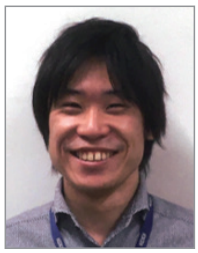
特集
Special Feature

［AI 時代のサイバーセキュリティ］

3 AI によるサイバーセキュリティ防御
― AI を活用したセキュリティ対策研究の最前線―

応用

清本晋作　中原正隆　成定真太郎　長谷川健人　（株）KDDI 総合研究所

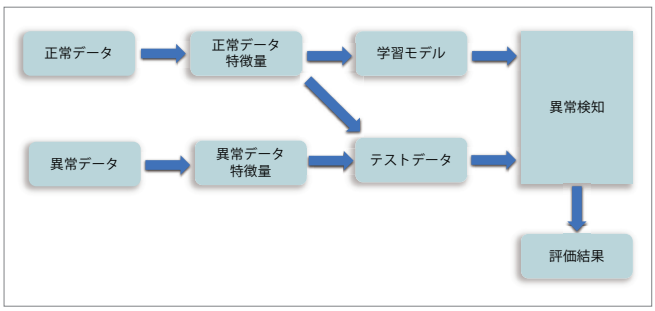
AI 活用の時代へ

深層学習等による機械学習技術の進歩により，実社会での AI 活用が加速度的に進展している．セキュリティ対策においても，それは例外ではなく，AI を活用したさまざまなアプローチが試みられている．具体的には，AI を使って攻撃を検知するといった試みに加え，AI を使って膨大なオープンソース情報を分析する試みや，AI を使って脆弱性の有無についての検査を効率化するような手法も検討されている．セキュリティ対策における AI 活用の目的は，監視対象の爆発的な増加に伴って省力化を図りたい，人手では解析が困難な分量の情報の分析を自動化したい，人が見落とす可能性が高い部分を効率良く発見したい，将来の高度セキュリティ人材の不足に備え対策システムの高度に自律化を図りたい，等が考えられ，課題解決の救世主として AI に対する期待も大きくなってきている．

一方で，AI をセキュリティ対策に活用する場合の課題，たとえば AI による判断の正確さを十分保証できるかどうか，なども徐々に浮き彫りとなってきており，課題を解決するための研究開発の活発に行われている．加えて，AI が攻撃者に悪用される事態も想定し，先回りした研究開発により対策を進めなければならない．本稿では，AI のセキュリティ対策への活用，特に AI を活用した異常検知について，その概要，ならびに，ネットワーク，ソフトウェア，ハードウェア，それぞれのセキュリティ対策への適用事例を述べる．また，本稿の最後で，AI のセキュリティ対策への活用における課題と，今後の研究開発の方向性について論じる．

AI による異常検知

本章では，AI による異常検知の一般的な流れを解説する．精度評価を含む処理フローの一例を図 -1 に示す．この例では，正常な挙動のみを学習データとしている．



■図 -1　異常検知の処理フロー例

AI を異常検知に用いるためには，まず学習および検知に必要なデータを取得する（図 -1 の「正常データ」，「異常データ」に相当）．たとえばネットワークトラフィックの異常検知の場合，図 -1 に示すように，トラフィックデータから正常データと異常データをそれぞれ収集する必要がある．

続いて，取得したデータのうち，正常データを対象としたデータから 学習に用いる特徴量を抽出する．トラフィックデータの場合，送信元・宛先ポート番号，送受信パケット数・バイト数などが用いられることが多い．特徴量は必要に応じて正規化や二値化などの処理が行われる．続いて，データから抽出した特徴量を機械学習や深層学習のアルゴリズムに入力し，異常検知のためのモデルを学習する．

異常検知のタスクにおいては，教師なし学習にて正常データのみを学習し，学習データから逸脱したデータを異常と判断するのが一般的であった．しかし，近年，教師あり学習にて正常データと異常データの両方を用いて分類モデルを学習させて異常検知を行う研究も少なくない．

精度評価においては，学習済みのモデルに対し，正常・異常の両方が含まれるテスト用データから抽出した特徴量を入力することで，異常検知の評価を行う．そのため，図 -1 における正常データ特徴量は学習データとテストデータの両方に分けて使われており，異常データ特徴量はテストデータのみに使われる形となる．異常検知の評価では，正常／異常を正しく判定できた割合である正解率（Accuracy），異常と判断したデータのうち実際に異常であるものの割合である適合率（Precision），異常データを正しく異常と判定できた割合である再現率（Recall）等が用いられる．重視される評価指標はユースケースによって異なり，たとえばアラート監視の負荷を軽減したい場合には正常データを異常と誤検知する数を少なくするよう，チューニングやアルゴリズムの改良が行われる．

ネットワークに対する異常検知

AI を用いた異常検知技術が有用であるケースとして，ネットワークにおけるセキュリティ対策への活用が考えられる．特に，IoT は従来の PC やスマートフォンと比べて台数が多く，また画面を持たないデバイスも多いため，利用者が自発的に異常に気付き，対策を行うことが難しい．そのため，AI を用いて自動的に異常検知を行うシステムが特に有効である．本章では，機械学習を用いたネットワークにおける異常検知，特に IoT 向け異常検知技術の事例を紹介する．

IoT の主なユースケースの 1 つとして，工業 IoT がある．工業 IoT はその性質上，攻撃を受けた際の物理的な被害が大きく，また通信形態も通常の IT システムとは異なるため，特別な対策が必要である．また，用途が限定されており，あらかじめ定められた動作を定常的に行うものも少なくないため，正常通信をパターン化しやすいという側面もある．

文献 1）では，工業 IoT の脆弱性を対象としたさまざまな攻撃通信を対象に，各種アルゴリズムにおける検知の精度を評価している．評価においては，まず正常系として貯水タンクの水位と濁度を監視する IoT システムを構築している．タンク，ポンプ，バルブ，センサが PLC（Programmable Logic Controller）を通して接続されており，Modbus というプロトコルで制御のための通信が行われているシステムである．

そして，攻撃シナリオとして，攻撃者からシステムへの接続を確立するバックドア攻撃，システムやデータベースの脆弱性を用いて悪意のあるコマンド・クエリを実行するインジェクション攻撃を実行し，正常時・攻撃発生時のトラフィックを取得し，評価用のデータセットを作成している．

続いて，用意したデータから特徴量を抽出している．ここでは，送信元･宛先ポート番号，パケット数，バイト数などから 23 次元の特徴量を抽出している．

データセットのうち，学習用に分割したデータを用いてモデルを訓練し，テスト用に分割したデータを用いて汎化性能を評価している．

アルゴリズムとしては，SVM（Support Vector Machine），KNN（K-Nearest Neighbor），NB（Naive Baize），RF（Random Forest），LR（Logistic Regression），ANN（Artificial Neural Network）を用いて比較している．結果として，RF が正解率や誤検知率などさまざまな指標で最もよい性能を示していた．

また，2016 年にマルウェア Mirai に感染したデバイスによる大規模な DDoS（Distributed Denial of Service）攻撃が起こったように，IoT デバイスにおけるマルウェア感染に起因する異常トラフィックの検知も重要な課題である．トラフィックの常時監視が望ましいが，前述のように IoT は台数が多いため，異常検知に必要なデータ量の削減が求められる．

そこで中原らは，軽量なデータを用いた異常検知手法を提案している [2]．本手法ではモデルの学習および異常検知に，通信の一連のやりとりを集約した情報であるフローデータを用いる．フローデータには通信の送信元・宛先の IP アドレスやポート番号など，通信の振る舞いを示す情報が含まれており，かつデータの中身は含まれていないため軽量である．一方で，データの中身を含めた場合に比べて異常検知に使えるデータが限定されるため，限られた情報で精度よく異常検知を行うための工夫が必要となる．

本手法では，IoT デバイスの挙動が限定的であることに着目して特徴量を設計したり，ホワイトリストを併用して異常検知対象とするデータを絞ったりすることで，誤検知数の削減に成功している．

ソフトウェアに対する異常検知（マルウェア検知）

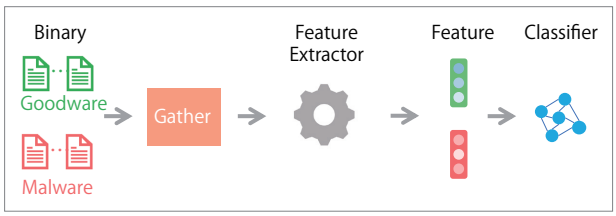
企業や個人を狙ったマルウェアとその亜種による攻撃手法は高度化の一途を辿っている．情報処理推進機構（IPA）が毎年発表している「情報セキュリティ 10 大脅威」では，2022 年度の組織に対する 10 大脅威の第一位にランサムウェアがランクインしていることからも，高度化したマルウェアを検知するシステムを開発することは急務となっている．

マルウェア検知システムには，大きく分けて，従来型の「パターンマッチング方式」と，近年研究が盛んな「機械学習（AI）に基づく方式」の 2 つがある．前者は，マルウェアが持つ特徴を「パターンファイル」としてあらかじめ準備しておき，このファイルに基づいてマルウェア分類を行う．しかし，この方式では検知が困難な未知のマルウェアやミューテーションと呼ばれる高度なマルウェアに対しては，AI に基づく検知システムによってより詳しい分析を行う必要がある．

AI に基づく検知システムの大まかな流れを図 -2 に示す．AI におけるマルウェア検知においては，通常はバイナリデータが対象となる．そして，バイナリデータから特徴量を抽出し，AI に入力するという流れになる．

ミューテーションがパターンマッチングによる検出を回避することを目的とする一方で，AI を誤判定させることを目的とする高度なマルウェアが存在し得ることが近年の研究により明らかとなっている [3]．これらの文献では，敵対的機械学習が AI によるマルウェア検知を誤判定させる領域に応用されている．

敵対的機械学習とは，機械学習システム全般に対する攻撃手法の総称であり，著名なものに，機械学習システムの入力データに微小なノイズを加えることにより，システムの誤検知を誘発させる「敵対的サンプル」や，「バックドアポイズニング攻撃」と



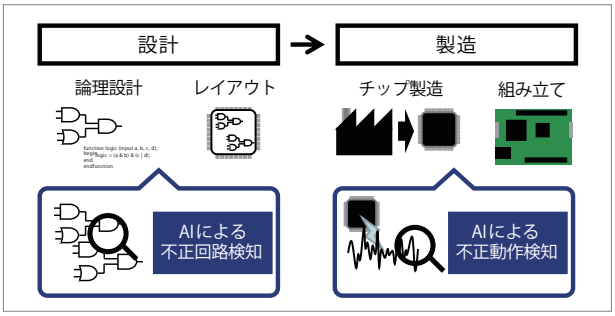
■図 -2　AI によるマルウェア検知

呼ばれる手法がある．

バックドアポイズニング攻撃とは，攻撃者が「トリガ」と呼ばれるノイズが付与されたデータ（毒データ）を，機械学習システムの訓練データに混入させることにより，システムが「トリガ」の付与されたデータのみを誤分類することを目的とした攻撃手法である．文献 3）では，実際のマルウェアに対して，マルウェアの機能を損なわずにトリガとなる微小なバイナリが付与できることを示した．さらに，トリガ付きのマルウェアが万一訓練データとして収集されてしまうと，攻撃が成功し，AI がトリガ付きマルウェアを良性ソフトウェアと誤判定する可能性があることが示された．

マルウェア検知システムへの上記の攻撃の対策として，Spectral Signature や Isolation Forest といった既存の異常データ検知アルゴリズムを使用して毒データを除去する方法が考えられる．しかし，一度訓練データが汚染されてしまうとこれらの対策では攻撃を完全に防ぐことが難しい．

一方で，AI によるマルウェア検知システムにおいては，分類モデルを訓練するためにソフトウェアデータを収集することは必要不可欠である．このとき，信頼できる提供元からのみデータを収集することが可能ならば，このような攻撃を防ぐことができるが，それができない場合は，訓練データに毒データが含まれていることを仮定した上で，「データの無毒化」を実施することで毒データの影響を低減できる [4]．



■図 -3　ハードウェアサプライチェーンと設計，製造工程おける不正検知

具体的な方法としては，分類モデルに与える特徴ベクトルから攻撃者による汚染が想定される次元を削減することや，特徴ベクトルをオートエンコーダや GAN による擬似データに置換することが挙げられる．しかし，これらの方法では分類精度が犠牲になるといった課題もある．

ハードウェアに対する異常検知

ハードウェア製品の大量生産，低価格化，開発サイクルの高速化に伴い，ハードウェアサプライチェーンは複雑化している．しかしながら，ソフトウェアと比較してハードウェア産業は参入障壁が高く，またサプライチェーンも多くがクローズドであることから，これまでは脅威として顕在化していなかった．昨今の国際情勢，半導体供給の不安定性，ならびにサプライチェーンのオープン化を鑑みれば，今後はハードウェアサプライチェーンにおけるセキュリティ技術は重要な位置づけとなる．そこではAI を活用した効率化が有望な解となり得る．

従来，ハードウェアは暗黙的に Root of Trust（システムの信頼性を保証する基点）と見なされ無条件で「信頼できるもの」とされていた．しかし，近年ではハードウェアレベルでのセキュリティ懸念が指摘されており，ハードウェアサプライチェーンにおけるセキュリティへの関心は高まっている [5]．

図 -3 にハードウェアサプライチェーンの代表的な工程と，不正検知の例を示す．ハードウェアサプライチェーンにおける防御を効率化するため，AI を活用した手法が数多く研究されている．

ハードウェアはハードウェア記述言語（HDL）を用いて設計される．サプライチェーンの複雑化により，この設計情報に悪意のある機能が組み込まれる危険性が指摘されている．ハードウェアに組み込まれた悪意のある機能はハードウェアトロイと呼ばれ，この十年で多数の検知手法が提案されている．通信インタフェースやプロセッサ等，さまざまな

ハードウェアに共通して利用できる部品は Intellectual Property（IP）コアと呼ばれるモジュール形式で提供されることがある．加えて，RISC-V を始めとして回路のオープンソース化が進むなど，ハードウェアコミュニティは広がり続けている．

こうした背景のもとで，HDL で記述された IP コアにハードウェアトロイが仕込まれる可能性がある．ハードウェアトロイはごく小規模な回路で普段は活性化しないため，検知が難しい．情報漏洩や機能停止を引き起こすことで，利用者の業務妨害や製品ベンダの信用失墜，さらにはインフラの停止を招く危険性が指摘されている．

ハードウェアトロイの機能的な特徴から，回路の動作や構造に特有の特徴が見られることが知られている．この特徴を利用することで，HDL の解析によりハードウェアトロイを検知できる．ところが，大規模なハードウェア設計情報からごく小規模なハードウェアトロイを検知する必要があり，そのためにはハードウェアトロイにだけ共通するハードウェアトロイ検知に効果的な特徴を見つけなければならない．

そこで，ハードウェアトロイに特有な特徴を AI で学習し，ハードウェアトロイを検知する手法が研究されている．近年では，回路をグラフ構造として表現できることを利用し，Graph Neural Network（GNN）を活用した検知手法 [6] が提案されており，高い精度での検知を実現している．

一度ハードウェアが IC チップとして製造されると，その内部を完全に明らかにするのは困難である．そこで，AI を活用して IC 内部の動作を推測する方法が研究されている．

代表的な手法として，サイドチャネル解析が挙げられる．サイドチャネル情報とは，IC チップを利用する際に入出力する情報とは別に，消費電力や漏洩電磁波のように自然に漏洩してしまう情報を指す．サイドチャネル情報を正確に観測することで，IC チップ内部の動作を推測することが可能となる．

一般に，こうしたサイドチャネル情報の観測では観測データにノイズが重畳しており，データを適切に解析する必要がある．そこで，AI を活用して観測した波形を処理する手法が採られる．たとえば，正常な IC チップの波形を AI で学習し，与えられた波形が学習した波形と比較して異常かを判定する手法が提案されている [6]．

今後の展望

AI の活用によるセキュリティインシデント発見・分析の省力化・自動化は，急速に研究が進展している分野である．一方で，まだその精度や，継続的な運用におけるトレンド変化への対応，等に課題を残しており，現状ではさらなる技術の進歩が望まれている．

さらには，AI の弱点を突いて検知をかいくぐろうとする試みも登場するなど，AI に対抗する技術の潮流も生まれている．実利用の場面においては，単一の対策手法を用いて検知システムの可用性と攻撃への頑健性を両方担保しようとするのではなく，複数の対策手法を組み合わせたり，いくつかの検知モデルを併用したりしてシステム全体の精度や頑健性を高めることがより重要であると考えられる．

一方で，異常検知等の AI をさまざまなセキュリティ対策に利用する場合の最も大きな課題の 1 つは，実データセットの不足である．特に，実際の不正トラフィック，マルウェア，ハードウェアトロイ，など，異常な状態を学習するためのデータを十分な量を入手することは容易ではない．このため，自ら異常な状態を作り出すことで学習データを生成するとともに，強化学習を導入するといった新たなアプローチが必要になると考えられる．

最後に，AI を実運用する場合の課題を 1 つ述べたい．AI を実運用する場合には，その判断の妥当性を，根拠を持って確認できることが望ましい．たとえば，ネットワークにおいて異常を検知した

場合，次のアクションとして特定の通信を遮断するなどの対処を行うことが考えられるが，それには第三者に対して説明可能な判断の根拠を求められる場合がある．

このような課題を解決する技術として，いわゆる説明可能 AI の研究が盛んに行われているが，クリティカルな判断と対処が求められるセキュリティ分野においては，とりわけ重要な意味を持つ技術ではないかと考えられる．まだ，その実用性が十分検証されてはいないが，今後の技術の発展が大いに期待されるものである．

参考文献

1）Zolanvari, M., Teixeira, A. M., Gupta, L., Khan, M. K. and Jain, R. : Machine Learning-Based Network Vulnerability Analysis of Industrial Internet of Things, IEEE Internet of Things Journal, Vol.6, No.4, pp.6822-6834 (2019).
2）Nakahara, M., Okui, N., Kobayashi, Y. and Miyake, Y. : Malware Detection for IoT Devices using Automatically Generated White List and Isolation Forest, Proc. IoTBDS 2021, INSTICC, pp.38-47 (2021).
3）Severi, G., Meyer, J., Coull, S. and Oprea, A. : Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers, USENIX Security 21 (2021).
4）Narisada, S., Matsumoto, Y., Hidano, S., Uchibayashi, T., Suganuma, T., Hiji, M. and Kiyomoto, S. : Countermeasures Against Backdoor Attacks Towards Malware Detectors. CANS2021 (2021).
5）長谷川健人，戸川　望：スパイチップはあるのか　ハードウェアセキュリティの必要性，情報処理，Vol.60, No.1, pp.4-6 (Jan. 2019).
6）Hasegawa, K., Yamashita, K., Hidano, S., Fukushima, K., Hashimoto, K. and Togawa, N. : Node-wise Hardware Trojan Detection Based on Graph Learning, arXiv (2021).

（2022 年 5 月 9 日受付）

■清本晋作　sh-kiyomoto@kddi.com

2000 年筑波大学工学研究科博士前期課程修了．博士（工学）．同年 KDDI（株）入社．現在，（株）KDDI 総合研究所に所属．

■中原正隆　ms-nakahara@kddi.com

2016 年京都大学大学院情報学研究科修士課程修了．同年 KDDI（株）入社．現在，（株）KDDI 総合研究所に所属．

■成定真太郎（正会員）　sh-narisada@kddi.com

2018 年東北大学大学院情報科学研究科博士前期課程修了．同年 KDDI（株）入社．現在，（株）KDDI 総合研究所に所属．

■長谷川健人（正会員）　kt-hasegawa@kddi.com

2020 年早稲田大学大学院基幹理工学研究科博士後期課程修了．博士（工学）．同年 KDDI（株）入社．現在，（株）KDDI 総合研究所に所属．