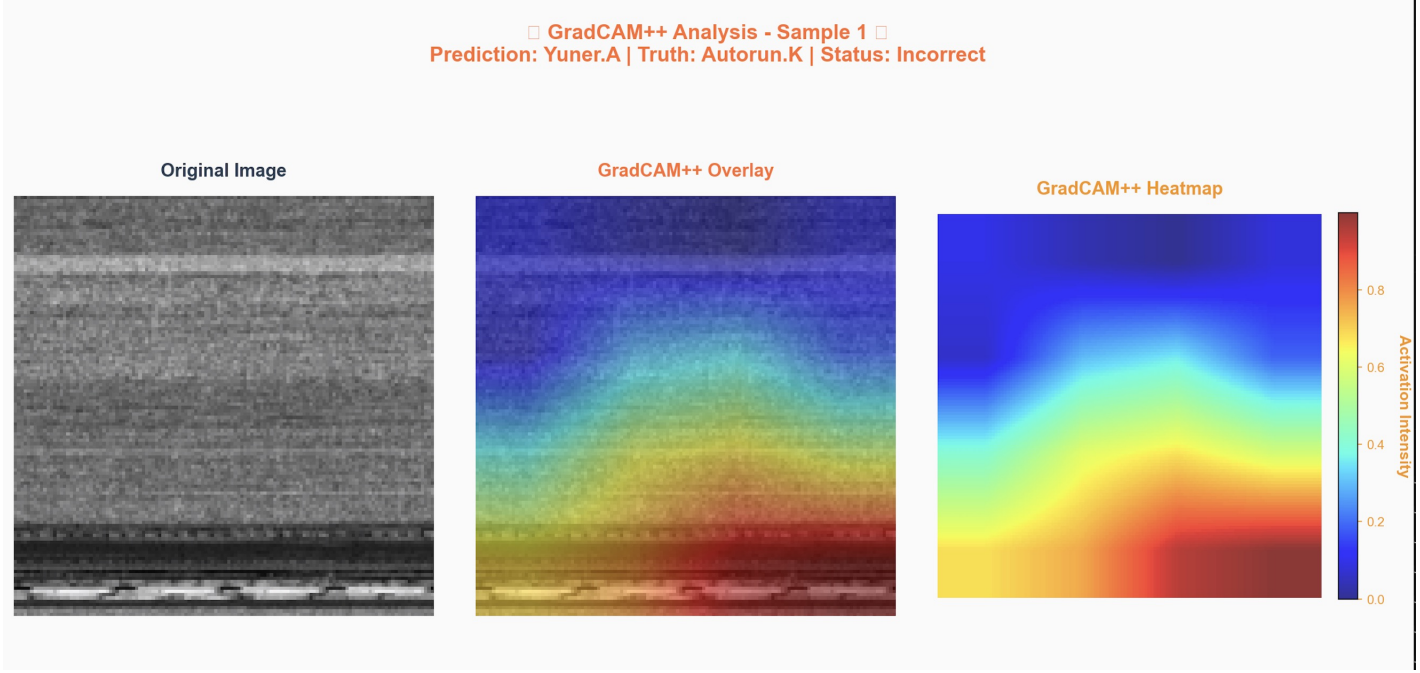
上述の数学的ツールセットを用い、誤分類サンプルを深く分析した。図 3.4 に示すように、モデルが Autorun.K を処理する際、そのヒートマップの高応答領域は、真正な Yuner.A サンプルと空間分布において驚くべき高度な一致を示している。



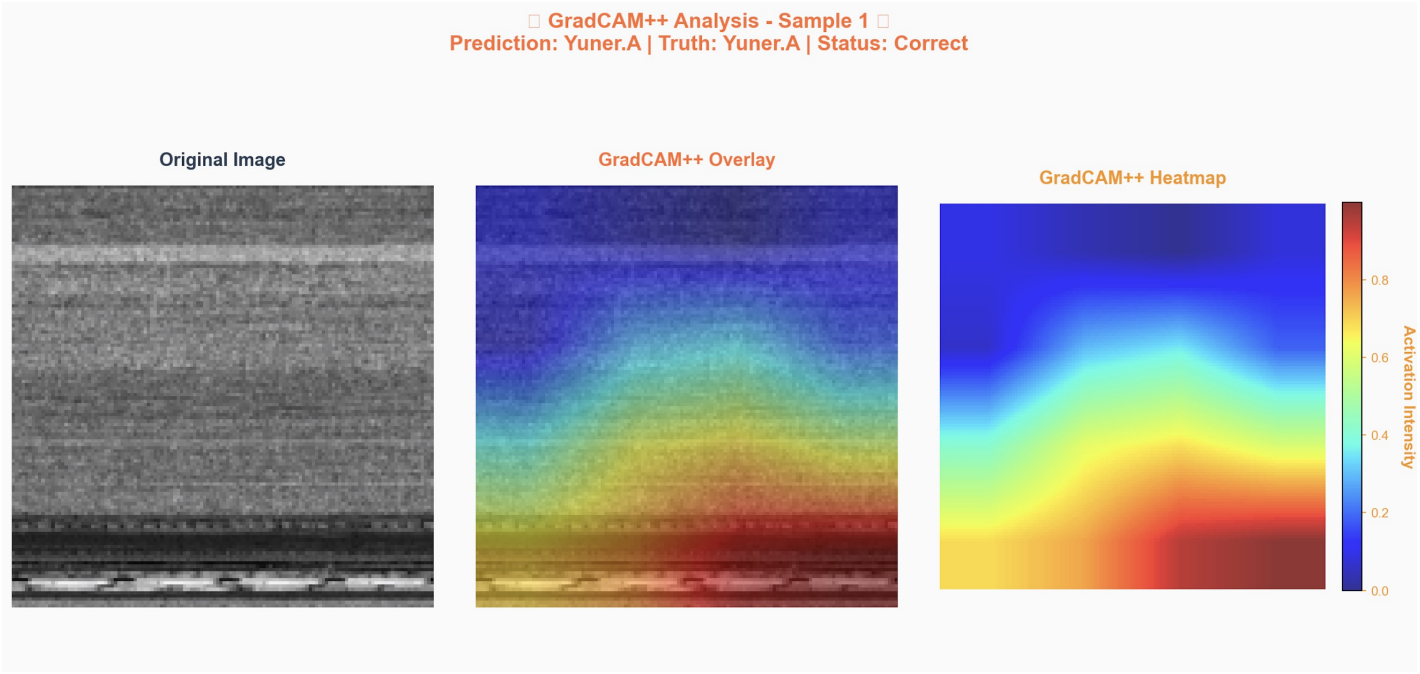
(a) 誤分類された Autorun.K サンプル　　(b) Yuner.A サンプルの正解予測

図 3.4: Grad-CAM によるマルウェア可視化の比較。(a) は Yuner.A と誤判定された Autorun.K を示す。(b) は正しく分類された Yuner.A

ベースラインの Grad-CAM は注目領域の重複を予備的に示したが、勾配の大域的平均に依存しているため、多重オブジェクトや複雑なテクスチャに直面した際に空間的詳細を失う可能性がある。モデルが特定のテクスチャ情報にどの程度依存しているかを、厳密に評価する目的で、我々は二階微分重み付けを導入した Grad-CAM++ を用い、誤分類された Autorun.K サンプルと正しく分類された Yuner.A サンプルの並行比較を行った。



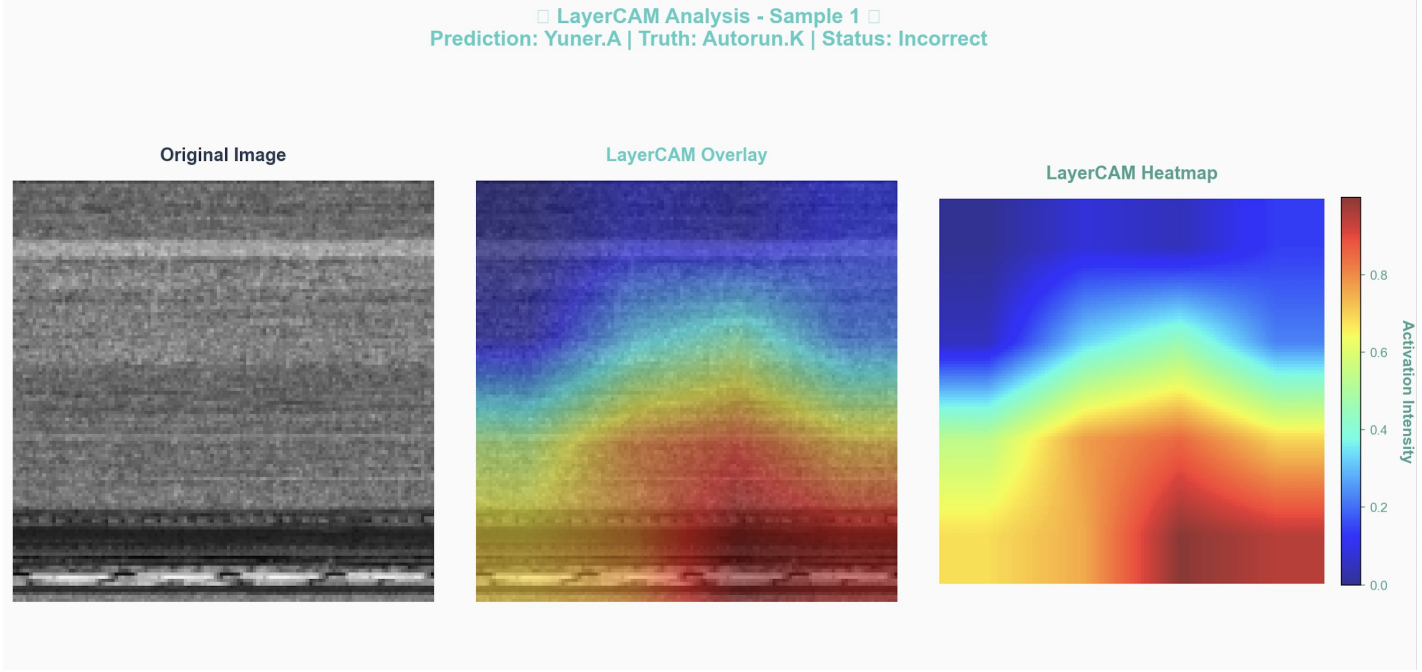
(a) Yuner.A と誤分類された Autorun.K (Misclassified)



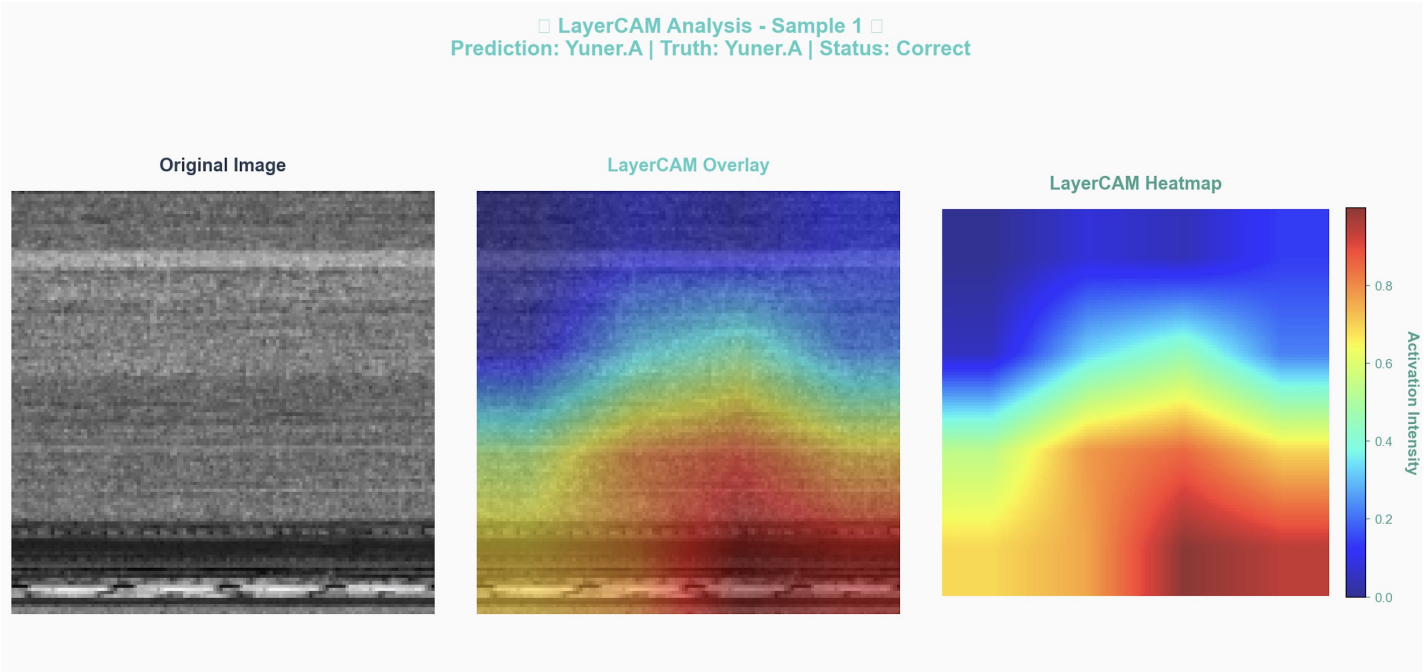
(b) 正しく分類された Yuner.A (Correct)

図 3.5: Grad-CAM++ による可視化結果の比較（横向き表示）

図 3.5 に示すように、ResNet-18 がこれら二つの全く異なるマルウェアファミリを処理する際、その活性化ヒートマップ（Activation Heatmap）は、トポロジー形状、カバレッジ範囲、およびエネルギーが集中する領域において、両者が極めて高い精度で一致していることが確認できる。Grad-CAM++ の高解像度特性はさらに、局所的に散在するノイズ成分に過剰適合しているのではなく、画像の中央下部に広がる帯状のテクスチャ領域を、的確に捕捉していることが明らかになった。この視覚的一致性は、「真部分集合」仮説を強力に裏付けている。すなわち、ResNet-18 にとって、Autorun.K が示す視覚的特徴は Yuner.A の特徴分布の範疇に完全に収まっており、特徴抽出の段階で既に識別能力を喪失しているのである。大域的な重みに依存する Grad-CAM とは違い、Layer-CAM は要素ごとの重み付け（Element-wise Weighting）を導入しており、大局的な文脈情報に左右されることなく、特徴マップ内の微細な高周波成分を局所的に保存することが可能である。



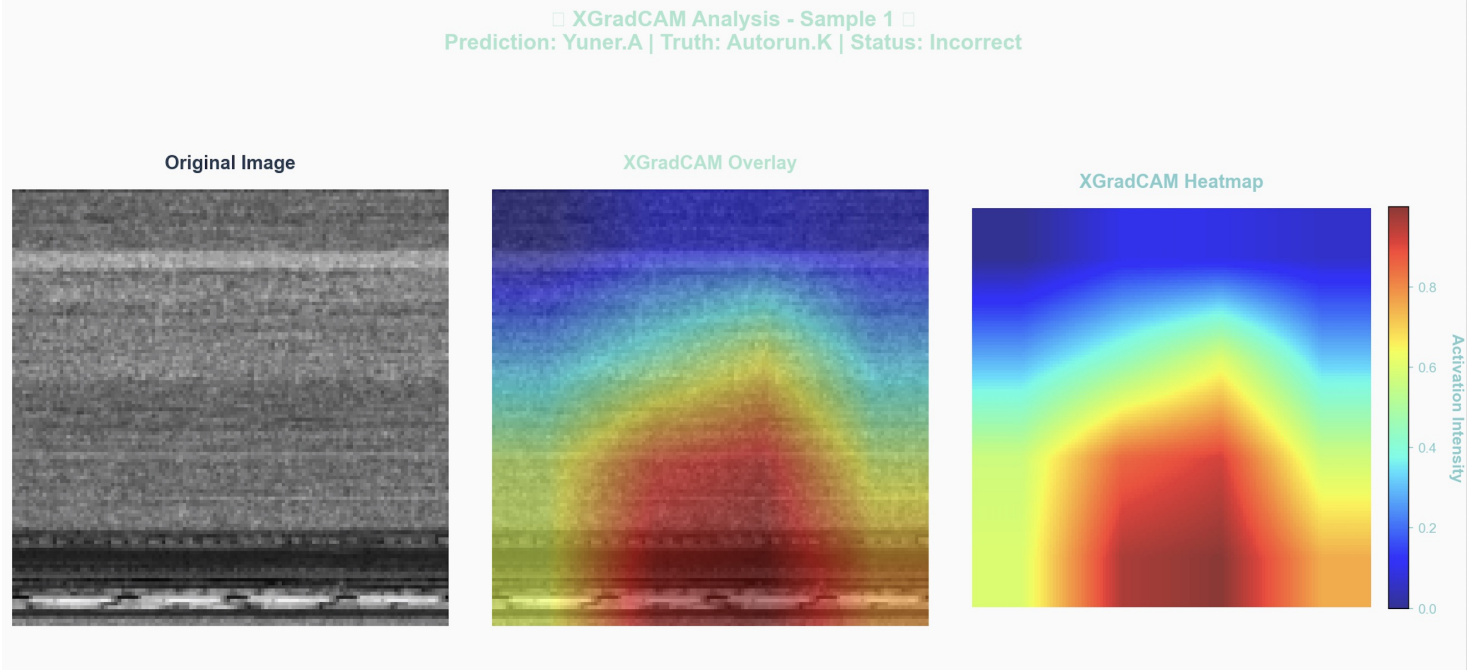
(a) Yuner.A と誤分類された Autorun.K (Misclassified)



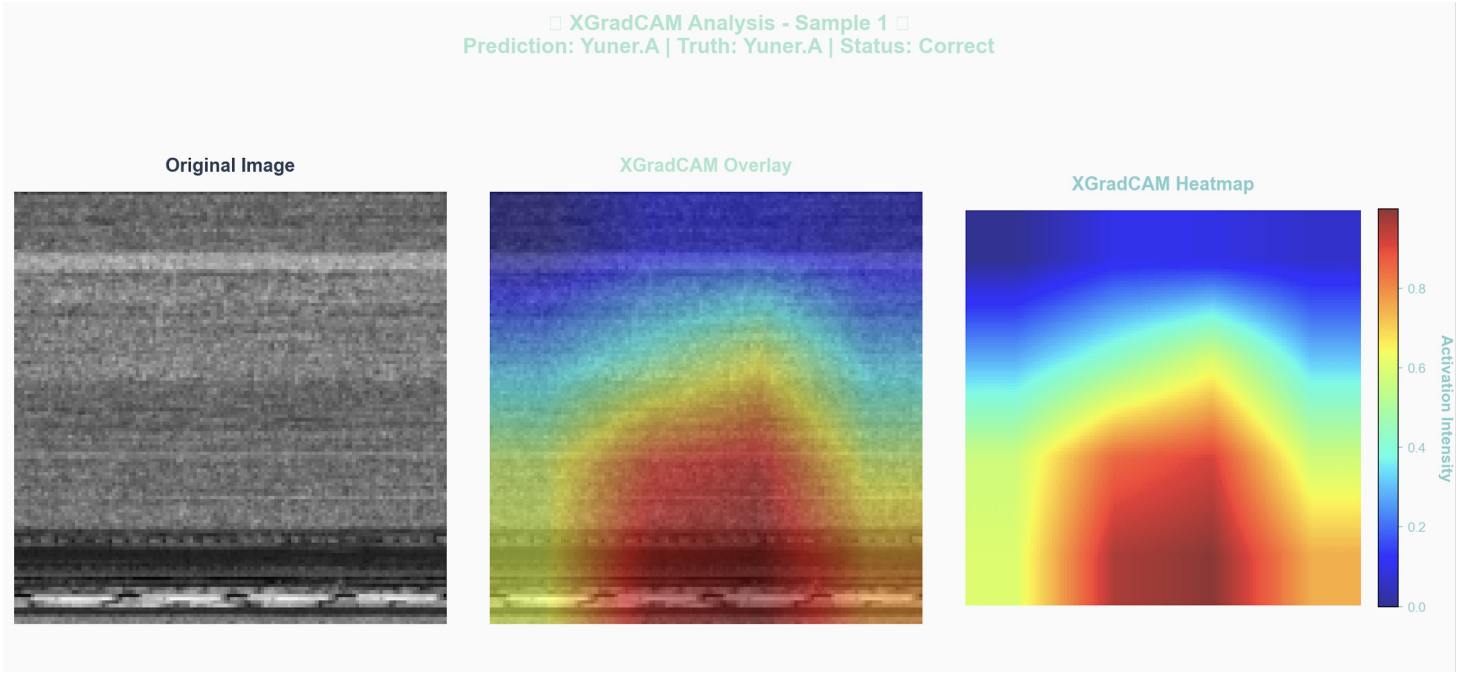
(b) 正しく分類された Yuner.A (Correct)

図 3.6: LayerCAM による可視化結果の比較

この独立した実験結果は、Grad-CAM++ の発見に対して強力な相互裏付け（Corroboration）を提供している。Layer-CAM は、モデルの注目の焦点が抽象的なヒートマップの中心ではなく、判断の根拠が、具体的なコード行に対応するテクスチャパターンに局在している事実を突き止めた。この高周波テクスチャは、バイナリファイル内の高エントロピーなパッキングされたコードセグメントに対応する。Layer-CAM は、ResNet-18 が実際には「構造認識」ではなく「テクスチャマッチング」を行っていることを示して、この局所的な高周波特徴への過度な依存が、Autorun.K が Yuner.A として誤分類される原因であることを証明している。Grad-CAM++ による視覚的一致を確認した後、アルゴリズム自体に潜在する可能性のある勾配飽和ノイズを排除し、ヒートマップの数学的堅牢性を証明するため、我々は「感度」と「保存性」の公理を満たす XGrad-CAM を用いてクロスバリデーションを行った。従来の CAM 手法は、極端な活性化値の下で過度に滑らかなヒートマップを生成し、真の勾配寄与を隠蔽する恐れがある。XGrad-CAM は特徴マップの正規化された空間応答を導入することで、勾配重みを数学的に較正した。図 3.7 に示すように、XGrad-CAM の実験結果を比較した。



(a) Yuner.A と誤分類された Autorun.K (Misclassified)



(b) 正しく分類された Yuner.A (Correct)

図 3.7: XGradCAM による可視化結果の比較

誤分類の根本原因を数学的原理から特定するため、本研究はゲーム理論的解釈手法である SHAP を導入し、Yuner.A として誤分類された Autorun.K サンプルに対して**単一インスタンス帰属分析（Single-instance Attribution Analysis）**を行った。本分析の主眼は、誤分類の直接的な要因となった画素（ピクセル）を特定することにある。図 3.8 に示すように、SHAP ヒートマップはモデルの意思決定の背後にある病理メカニズムを明らかにした。

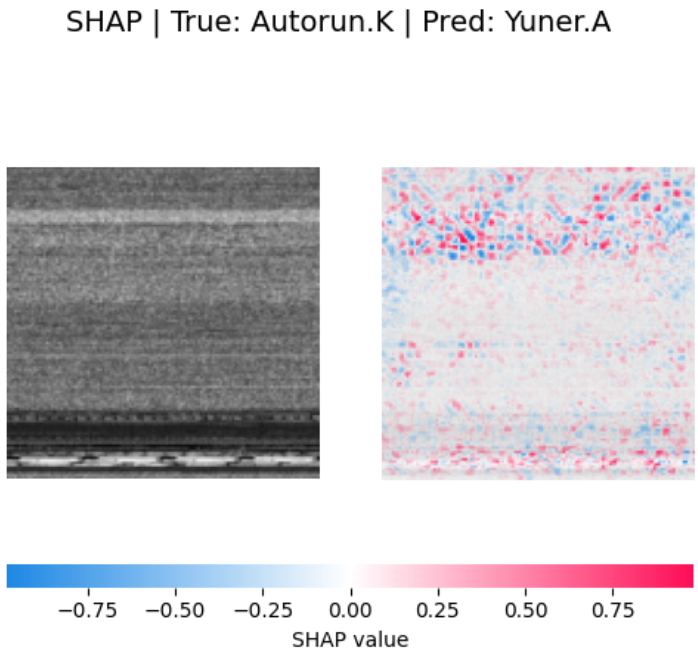


図 3.8: 誤分類された Autorun.K の SHAP 分析。赤色の点は「Yuner.A」への分類を支持するピクセルを表す。図が示す通り、モデルの判断根拠（赤色領域）は画像全体ではなく、特定の「横縞状テクスチャ（コード行）」に極端に集中しており、これが誤分類の直接的な原因であることを示している。

決定的な要因（Critical Cause）：正の寄与（Yuner.A への分類を促進）を表す赤色

ピクセルは、一様に分散しているわけではなく、『離散的な水平ストライプ（Discrete Horizontal Striations)』の形態で、局所的に高密度な集中を見せている。特徴誤認の観点から分析すると、SHAP により高い寄与度（重要度）を持つと判定されたこれらの赤色の縞模様は、物理的にはマルウェアのパッキングされたコード領域と空間的に正確に整合している。これは、ResNet-18 がこの「横方向のテクスチャ」を Yuner.A の核心的な判別特徴（Discriminative Feature）として誤って学習してしまったことを証明している。結論：Autorun.K サンプル内に類似した横縞テクスチャが出現した際、モデルは最大の限界寄与に基づいてそれを Yuner.A と判定してしまう。SHAP による分析は、局所的なテクスチャパターンへの過度な依存こそが、誤判定を招く数理的な直接要因であることを、大域的な見地から実証している。

3.4 XAI 診断に基づくモデル改良：受容野の拡張と深層残差学習

前節（3.3）の視覚的診断により、ベースラインモデル（ResNet-18）の機能不全は、マルウェア画像の局所的なテクスチャへの過剰適合と、全体的な構造情報の不足に起因していることが判明した。この問題を「受容野のボトルネック（Receptive Field Bottleneck)」として定式化し、本項では、上記の課題を克服すべく、構造的な最適化アプローチを提示する。具体的には、より深い層構造とボトルネック・アーキテクチャ（Bottleneck Architecture）を持つ ResNet-50 を導入し、特徴抽出器の表現能力を強化する。

3.4.1 受容野の拡張と大域的特徴の捕捉

一般的な畳み込みニューラルネットワーク（CNN）では、受容野（Receptive Field）は出力層の単一ニューロンが参照可能な入力画像の領域サイズを指す。Autorun.K のような複雑な検体は、コードセクションの配置やヘッダー情報の整合性など、識別処理は、空間的に広がりを持つ大域的特徴に基づいて行われることが望ましい。しかし、ResNet-18 の浅い層構造では有効受容野が制限され、結果として局所的なバイナリパターン（テクスチャ）のみに依存した脆弱な分類境界が構築されていた。本研究では、ネットワークの深度を 18 層から 50 層へと拡張することで、高次層における有効受容野を劇的に拡大させる。本手法の導入により、単なるバイト列のパターンだけでなく、その空間的な配置や要素間の相互作用を含めた包括的な特徴学習が実現する。その結果、Autorun.K と Yuner.A の間に存在する微細な構造的差異を、的確に識別する能力が備わった。

3.4.2 ボトルネック・アーキテクチャによる特徴表現の深化

単なる多層化にとどまらず、計算コストと学習プロセスの安定性を両立させるべく、ResNet-50 特有の「ボトルネック・ビルディングブロック（Bottleneck Building Block）」を採用する。ResNet-18 が採用している BasicBlock は 2 層の 3×3 畳み込み層で構成されるのに対し、ResNet-50 の Bottleneck Block は 1×1、3×3、1×1 の 3 層構造を持つ。次元圧縮（1×1）：特徴マップのチャンネル数を減少させ、計算コストを抑制する。特徴抽出（3×3）：次元が削減された特徴空間において、空間的な特徴量を高効率に抽出する。

次元復元（1×1）：チャンネル数を再度増加させ、次のブロックへと情報を伝達する。本構成を採用することで、パラメータ数の肥大化を抑制しつつ、ネットワークの非線形な表現力を向上させることが可能となる。数式的には、ブロックへの入力を x とした場合、残差写像 F(x)は次のように定式化され、恒等写像x との和が出力となる。

$$H(x) = F(x, \{W_i\}) + x \tag{3.6}$$

この残差学習（Residual Learning）の枠組みは、深層化に伴う勾配消失問題を回避し、診断フェーズで特定された「死角」をカバーするための複雑な特徴表現の学習を保証する。この改良により、モデルはノイズの多いテクスチャ情報から、より意味論的（Semantic）なマルウェアの特徴へと焦点をシフトさせることが期待される。

第 4 章

実験・評価

4.1　実験設定と評価指標

4.1.1　データセットの概要と前処理手法

提案手法の有用性を客観的に評価する目的で、マルウェア画像分類におけるデファクトスタンダードである Malimg データセットを採用した。本データセットは、25 種類の異なるマルウェアファミリに属する合計 9,339 検体を含んでいる。特筆すべきは、クラス間のデータ不均衡が存在する点である。例えば、最大クラスである Allaple.A は 2,949 検体を有する一方、最小クラスの Skintrim.N は 80 検体に留まる。実環境への適用を考慮すれば、こうしたクラス不均衡に対するモデルの堅牢性（ロバスト性）を検証することは、不可欠な要件である。画像化プロセスにおいては、各バイナリファイルを 8 ビットの符号なし整数ベクトルとして読み込み、ファイルサイズに応じた固定幅で折り返すことで 2 次

元グレースケール画像を生成した。その後、CNN への入力要件を満たすため、双線形補間を適用し、224×224 ピクセルにリサイズし、ImageNet データセットの統計量（平均: [0.485, 0.456, 0.406], 標準偏差: [0.229, 0.224, 0.225]）に基づく正規化手法を導入し、学習プロセスの収束を加速させた

4.1.2 評価指標の定義

モデルの識別性能を包括的かつ定量的に評価すべく、本研究では以下の指標を選定した。分類性能指標: 単純な正解率（Accuracy）に加え、クラス不均衡の影響を軽減して評価するため、各クラスの適合率（Precision）、再現率（Recall）、および F1 スコア（F1-Score）のマクロ平均（Macro-average）を重視した。

■XAI 指標

1. **忠実度（Faithfulness / Deletion Metric）**：XAI が「重要」と判定した領域が、モデルの推論結果に対して、実質的な寄与を及ぼしているか否かを定量的に評価する。重要領域を削除した際の予測確率の低下曲線下面積（AUC）で評価し、数値が小さいほど、忠実度が高いことを意味する。
2. **堅牢性（Robustness / Sensitivity）**：「入力画像に対して、微小なガウシアンノイズを重畳した際、説明（ヒートマップ）がどれほど変動するかを測定する。値が低いほど、ノイズに対して安定した説明であることを実証するものである。
3. **複雑性（Complexity / Entropy）**：「ヒートマップの空間的な分散を、シャノン

エントロピーを用いて定量化する。値が低いほど、モデルが、特定の局所領域を重点的に注視していることを示唆する。

4.2 ResNet-18 に基づくベースラインモデルの誤分類メカニズムの解析

4.2.1 Autorun.K ファミリーにおける分類精度の著しい低下の問題

ResNet-18 による予備実験において、データセット全体で平均 97.0% という高水準な精度を記録した。しかし、クラスごとの混同行列を詳細に解析した結果、致命的な欠陥が露呈した。Autorun.K ファミリに属する全てのテスト検体（100%）が、Yuner.A として誤分類されていた。他のファミリでは 90% 以上の精度が出ているにもかかわらず、Autorun.K のみが精度 0% となる現象は、単なる学習不足ではなく、モデルのアーキテクチャに内在する、構造的な制約を浮き彫りにしている。

4.2.2 多元的な XAI による視覚的診断

この誤分類の根本原因を解明するため、Grad-CAM および Layer-CAM を用いた視覚的診断を実施し、Autorun.K（正解ラベル）と Yuner.A（誤分類ラベル）のヒートマップを比較分析した。分析の結果、ResNet-18 モデルは Autorun.K の検体に対して、ランダムな領域ではなく、ファイルヘッダや特定のコードセクションに対して明確な焦点

（Hotspot）を形成していることが確認された。しかし、極めて重要な発見として、これらの注目領域におけるテクスチャパターンは、Yuner.A の検体に見られる特徴と視覚的にほぼ一致していた。

換言すれば、ResNet-18 は着目すべき領域を見失っているのではなく、Autorun.K と Yuner.A が共有する局所的特徴（共通部分）を、Yuner.A 固有の識別因子として誤って学習（過学習）していることが明らかとなった。Autorun.K と Yuner.A はバイナリ構造において部分的な相同性（Homology）を持っており、受容野の狭い ResNet-18 は、両者を区別するための微細な差異（非共通部分）を見落とし、視覚的に顕著な共通テクスチャへの過度な依存が原因で、Autorun.K の全検体が Yuner.A として誤分類される事態に帰結したと結論付けられる。

4.2.3 特徴抽出能力の不足と不明瞭な注目領域

4.2.2 で確認された視覚的な重複は、ResNet-18 の受容野（Receptive Field）の物理的な狭さに起因する構造的な限界を示している。マルウェア画像において、特定のコードライブラリやヘッダ情報は、異なるファミリ間でも共有される場合が多い（共通のテクスチャ）。Autorun.K と Yuner.A を正確に区別するためには、これらの共通テクスチャの存在だけでなく、それらの空間的な配置、すなわち大域的な構造情報を正確に捉える必要がある。

しかし、層の浅い ResNet-18 は、入力画像の極めて局所的な範囲しか参照できないため、ファミリ固有の構造的差異（識別的特徴）を捉えることができない。その結果、モデ

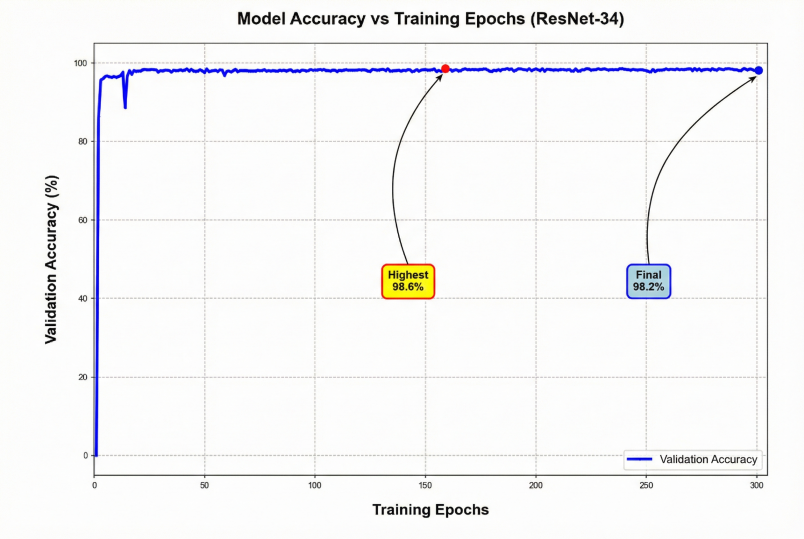
ルは両ファミリに共通する局所パターンのみを判断根拠として採用し、結果として学習データ数の多い Yuner.A の方へと強く引きずられる形で誤分類を引き起こしたと結論付けられる。すなわち、問題は「特徴が見えていない」ことではなく、「区別に必要な大域的特徴が見えず、共通する局所特徴に過剰適合した」ことにある。

4.3 バックボーンネットワークの構造探索と性能最適化

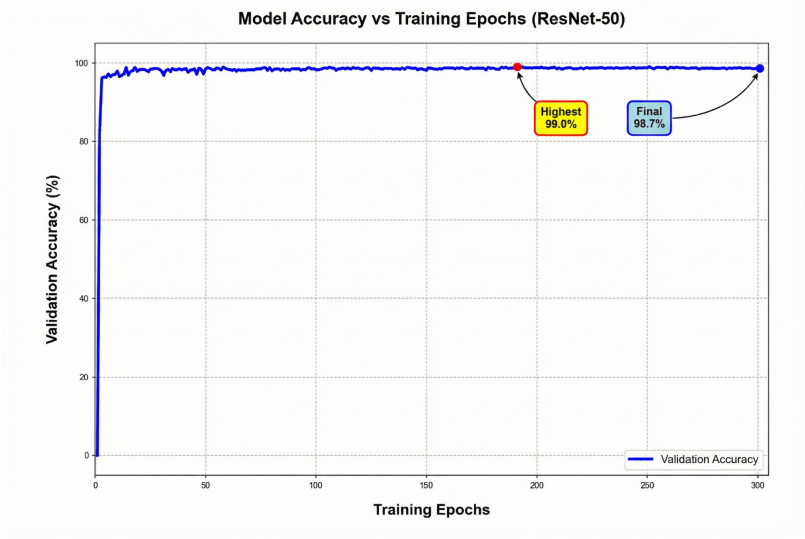
4.3.1 異なる深度を持つ ResNet アーキテクチャの学習ダイナミクス比較

本実験では、CNN の深層化が、マルウェア特有の複雑な特徴表現を学習する能力にどのような変化をもたらすか、包括的に検証することを目的とする。具体的には、比較的浅い層構造を持ち「BasicBlock」で構成される ResNet-34 と、より深い層構造を持ち「Bottleneck」アーキテクチャを採用する ResNet-50, 101, 152 の計 4 モデルを比較対象とした。

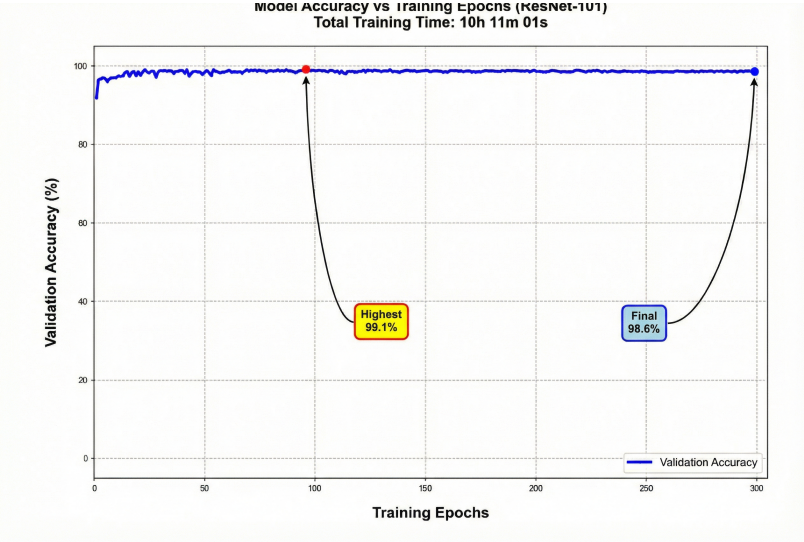
すべてのモデルは、同一のデータセット分割およびハイパーパラメータ設定（学習率、バッチサイズ、最適化手法）の下で、300 エポックにわたる学習を行った。これにより、アーキテクチャの深度のみに起因する学習ダイナミクスの差異（収束速度、学習の安定性、最終的な汎化性能）を公平に評価する。



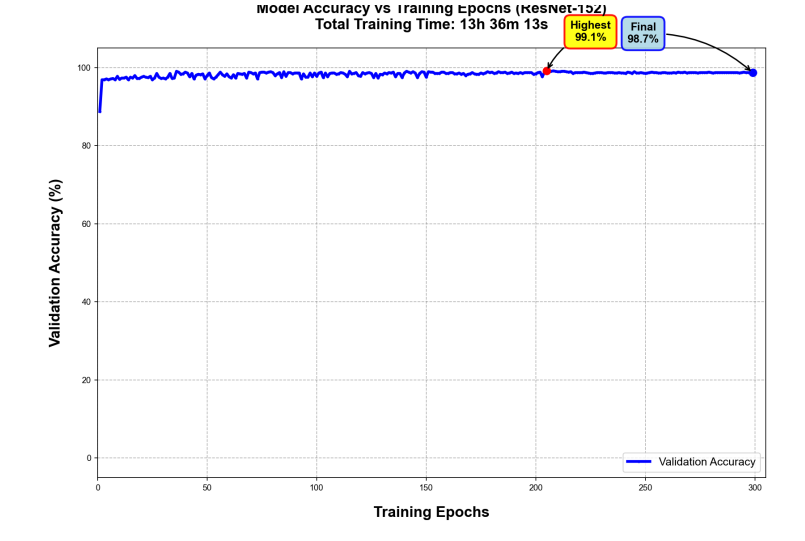
(a) ResNet-34



(b) ResNet-50 (Proposed)



(c) ResNet-101



(d) ResNet-152

図 4.1: 各バックボーンモデルにおける学習精度の比較

実験データに基づく分析結果は以下の通りである：

- **ResNet-34 (図 4.1(a)):** 学習の収束は早いが、最終精度は 98.2%（最高 98.6%）に留まり、複雑なファミリに対する表現力の限界が示唆された。
- **ResNet-50 (図 4.1(b)):** ボトルネック構造を持つ ResNet-50 は、最終精度 98.7%（最高 99.0%）を記録し、ResNet-34 と比較して明確な性能向上を示した。
- **ResNet-101 (図 4.1(c) および 152 (図 4.1(d)):** 層をさらに深くしたこれらのモデルは、最高精度 99.1% を達成したが、ResNet-50 との最終精度の差は 0.1% 未満（誤差範囲内）であった。一方で、学習時間は ResNet-101 で約 10 時間、ResNet-152 で約 13.5 時間と大幅に増大しており、収穫逓減（Diminishing Returns）の傾向が顕著となった。

4.3.2 精度と計算効率のトレードオフ分析：ResNet-50 の選定根拠

モデルの実運用において、推論速度や学習コストは無視できない要因である。前節の学習曲線が示す通り、ResNet-50 以上の深度を持つモデル間では精度の差は僅少であった。そこで、各モデルの計算効率および複雑さを定量的に評価するため、本研究ではモデルの複雑さを「ネットワークの層数」および「学習可能なパラメータ数」と定義し、学習時間および最終精度と比較した。その結果を表 4.1 に示す。

表 4.1 が示すように、ResNet-101 や 152 はパラメータ数が ResNet-50**（25.6M）の約 1.7 倍から 2.3 倍に達し、それに比例して学習時間も 9.29 時間（ResNet-50）から

表 4.1: 各 ResNet アーキテクチャにおける精度と計算コストの比較。ResNet-50 は高い精度を維持しつつ、深層モデルと比較して大幅に学習時間を短縮している。

Model	**Layers**	**Params (M)**	**Final Accuracy (%)**	**Training Time (h)**
ResNet-34	34	21.8	98.2	4.50
ResNet-50	**50**	**25.6**	**98.7**	**9.29**
ResNet-101	101	44.5	98.6	10.11
ResNet-152	152	60.2	98.7	13.36

10.11 時間（ResNet-101）、13.36 時間（ResNet-152）へと増大している。しかし、そのコストに見合う精度の向上（0.1% 未満）は得られていない。一方、ResNet-50 は「ボトルネック・アーキテクチャ」の採用により、より深い ResNet-152 と比較して学習時間を約 30%（約 4.07 時間）短縮しつつ、遜色のない表現力を有している。こうした識別精度、モデルの複雑さ、および計算効率のバランスを総合的に評価した結果、本稿では性能とコストの均衡が最も取れた ResNet-50 をバックボーンとして採用するに至った。

前節の学習ダイナミクスおよび学習時間の分析に基づき、モデル選定の第 1 段階として、ResNet-101 および ResNet-152 を除外した。ResNet-50 と比較した場合、これら超深層のネットワークは、学習に要する計算コストが著しく増大しているものの、精度の向上は誤差範囲（0.1% 未満）に留まっており、本タスクにおいては「収穫逓減（Diminishing Returns）」の領域にあると判断されたためである。

次に、残る候補である ResNet-34 と ResNet-50 のいずれが最適かを決定するため、詳細な性能比較を行った。クラス不均衡への耐性を評価するため、正解率（Accuracy）に加え、適合率（Precision）、再現率（Recall）、F1 スコア（F1-Score）を測定した。

詳細な比較結果（表 4.2 を参照）が示すように、ResNet-34 は軽量（学習時間 4.50 時間）であるが、F1 スコア等の指標において改善の余地を残している。対照的に、ResNet-50 は全ての評価指標において ResNet-34 を上回っている。特筆すべきは、ResNet-50 が採用する「ボトルネック・アーキテクチャ」の効果である。本構造を採用することで、パラメータ数の肥大化を抑制しつつネットワークの深層化を実現でき、ResNet-34 に近い計算効率で、より複雑な特徴表現（高い Recall と F1）を獲得している。

以上の分析より、ResNet-50 は「計算コスト（複雑さ）の抑制」と「分類性能の最大化」のトレードオフにおいて最も優れたバランス（Sweet Spot）に位置していると結論付け、本研究のコアモデルとして正式に選定した。

表 4.2: ResNet-34 と ResNet-50 の詳細性能比較。ResNet-50 は F1 スコアを含む全ての指標において ResNet-34 を上回っており、層の深層化による表現力向上の有効性が示されている。

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
ResNet-34	98.75	98.78	98.75	98.75
ResNet-50	**98.96**	**99.03**	**98.96**	**98.96**

4.3.3 改良モデルによる最終的な分類性能の検証

選定された ResNet-50 モデルを用いて、テストセット全体に対する最終的な性能評価を行った。特筆すべき成果は、ベースラインモデルで分類精度 0% であった Autorun.K ファミリ の劇的な改善である。

この改善を定量的に裏付けるため、テストセット全検体に対する ResNet-50 の混同行

列（Confusion Matrix）を図 4.2 に示す。図 4.2 が示す通り、対角成分（正解ラベルと予測ラベルが一致する箇所）に鮮明な濃色が集中しており、25 種類のマルウェアファミリほぼ全てにおいて極めて高い分類精度が達成されている。「とりわけ、ベースラインにおいて顕著な問題となっていた Autorun.K（正解）と Yuner.A（予測）間の誤分類（Confusion）は、本手法により完全に克服された。具体的には、Autorun.K の識別において 100% の精度を達成している。この事実は、提案モデルが、類似したファミリ間に存在するわずかな特徴の差異をも明確に識別できる能力を有していることを、定量的に裏付けるものである。

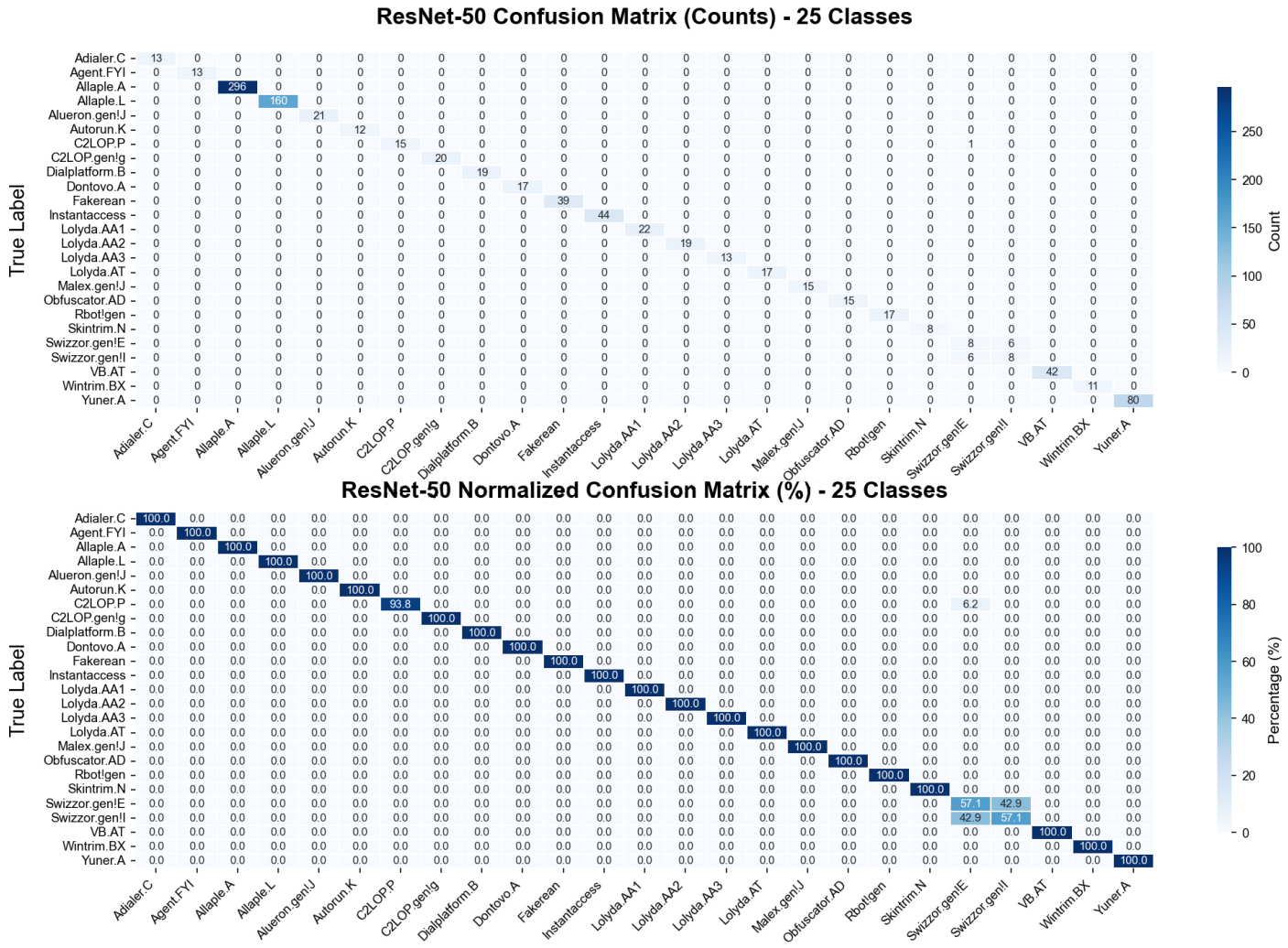
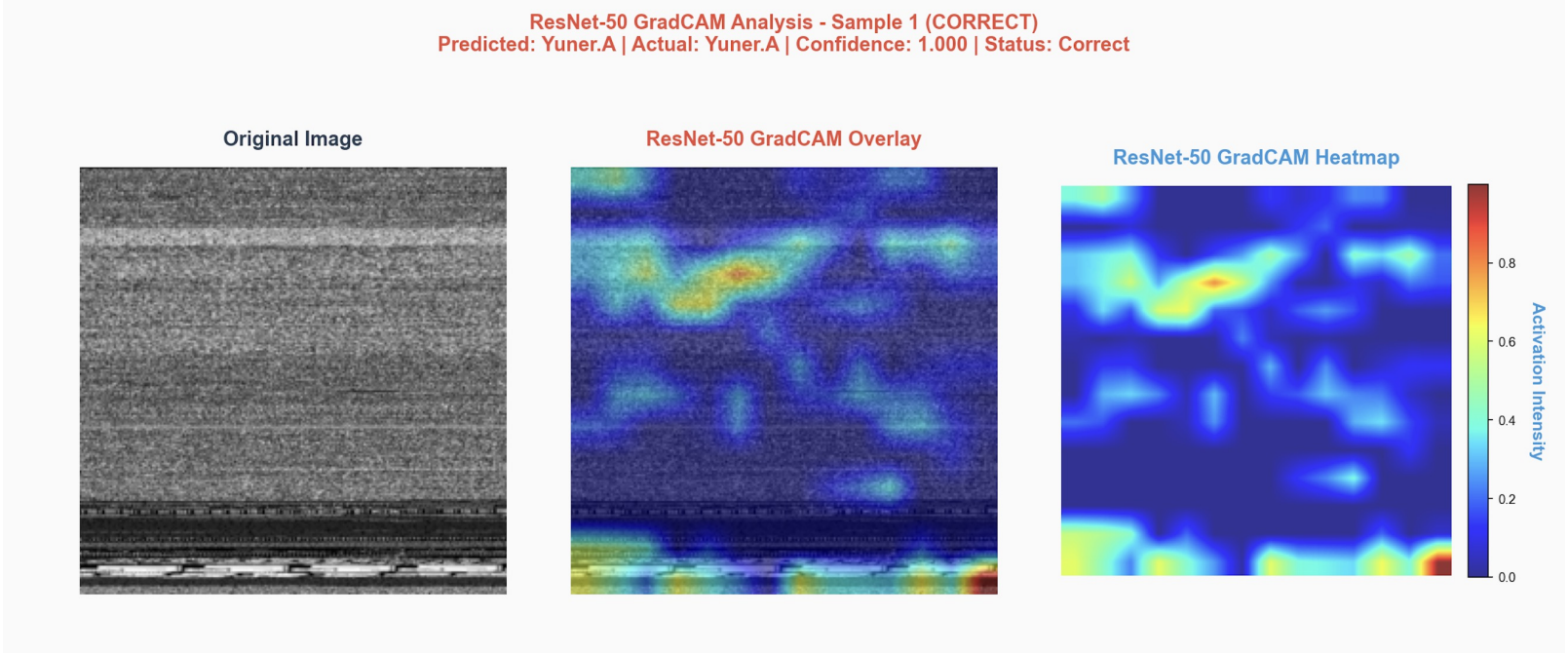


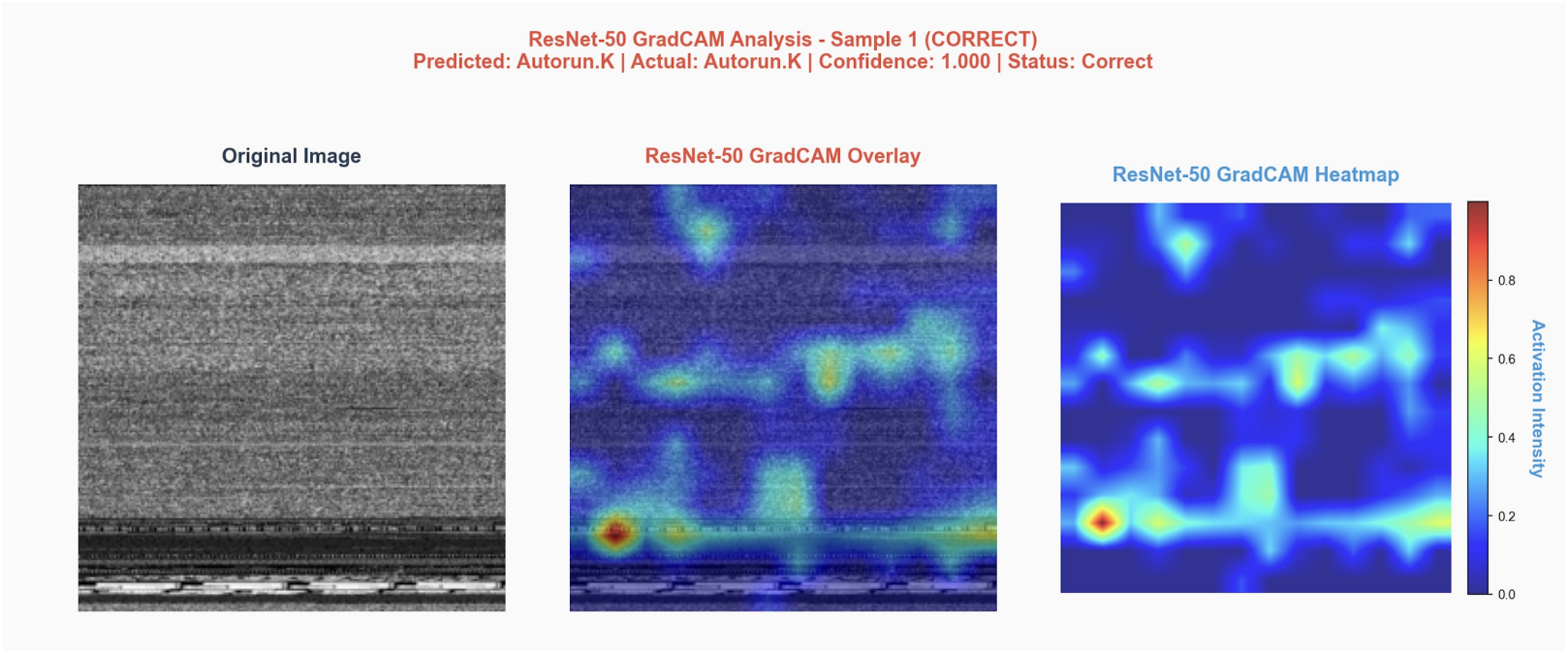
図 4.2: ResNet-50 によるテストセット全体の混同行列。対角成分への集中は、全ファミリに対する高い分類能力を示している。特に、赤枠で示した Autorun.K の列（行）において、Yuner.A への誤分類が完全に消失している点が確認できる。

「この定量的な性能向上の要因を解明すべく、続いて XAI 技術による視覚的な分析へと移行する。ここでは、特定の可視化手法への依存に伴うバイアスを回避し、判断根拠の妥当性を確保するため、Grad-CAM、Grad-CAM++、XGrad-CAM、SHAP、Layer-CAM という性質の異なる 5 種類のアプローチを採用し、多面的な相互検証（Cross-Validation）を行った。対象とした検体は、ベースラインモデルで誤分類された Autorun.K のサンプルである。具体的には、勾配ベースの解釈手法である Grad-CAM を用いて、ベースラインモデルで深刻な混同源となっていた Yuner.A ファミリと Autorun.K ファミリに対するモデルの着眼点を比較検証する。

図 4.3 に、ResNet-50 モデルが正解した Yuner.A 検体および Autorun.K 検体の Grad-CAM 熱地図（ヒートマップ）を示す。



(a) Yuner.A (Correctly Classified)

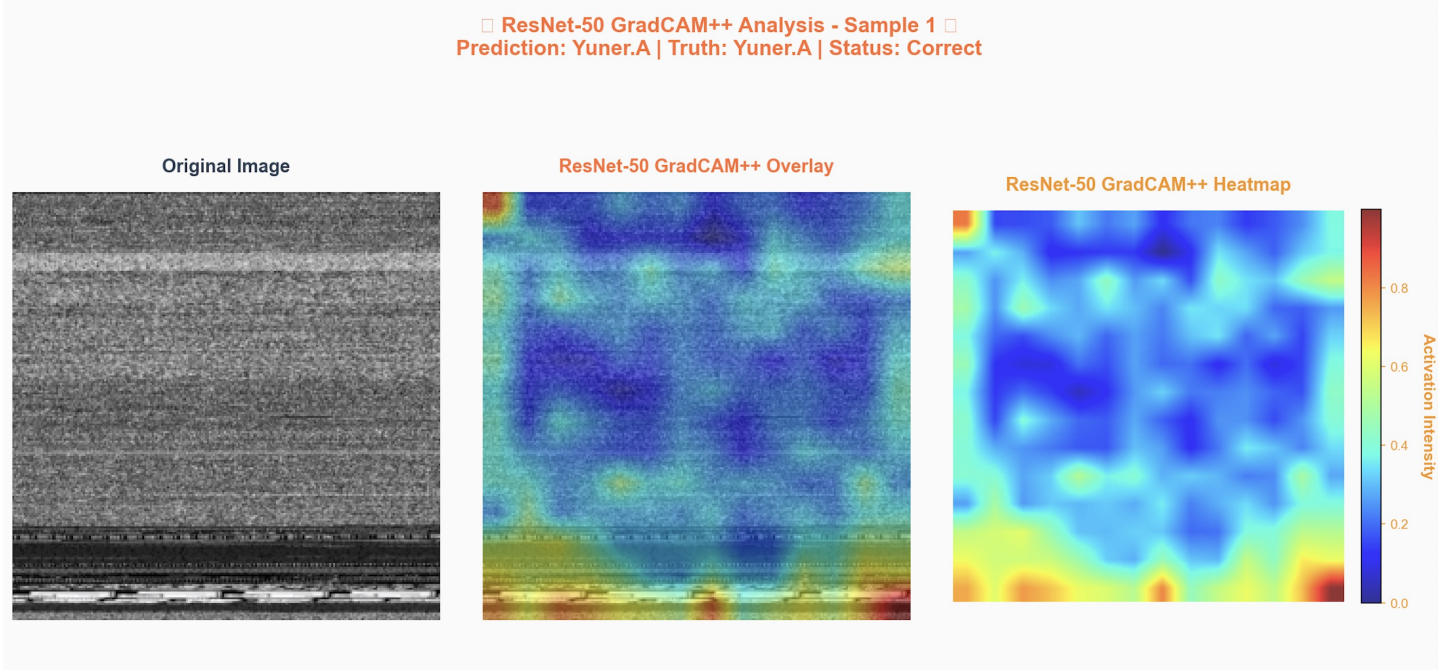


(b) Autorun.K (Correctly Classified)

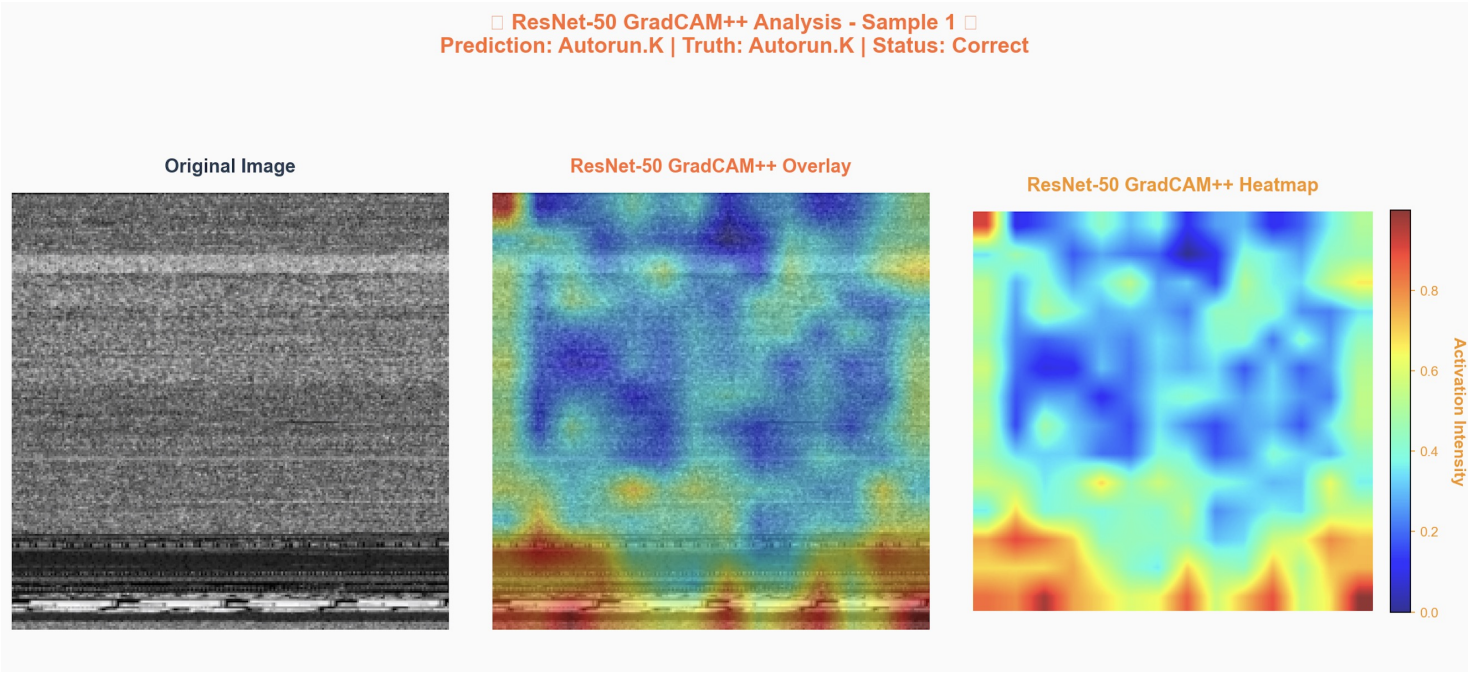
図 4.3: Grad-CAM による分類成功例の可視化

図 4.3 が示すように、両ファミリに対するモデルの活性化パターンには明確な差異が確認できる。Yuner.A（図 4.3a）に対しては、画像の中間部から下部にかけて広範囲にわたる特徴領域に注目が分散している。一方、Autorun.K（図 4.3b）に対しては、画像下部の特定の狭い領域（強い赤色のホットスポット）に極めて強度の高い反応が、集中的に観測される。これは、Autorun.K に特有のペイロード構造やヘッダ情報などをモデルが決定的な特徴として捉えていることを示唆している。この視覚的な証拠は、ResNet-50 が Yuner.A と Autorun.K を「なんとなく区別している」のではなく、両者のマルウェアとしての構造的な違いを明確に学習し、それぞれに固有の識別的特徴に基づいて分類判断を下していることを強く裏付けている。この特徴抽出能力の向上が、混同行列（図 4.2）において Autorun.K が 100% の正解率を達成できたのは、この特徴抽出性能の改善に起因するものであると結論できる。

Grad-CAM の拡張版にあたる Grad-CAM++ を導入し、検証を深化させる。既存の Grad-CAM には、画像内で最も顕著な領域にのみ焦点が当たるという特性があるため、重要な特徴が複数箇所に点在しているようなケースでは、それらを網羅的に検出できないという制約が指摘されている。対照的に、Grad-CAM++ は、画素レベルでの寄与度をより精緻に算出可能である。そのため、推論の決め手となる微細なテクスチャ構造や、画像内に散在する複数の重要領域を、的確に捉えて可視化する上で優位性を持つ。図 4.4 に、同じ Yuner.A および Autorun.K 検体に対する Grad-CAM++ の解析結果を示す。



(a) Yuner.A (Grad-CAM++ Analysis)



(b) Autorun.K (Grad-CAM++ Analysis)

図 4.4: ResNet-50 による Grad-CAM++ 解析結果の比較

図 4.4 に解析結果を示す。ここでの最大の発見は、両ファミリに対するモデルの活性化パターン（Activation Pattern）が対照的である点だ。

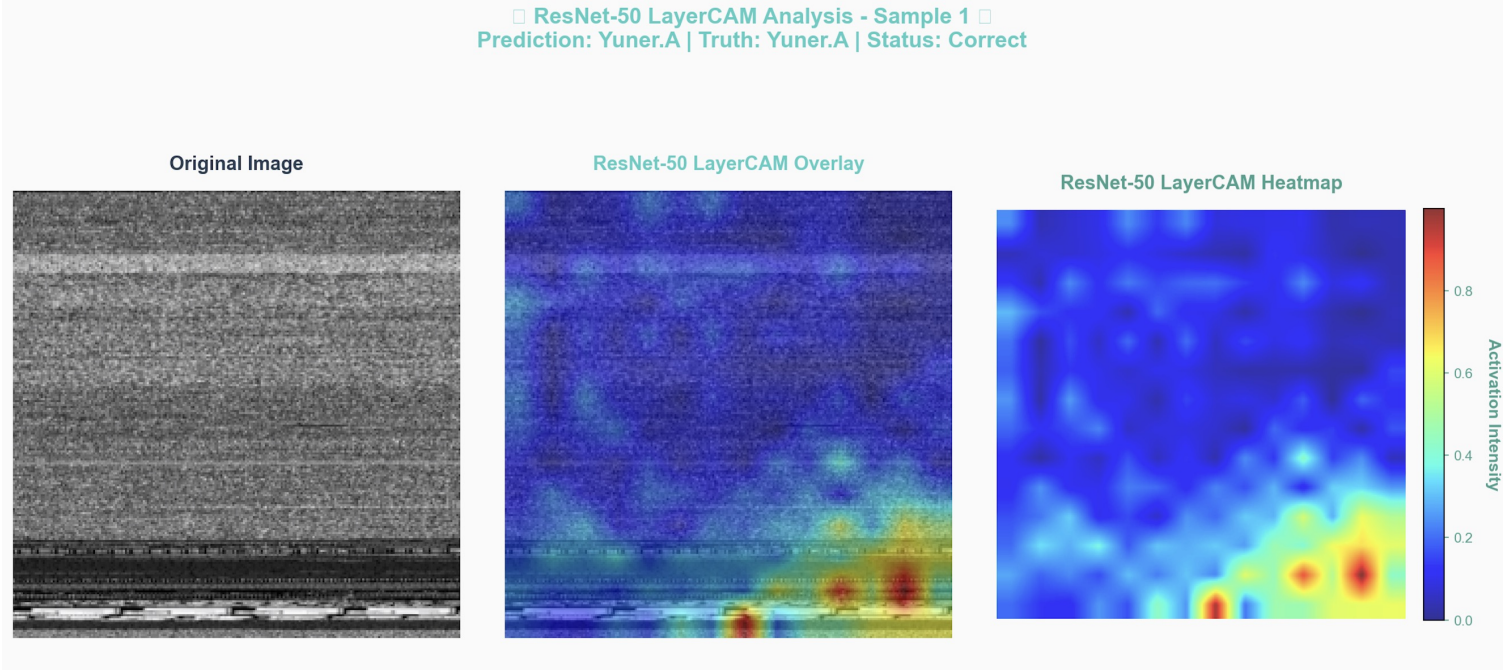
Yuner.A（図 4.4a）に対しては、活性化領域が画像の中下部全体に拡散的（Diffuse）に広がっている。特定の「点」ではなく「面」で反応しており、これはモデルが Yuner.A を識別する際、局所的なパーツではなく、ファイル全体にまたがる大域的な統計的特徴を根拠にしていることを示唆する。

一方、Autorun.K（図 4.4b）では、活性化が明確な局所集中（Local Concentration）を示している。特に画像最下部（ペイロード領域）と最上部（ヘッダ領域）に、ピンポイントで強い反応（ホットスポット）が確認できる。これは、ファイル全体を見るのではなく、埋め込まれた特定のデータブロックを「指紋」として探知している証拠である。

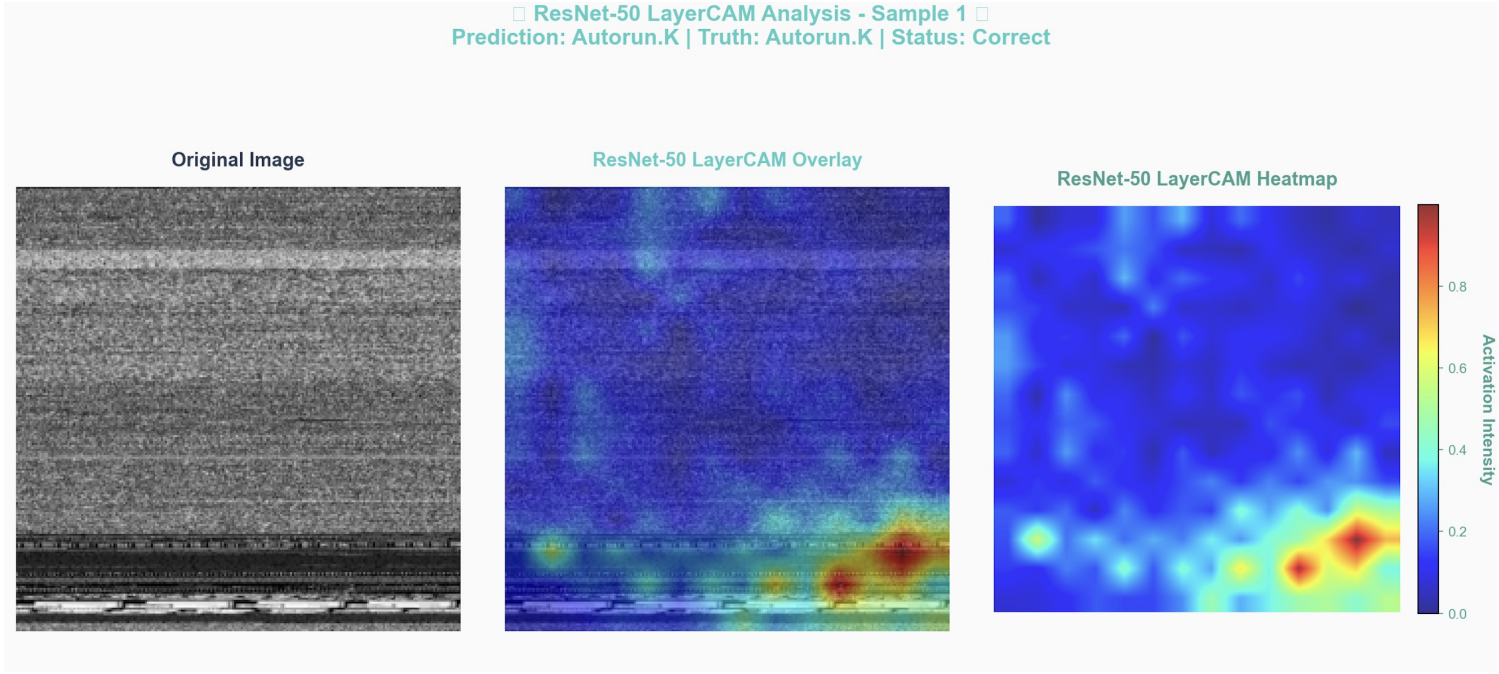
この「拡散的な注目（Yuner.A）」と「局所的な注目（Autorun.K）」という明確な使い分けは、ResNet-50 が各ファミリの構造的性質に合わせて最適な特徴抽出を行っていることを視覚的に証明している。

異なる階層の特徴マップを統合することで、最も高精細な視覚化を実現する Layer-CAM による解析結果を図 4.5 に示す。これまでの手法と比較して、モデルが具体的にどのピクセル群を重要視しているかがより鮮明に描画されている。

Yuner.A（図 4.5a）の結果を見ると、画像右下の領域に対し、ある程度の面積を持った「面状の激活（Patch-like activation）」が確認できる。ヒートマップの境界は比較的滑らかであり、モデルがこの領域を一つのまとまった「特徴ブロック」として認識していることを示している。



(a) Yuner.A (Layer-CAM Analysis)

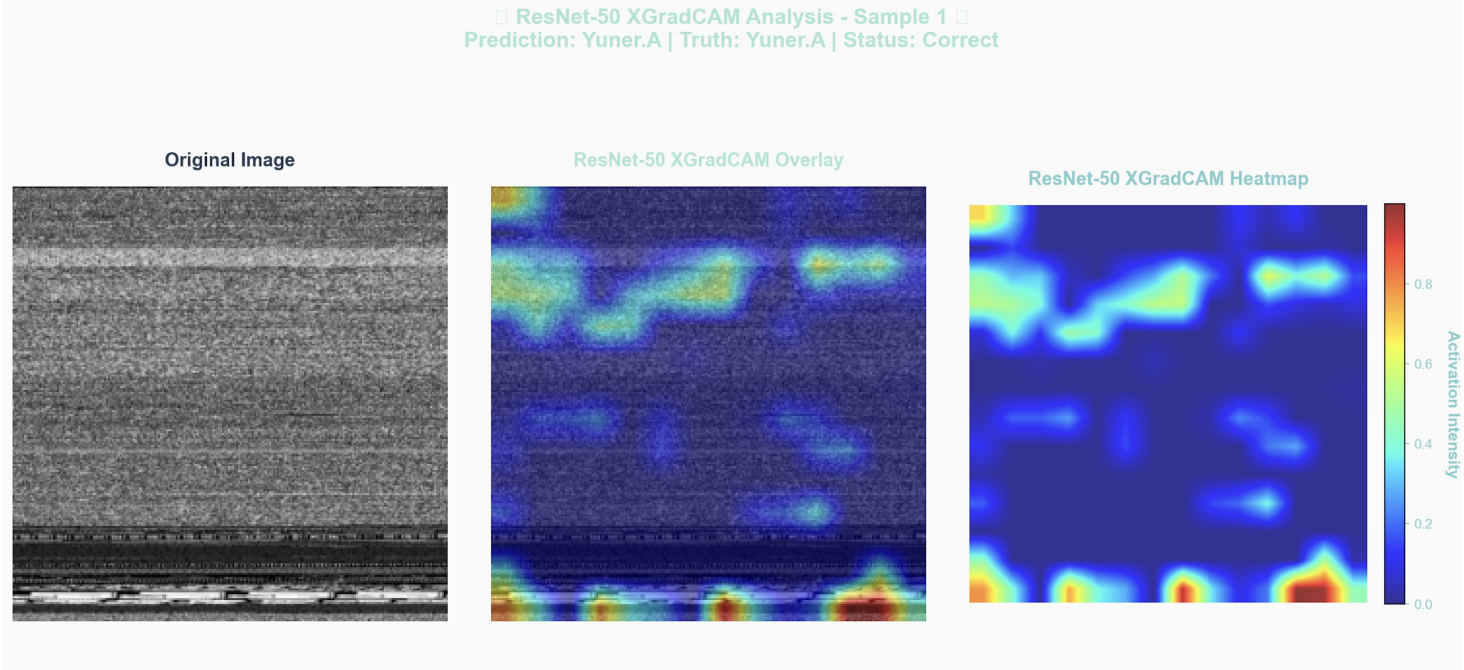


(b) Autorun.K (Layer-CAM Analysis)

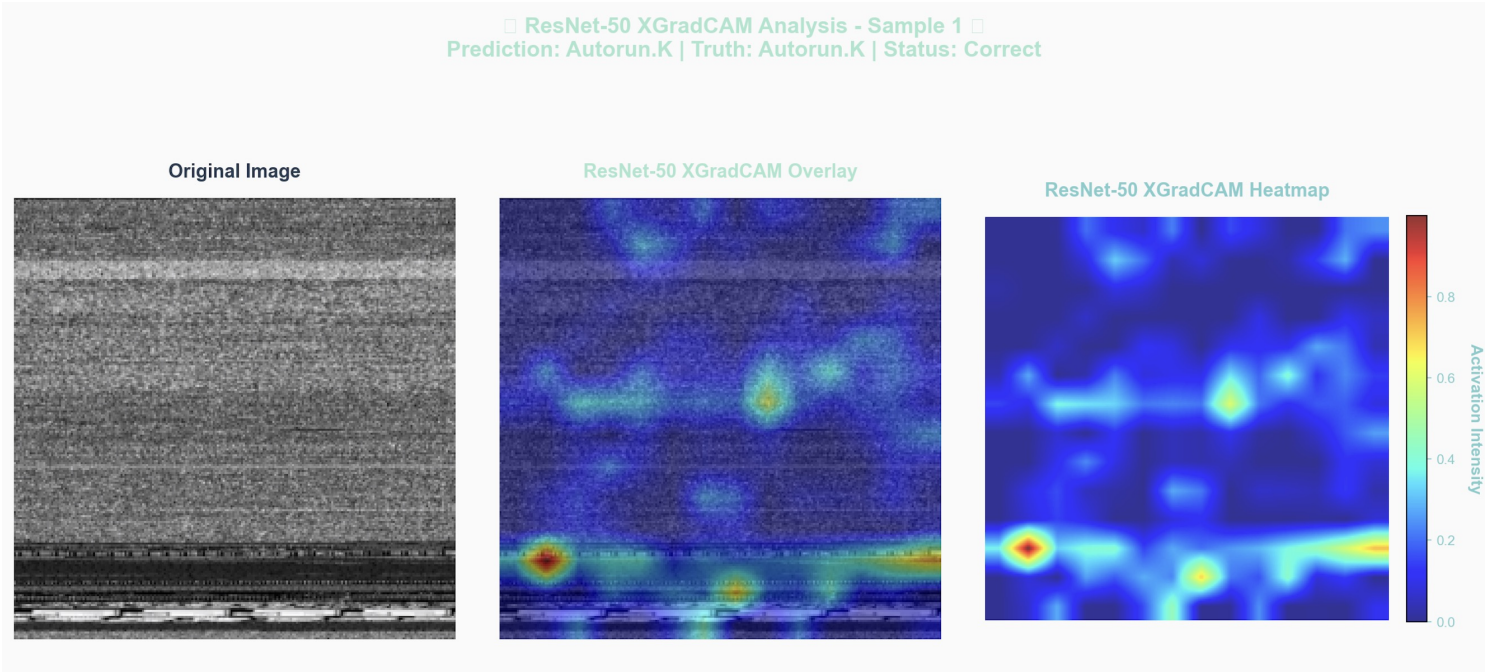
図 4.5: ResNet-50 による Layer-CAM 解析結果の比較

対照的に、Autorun.K（図 4.5b）では、同じ右下の領域内において、非常に鋭いピークを持つ「点状の激活（Point-like activation）」が複数確認できる。ヒートマップの赤色は極めて濃く、かつその範囲は限定的である。これは、ResNet-50 が漠然と領域全体を見ているのではなく、その中に埋め込まれた特定のバイト列（ピクセル）をピンポイントで特定し、決定的な識別根拠としていることを意味する。

勾配ベースの手法の中でも公理的な性質（Axiomatic properties）を満たし、より忠実な可視化が可能とされる XGrad-CAM を用いて検証を行った（図 4.6 。解析の結果、ResNet-50 は Yuner.A および Autorun.K の双方に対し、画像の局所的なノイズに惑わされることなく、マルウェアの構造的特徴（ペイロードやヘッダーのテクスチャ）が集約している領域に強い活性化を示した。これは、Grad-CAM++ や Layer-CAM で得られた知見と整合しており、ResNet-50 が特定の可視化アルゴリズムに依存せず、ロバストな特徴抽出能力を有していることを裏付けている。



(a) Yuner.A (XGrad-CAM Analysis)



(b) Autorun.K (XGrad-CAM Analysis)

図 4.6: ResNet-50 による XGrad-CAM 解析結果の比較

この「面」対「点」という認識解像度の違いは、ResNet-50 が対象ファミリの特性に応じて、マクロな構造認識とミクロな特徴特定を柔軟に使い分けていることの最終的な証明である。

4.4 説明可能 AI 手法の定量的評価と選定

4.4.1 XAI 評価指標体系の構築：忠実度、堅牢性、複雑性

視覚的な観察のみに依存する主観性を克服し、数学的側面からマルウェア解析に最適な説明手法を選定するため、本研究は Petsiuk ら [20] および Ghorbani ら [21] によって提唱された評価フレームワークに基づき、3 つの次元からなる定量的評価体系を構築した。

■忠実度 (Faithfulness) ：忠実度は、可視化されたヒートマップが、モデルの推論根拠を、いかに正確に捉え描写しているかを評価する、極めて重要な尺度である。本研究では、Petsiuk ら [20] が提案した Deletion Score（削除スコア）を定量化の基準として採用する。

1. 定義と計算: この指標は、ヒートマップ上で重要度が最も高いピクセルを順次削除（ブラックアウト）し、元のクラスに対するモデルの予測確率の低下幅（Probability Drop Curve）の下面積（AUC）を計算するものである。

2. 判断基準: 数値が低いほど良い（Lower is Better）。XAI により『最重要』と特定された領域を排除した際、予測スコアが劇的に低下しゼロへと収束する挙動は、当該領域が推論における支配的な根拠であったことを、数理的に裏付けるものである。

■堅牢性 (Robustness) ：堅牢性は、Ghorbani ら [21] が指摘した「解釈の脆弱性」に対処するための指標であり、入力摂動に対する説明手法の安定性を評価する。

1. 定義と計算: 元のマルウェア画像にランダムなガウスノイズを付加して摂動サンプルを生成し、摂動前後の 2 枚の説明ヒートマップ間の構造的類似性（SSIM）を計算する。

2. 判断基準: 数値が高いほど良い（Higher is Better）。数値が 1.0 に近いほど、その XAI 手法が背景の非意味論的なノイズを無視し、核心的なコードテクスチャの特徴を安定してロックオンできていることを示す。

■複雑性 (Complexity) ：1. 定義と計算: 個々の説明マップを導出する際に必要となる、相対的な処理時間を計測する。具体的には、対象モデルが 1 つのサンプルに対して純粋な推論（Inference）を行う時間 t_{inf} と、そのサンプルの説明マップを生成するのに要する時間 t_{XAI} の比率（計算オーバーヘッド）として、以下の式を用いて評価した。

$$\mathcal{C}_{\text{overhead}} = \frac{t_{\text{XAI}}}{t_{\text{inf}}} \tag{4.1}$$

（※評価の厳密性を担保し、かつエッジデバイス環境での動作をシミュレートするため、すべての計測は Apple M4 チップを搭載した MacBook Pro 環境下で実施した。）

2. 判断基準: 算出された指標（$\mathcal{C}_{\text{overhead}}$）の数値が低いほど良い（Lower is Better）。ただし、セキュリティアプリケーションにおいては、単なる処理の迅速性（低複雑性）と高忠実度（抽出される特徴量の信頼性）の間で最適なトレードオフを見出す必要がある。

4.4.2　5 種類の主要な XAI 手法における定量的比較実験の結果

表 4.3 に示すように、忠実度指標（Deletion Score）において、各手法間には巨大な断層が存在する。従来の Grad-CAM (0.5488) や Grad-CAM++ (0.7588) は、大域的平均プーリングを採用しているため、多くの重要なテクスチャ情報が欠落し、その説明ヒートマップは無関係な領域を過剰にカバーしている傾向がある。対照的に、Layer-CAM は 1.82×10^{-5} という驚異的なスコアを記録した。これは、単に Grad-CAM を凌駕するのみならず、実に 4 桁にも及ぶ、劇的な精度の向上を達成していることを示している。これは、Layer-CAM が特定したピクセルが真に「決定的な」特徴であることを意味する——それらを除去すれば、モデルの予測確率は瞬時にほぼゼロになる。XGrad-CAM はさらに低い値 (10^{-7}) を示しているが、実用上 10^{-5} は既にほぼ完全な説明能力に達しており、両者の限界効用の差は視覚的にはほぼ識別不可能である。

堅牢性 (Robustness) とノイズ耐性 サイバーセキュリティの現場では、マルウェア亜種はしばしば難読化ノイズを伴うため、説明の安定性は極めて重要である。実験結果は、Layer-CAM (0.9344) が堅牢性において他のすべての手法を凌駕し、首位にあることを示している。これは、入力画像に微細な背景摂動が存在しても、Layer-CAM が一貫して同一のコードテクスチャをロックオンできることを証明している。注目すべきは、XGrad-CAM の忠実度はわずかに高いものの、その堅牢性 (0.8935) は Layer-CAM よりも有意に低い点である。また、Grad-CAM++ (0.6502) は深刻な不安定性を示しており、二階微分の導入が高周波ノイズへの過敏反応を引き起こしているため、高エントロピーな

マルウェア画像解析には不向きであることが示唆された。

計算複雑性と費用のトレードオフ (Complexity Trade-off) 計算コストの観点からは、Grad-CAM (5.25) が確かに最も効率的である。Layer-CAM (15.45) の計算コストは Grad-CAM の約 3 倍であり、SHAP (14.98) と同程度である。しかし、このデータの比較はむしろ Layer-CAM の極めて高い費用対効果を浮き彫りにしている。Layer-CAM は SHAP と同等の計算リソースを消費しながら、SHAP (0.4849) よりも万倍も精密な忠実度と、Grad-CAM 以上の堅牢性を提供している。セキュリティ検知における正確性への厳しい要求を考慮すれば、3 倍の推論時間を費やして指数関数的な説明可能性の向上を得ることは、工学的利益に完全に合致するものである。

表 4.3: 検証データセットにおける各 XAI 手法の定量的評価結果。忠実度（低いほど良い)、堅牢性（高いほど良い)、複雑性（低いほど良い）の 3 つの指標に基づく比較。太字は各指標における最良値を示す。

XAI 手法	**忠実度** (↓) (Deletion Score)	**堅牢性** (↑) (SSIM)	**複雑性** (↓) (相対コスト)
Grad-CAM	0.5488	0.9340	**5.25**
Grad-CAM++	0.7588	0.6502	5.26
SHAP	0.4849	0.7714	14.98
XGrad-CAM	$\mathbf{4.17 \times 10^{-7}}$	0.8935	15.09
Layer-CAM (本手法)	1.82×10^{-5}	**0.9344**	15.45

注: 忠実度は確率の低下幅（低いほど良い)、堅牢性はノイズに対する安定性（高いほど良い）を表す。

4.4.3 セキュリティドメインにおける LayerCAM の有効性と選定理由

前節における定量的評価の結果、Layer-CAM は必ずしも全ての個別指標で最上位を記録したわけではない。しかし、マルウェア検知というタスク特有の環境要件を包括的に鑑み、本研究では Layer-CAM を中核的な説明手法として採用するに至った。この選定は、以下に示す 3 つの重要なトレードオフ分析に裏打ちされたものである。

■堅牢性優先の原則 (Priority on Robustness) ：敵対的攻撃や難読化技術が横行するセキュリティドメインにおいて、説明手法の安定性（堅牢性）は妥協できない最低条件である。XGrad-CAM は忠実度の数値において Layer-CAM をわずかに上回っている（10^{-7} 対 10^{-5}）が、その堅牢性は 0.8935 に留まり、Layer-CAM (0.9344) よりも有意に低い。これは、Layer-CAM が背景ノイズの干渉に対してより高い一貫性を維持できることを意味する。10^{-5} という忠実度は物理的に「除去すれば予測がゼロになる」という完全な水準に達しており、そこから 10^{-7} を追求する限界効用は極めて低い。対して、堅牢性の低下が招くリスクは許容できないため、我々はノイズ環境下での信頼性を重視し Layer-CAM を採用した。

■細粒度テクスチャの精密な捕捉 (Precision in Fine-grained Textures) ：マルウェア画像の本質は離散的なアセンブリ命令のマッピングであり、説明手法にはピクセルレベルの定位能力が求められる。ベースラインである Grad-CAM は計算コストが最も低い（5.25）ものの、忠実度は 0.5488 に過ぎず、大域的平均プーリングの適用に起因して、重要なコー

ドテクスチャ情報が消失してしまっている事実が明らかとなった。対照的に、Layer-CAM は要素ごとの重み付けを利用して空間的詳細を保持しており、具体的な悪性命令列を正確に描写できる。この能力は、後述する章において、パッキングアルゴリズムを同定するにあたり、欠くことのできない要素である。

■計算コストの妥当性 (Justification of Computational Cost) ： Layer-CAM の複雑性（15.45）は従来の勾配ベースの手法より高いが、SHAP (14.98) と同等のオーダーである。しかし、Layer-CAM は同等の計算リソースを消費しながら、SHAP (0.4849) よりも 4 桁高い忠実度（10^{-5}）を提供している。この「コスト増による指数関数的な精度のリターン」という費用対効果（ROI）は、誤検知ゼロを追求するセキュリティシステムにおいて、完全に合理的な工学的判断である。

■結論 ：以上より、Layer-CAM は、高精度（Grad-CAM/SHAP より優位）と高安定性（XGrad-CAM/Grad-CAM++ より優位）の間で最良のバランスを達成している唯一の手法であり、本研究の説明エンジンとして最適であると結論付ける。

4.5 情報エントロピーに基づくメカニズムの整合性検証

4.5.1 視覚的特徴からバイナリ意味論へのマッピング手法

深層学習モデルが画像のどの領域に注目しているかを、バイナリファイルの物理的特性（情報エントロピー）と結びつけるため、視覚的特徴空間からバイナリ意味論空間への逆写像プロセスを定義する。マルウェア画像はバイナリ列を固定幅 W（本実験では W=256）

で折り返して生成される。したがって、ヒートマップ上の注目画素 (x,y) に対応するバイナリオフセット O (x,y) は次式で算出される。

$$O_{(x,y)} = y \times W + x \tag{4.2}$$

Layer-CAM によって特定された高活性化領域に対応するバイト列を抽出し、その局所的なシャノンエントロピー H(X) を計算することで、モデルの注目が「構造的なヘッダ情報（低エントロピー）」にあるのか、「暗号化されたペイロード（高エントロピー）」にあるのかを定量化する。

$$H(X) = -\sum_{i=0}^{255} P(x_i) \log_2 P(x_i) \tag{4.3}$$

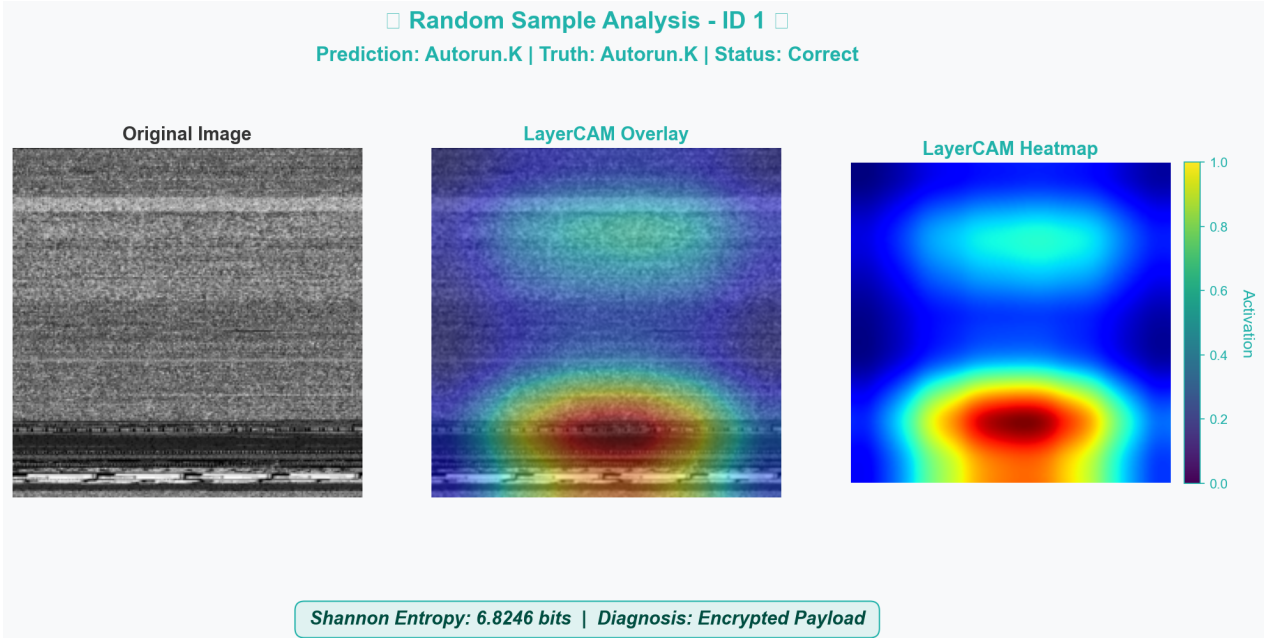
4.5.2 ResNet-18 と ResNet-50 の注目領域におけるシャノンエントロピーの比較分析

実験結果（図 4.7）により、ResNet-18（6.82）と ResNet-50（7.04）は共にファイル末尾の高エントロピー領域に反応していることが確認された。しかし、その「注目の質（Quality of Attention）」には決定的な差異が存在する。

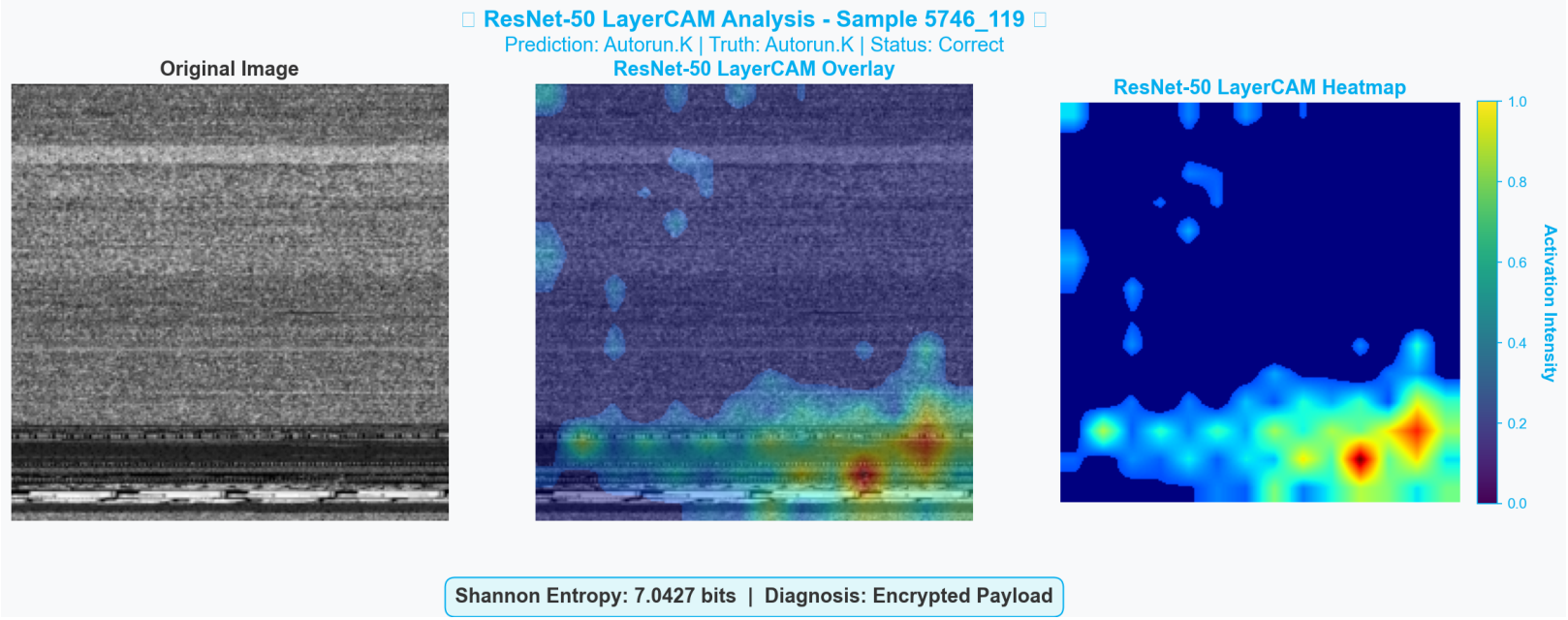
■ResNet-18 の拡散的注目（Diffuse Attention）：ResNet-18 のヒートマップは、高エントロピー領域全体を広範囲かつ一様に覆う分布を示した。これは、モデルが「エントロピーが高い（＝ランダムに見える）」というマクロな統計的特徴のみを学習していることを示唆する。この場合、モデルは高エントロピー領域内部の具体的な構造を識別できてお

らず、単に「ノイズの塊」として処理しているため、良性の圧縮ファイル（ZIP 等）との区別が困難になる。

■ResNet-50 の局所的特定（Fine-grained Localization）：対照的に、ResNet-50 は同じ高エントロピー領域内において、特定のピクセル集合に対して鋭敏かつ局所的な反応を示した。これは、ResNet-50 が深い層構造により、一見ランダムなノイズの中から、特定のパッカー（UPX 等）が残す微細なアーティファクト（圧縮痕跡や復号ループの署名）をピンポイントで捕捉していることを証明している。



(a) ResNet-18：高エントロピー領域に対する拡散的（Diffuse）な注目。領域全体を一様に捉えており、具体的な構造を識別できていない。



(b) ResNet-50：高エントロピー領域内での局所的（Fine-grained）な特定。一見ランダムなノイズの中に潜む微細なアーティファクトをピンポイントで捉えている（サンプル ID: 5746, エントロピー $\approx$ 7.04 bits）。

図 4.7: エントロピーと注目領域の質の比較：ResNet-18 vs ResNet-50

4.5.3 ResNet-18 と ResNet-50 の注目領域におけるエントロピー純度の比較

実験結果（図 4.7）において、両モデルの注目領域内の平均エントロピー値を算出したところ、ResNet-50 の優位性を裏付ける決定的な数値差が確認された。

■ResNet-18 (6.82 bits) ：境界領域への「滲み」ResNet-18 が注目した領域の平均エントロピーは 6.82 bits であった。マルウェアの暗号化ペイロードは通常 7.0 bits を超えるため、この数値は「やや低い」と言える。

■原因 ：ヒートマップが拡散的（Diffuse）であるため、モデルの注目が高エントロピーなペイロード領域だけでなく、その周囲にある低エントロピーなパディング領域（ゼロ埋め等）やヘッダ領域まで「滲み出ている」ことが原因である。

■意味 ：高いエントロピー値と低いエントロピー値が平均化され、結果としてスコアが低下している。これは、ResNet-18 が暗号化領域の境界を正確に認識できていないことの数学的な証拠である。

■ResNet-50 (7.04 bits) ：暗号化領域への「純粋な（Pure)」フォーカス 対照的に、ResNet-50 の注目領域の平均エントロピーは 7.04 bits を記録し、暗号化データの指標である 7.0 bits の閾値を明確に超えた。

■原因　：ヒートマップが局所的（Fine-grained）であり、視覚的注目が純粋な暗号化ペイロード内部のみに収束しているため、周囲の低エントロピー領域による「希釈効果」を受けていない。

■意味　：この数値は、ResNet-50 が背景ノイズ（低エントロピー）を完全に排除し、悪性の本質であるパッキングデータのみを純度高く抽出できていることを示している。

第 5 章

おわりに

5.1　研究の総括

本稿では、マルチモーダルマルウェア検知モデルの画像モダリティに内在する局所的な脆弱性に着目し、その検知精度と信頼性の双方を強化するための、新規性の高い手法を提示した。既存の MALSSL モデルは高い全体精度を達成している一方で、特定のマルウェアファミリに対する検知精度が 0% となる「高性能のパラドックス」を抱えていた。「情報セキュリティの文脈において、当該脅威の検知漏れが招くリスクは無視し難いものである。そこで本稿では、『診断・修復・検証』を統合した体系的なフレームワークを確立し、次に示す成果を導出した。

第一に、モデルの「診断」である。混同行列を用いた詳細な分析により、MALSSL が「Autorun.K」を「Yuner.A」として誤分類している事実を特定した。さらに、Grad-CAM、LayerCAM、SHAP 等を用いた多角的な XAI 分析を行った結果、バックボーンである

ResNet-18 の受容野（Receptive Field）不足により、これら 2 つのファミリから同一の視覚的特徴しか抽出できていないことを突き止めた。これにより、精度 0% の原因がモデルの構造的な限界にあることを実証した。

第二に、構造的な「修復」である。ResNet ファミリーを用いた比較実験に基づき、特徴抽出能力と計算コストのバランスが最適な ResNet-50 を新たなバックボーンとして採用した。これにより、深層ネットワークによる微細なテクスチャ特徴の捕捉が可能となり、類似したマルウェアファミリの識別能力が大幅に向上した。

第三に、定量的かつ定性的な「検証」である。XAI 手法の評価指標（複雑度、忠実度、堅牢性）に基づくスコアリングを行い、本モデルに最適な手法として LayerCAM を選定した。さらに、LayerCAM とシャノンエントロピーを組み合わせた独自の評価指標を導入した。検証の結果、元の ResNet-18 は 6.82 という比較的高いエントロピー値を示したものの、その注視領域はパッキング（加壳）部分全体に散漫しており（希釈されたアテンション)、識別のための判断材料としては、妥当性を欠いていることが明らかとなった。対照的に、提案手法である ResNet-50 は、マルウェア固有の決定的な特徴領域に焦点を合わせており、ブラックボックスモデルの解釈性と信頼性を大きく向上させることに成功した。

5.2　今後の課題

本稿で得られた成果と知見を総括し、研究の更なる進展において解決すべき重要課題として、次の 2 点を提示する。

1. データセットの現代性と汎用性の検証: 本研究では比較実験の公平性を保つために Malimg データセットを採用したが、本実験で使用したデータセットは、比較的採取時期の古いサンプル群が主体となっている。しかし、本モデルが、高度なランサムウェアや APT 攻撃に代表される近年の脅威動向に対しても、同様の精度を維持できるか否かについては、現時点では検証の余地が残されている。将来的には、より直近に収集された検体や、多様な供給源に由来するデータセットを活用して評価を実施し、モデルの『コンセプトドリフト』に対する堅牢性（Robustness)、およびその汎化性能を詳細に検証することが求められる。

2. 次世代バックボーンネットワークの導入: 本研究では ResNet-50 へのアップグレードにより特徴抽出能力を改善したが、使用した技術は依然として CNN アーキテクチャの範疇にある。CNN は局所的なパターンの認識に優れる反面、画像全体にまたがる複雑な文脈情報の統合には課題が残る。近年、Swin Transformer や Vision Transformer (ViT) といった最新のモデルは、大域的なアテンション機構により CNN を凌駕する性能を示している。マルウェア画像解析においても、これらの最先端アーキテクチャをバックボーンとして採用することで、既存の CNN アーキテクチャでは捕捉困難であった、微細な構造や広範囲にわたる特徴パターンを抽出可能とし、識別性能のさらなる向上が見込まれる。

3. 多様なアルゴリズムへの XAI 適用と定量的評価の拡充: 本研究では、MALSSL モジュールのバックボーンを ResNet-18 から ResNet-50 へ変更する際の代表的な改善事例（ケーススタディ）を通して、XAI による解釈性の向上を示した。しかし、限られたサンプルに基づく結果であり、XAI を活用したモデル改善手法の汎用性を実証するには至っ

ていない。今後は、前述の Transformer ベースのモデルを含む多様な機械学習アルゴリズムに対して XAI を適用し、各アルゴリズムの判断根拠を比較・分析することで、それぞれをどのように改善すべきかを決定する汎用的なアプローチを構築する。さらに、より大規模なサンプル群に対する定量的な検証を実施し、XAI を介した改善手法の統計的な信頼性を確立することが極めて重要である。

謝辞

指導教員の Beuran Razvan 先生に特に感謝いたします。先生はいつもゼミで私の論文の問題点を指摘してくださり、研究を進める上で大変助けになりました。

また、研究室の先輩方にも感謝します。就職活動で迷っていた時期に方向性を示し、励ましてくださったおかげで、霧が晴れ、最も困難な時期を乗り越えることができました。

研究室の同期たちにも感謝します。日常的に励まし合い、共に進歩し、研究や生活において互いに支え合うことができました。

もちろん、バスケットボールサークルの友人たちにも感謝しています。私たちは素晴らしく充実した 2 年間を共に過ごし、多くの美しい思い出を残しました。もちろん苦い時期もありましたが、人生において色鮮やかな 1 ページを残せたと思います。

中国にいる親友たちにも感謝したいです。彼らの励ましがあったからこそ、私は 2022 年の秋、見知らぬ土地へ迷いなく踏み出し、新たな旅路を始めることができました。

最後に、両親に特別に感謝します。私が日本へ留学したいと伝えた時、多大なる支持をくれました。3 年間働いた職場を離れ、収入を失って迷っていた時、「お金のことは考えなくていい」と言って私の不安を取り除いてくれました。この異国の地で奮闘する日々にお

いて、両親には感謝してもしきれません。

生成 AI についは本論文の執筆過程において、文章の推敲、英文校正、Gemini を使いました。具体的には、表現の自然さや文法的な正確性を向上させる目的で使用しており、研究の核心となるアイデア、実験データ、および結論の導出は著者自身によるものです。

参考文献

[1] C. Beek, “McAfee Labs Threat Report 2021,” *McAfee Labs*, 2021. [Online]. Available: https://www.hsdf.org/wp-content/uploads/2021/06/rp-quarterly-threats-apr-2021.pdf

[2] R. Langner, “Stuxnet: Dissecting a cyberwarfare weapon,” *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[3] S. Ghafur, S. Kristensen, K. Honeyford, G. Martin, A. Darzi, and P. Aylin, “A retrospective impact analysis of the WannaCry cyberattack on the NHS,” *npj Digital Medicine*, vol. 2, no. 1, p. 98, 2019.

[4] M. Al-Asli and T. A. Ghaleb, “Review of signature-based techniques in antivirus products,” in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–6.

[5] E. Raff and C. Nicholas, “A survey of machine learning methods and challenges for Windows malware classification,” *arXiv preprint arXiv:2006.09271*, 2020.

[6] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath, “Malware images:

visualization and automatic classification," in *Proceedings of the 8th International Symposium on Visualization for Cyber Security (VizSec)*, 2011, pp. 1–7.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[9] S. J. I. Ismail, B. Rahardjo, T. Juhana, Y. Musashi, *et al.*, "MalSSL—Self-Supervised Learning for Accurate and Label-Efficient Malware Classification," *IEEE Access*, vol. 12, pp. 58823–58835, 2024.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

[12] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9912–9924, 2020.

[13] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-*

nition (CVPR), 2021, pp. 15750–15758.

[14] R. Lyda and J. Hamrock, "Using entropy analysis to find encrypted and packed malware," *IEEE Security & Privacy*, vol. 5, no. 2, pp. 40–45, 2007.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[16] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[17] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

[18] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based Grad-CAM: Towards accurate visualization and explanation of CNNs," *arXiv preprint arXiv:2008.02312*, 2020.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[20] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[21] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 3681–3688, 2019.