# 証明書に問題があるドメインの自動的な分類の実現

野村 尚加[1]　岡村 耕二[1]

**概要**：ASM ツールによって検出された証明書不備を持つドメインを, サイトのコンテンツを考慮したリスクの重要度に基づく分類を自動化で行った. 従来の ASM ツールは技術的要因に一律の深刻度を付与するため、資産の背景に応じたリスクの差別化が困難だった. この課題を, 生成 AI を用いてサイトのコンテンツ内容を高精度で推測し, リスクの重要度に基づく分類を自動で行った. 本手法により, 技術的なリスク要因に加え, サイトの詳細な状況を考慮した高度なリスクの自動的トリアージが可能となり, 効率的なセキュリティ運用に寄与が期待できる.

**キーワード**：リスク分析・評価, Web・メールセキュリティ, 不正・異常検出

# Implementation of automatic classification of FQDN with certification errors

Nomura Naoka[1]　Okamura Koji[1]

***Abstract:*** This study automated the classification of domains with certificate vulnerabilities detected by ASM (Attack Surface Management) tools, based on risk significance derived from site content. Traditional ASM tools assign a uniform severity based solely on technical factors, making it difficult to prioritize responses according to the operational background of each asset. To address this challenge, we utilized Generative AI to accurately infer site content and automatically categorize assets based on their risk importance. This approach enables advanced, automated risk triaging that considers both technical risk factors and the specific context of the site, promising significant contributions to more efficient security operations.

***Keywords:*** Risk analysis and assessment, Web/Mail security, Fraud/error detection

## 1. はじめに

近年, デジタルトランスフォーメーション（DX）の加速やクラウドコンピューティングの普及により, 組織が保有する IT 資産は急速に拡大し複雑化している. そのため, 組織自身が自社の IT 資産の全体像を把握することは困難である. 加えて, 組織が所有する特定の IP アドレスや普及しているドメインのみを対象とした従来のスキャン手法では, 稼働しているサービスの約 50 %を見逃すなど, 既存の手法ではリスクの展望を断片的にしか捉えられないことが実験的に示されている [3]. このような背景から, 外部からの視点で多角的な情報源を組み合わせ, 自組織の IT 資産を自動的に探索・継続的監視することで, 潜在的な脆弱性を特定する手法である ASM（Attack Surface Management：攻撃対象領域管理）の重要性が増している. ベンダーが提供する ASM ツールを活用することで, 効率的な IT 資産管理が可能になる.

ASM ツールは, IT 資産の発見をすることは得意である一方で検出したリスク内容（脆弱性の内容）については, 同一の深刻度（セベリティスコア）を付与する. しかし, 同一のリスク内容が存在する資産でも, 資産の運用状況に

[1] 九州大学工学部電気情報工学科
Kyushu University Undergraduate school of Engineering Department of Electrical Engineering and Computer Science

よって, そのリスクの重要度は異なる. つまり, 既存の ASM ツールでは, 資産背景の詳細な状況を考慮したリスクの重要度を把握することは困難である.

加えて, セキュリティ分野におけるアラート疲れは深刻な問題である. セキュリティ運用センター（SOC）におけるアラート疲れに関する研究では, 誤検知の多さやアラート量の多さがが担当者の認知負荷を増大させ, 真に重大な脅威への対応を遅らせる要因となることが示されている [9]. そのため検出されたアラートの深刻度や重要度がより詳細かつ正確であることが求められる.

これらの課題を解決するために, 管理者が目視でサイトの運用状況やリスクの重要度を確認する手法も考えられるが, それには多大な人的コストと時間を要する. Web サイトや証明書の状況は日々変化するため, 手動調査では変化の激しい攻撃対象領域（Attack Surface）に対してリアルタイムな防御体制を構築することができず, 実用性の面で限界がある. 従って, 技術的なスキャンの迅速性を維持しつつ, 人間のような柔軟な文脈理解に基づいたリスクの自動選別を実現する新たな手法が求められている.

そこで本研究では, IT 資産が有するリスクについて, ASM が検出した技術的な要因に加え, 大規模言語モデル（LLM）である Gemini API を活用して, サイトの HTML コンテンツ等から運用上の背景的要因を自動的に推論・分類する手法を提案する. 本手法により, 膨大な資産の中から「ユーザ視点」で詳細な運用状況を把握し, 真にリスクの重要度の高いサイトを迅速に特定し, 効率的かつ詳細な分類を行うことが可能となることを検証する.

## 2. 関連技術の概要

### 2.1 ASM について

ASM（Attack Surface Management：攻撃対象領域管理）とは, インターネットに公開されており, 外部からの攻撃に曝される可能性のある組織の IT 資産を, 攻撃者の視点から継続的に発見, 監視, 分析, 評価しリスクを管理するプロセスである [4], [7], [10]. これには, 保護されていないサーバー, ドメイン名, クラウドサービス, 公開されているポートなど, 組織に認識されていない資産や放置されている資産も含まれる. ASM の目的は, 組織が持つ資産の包括的な全体像を把握し, 悪用される可能性のある外部攻撃対象領域を予測することにある. この技術により, 組織は未知のセキュリティリスクを低減し, 情報システムや資産をより適切に保護できるようになる [2]. なお, ASM においては外部からの攻撃に焦点を充てているプロセスを特に, 「外部の」という意味である「External」をつけて EASM と呼ぶことがあるが, 本論文では, ASM と EASM を同義として扱う.

本研究では, CyCraft 社が提供する ASM プラットフォームである「XCockpit EASM」を使用した. 本プラットフォームのダッシュボードでは, 組織が保有する IT 資産の状態を直感的に把握できるよう, 以下の構成で情報が可視化されている.

- IT 資産一覧表
  図 1 の上の表が, IT 資産の一覧表である. 発見されたドメインや IP アドレス等の資産がリスト形式で表示される. 各資産の左端には, その資産が抱える全体的なリスク状態を示すカラーバーが表示されており, 組織内のどの資産に優先的に対処すべきかを一目で判別できる.

- リスク要因の詳細
  図 1 の下の表が, 各 IT 資産のリスク要因の詳細を表す表である. IT 資産の一覧表から, 特定の資産を選択することで, その資産に関連する具体的なリスク要因（例：SSL/TLS 証明書の不備, 設定ミス等）の詳細を確認できる. ここでは、左端に各リスク要因に対して個別に算出されたセベリティスコアが表示される.

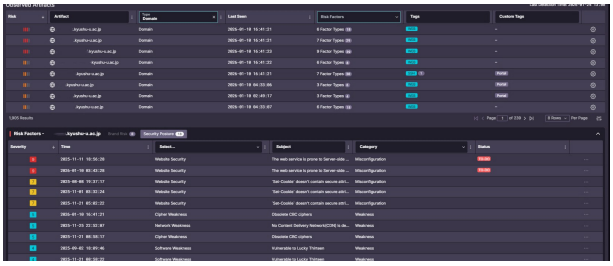

**図 1** ASM のダッシュボード
（※具体的なドメイン名は伏せている）

### 2.2 SSL/TLS サーバ証明書の概要

サーバ証明書とは, SSL/TLS プロトコルを用いた通信において, 通信相手であるサーバの正当性を証明するための電子証明書である. これは, 信頼できる第三者機関である認証局（CA: Certificate Authority）によって発行される. 証明書には, サーバの公開鍵や所有者の情報, 認証局のデジタル署名が含まれており, これによってデータの改ざん防止となりすましの検知が可能となる. 以下にサーバ証明書の役割と, 証明書に多く見られる不備に関して詳述する.

#### 2.2.1 サーバ証明書の役割

サーバ証明書は、SSL/TLS（Secure Sockets Layer / Transport Layer Security）プロトコルを用いた通信において, 主に「通信の暗号化」と「サーバの身元保証（真正性の証明）」という 2 つの極めて重要な役割を担っている. これらにより, インターネットという公開されたネットワーク上においても, 安全に機密情報をやり取りすることが可能となる.

(1) 通信の暗号化
サーバ証明書の最も基本的な役割は, クライアント（ブラウザ）とサーバ間でやり取りされ通信を暗号化し, 第三者が内容を解読できないようにすることである. サーバ証明書はこの暗号化のプロセスにおいて, サーバの「公開鍵」をクライアントへ安全に提供するための媒体として機能する.

(2) サーバの身元保証
第二の役割は、接続先のサーバが, 利用者がアクセスしようとしているドメインの正当な所有者によって運営されていることを証明する「身元保証」である.
サーバ証明書には, 信頼できる第三者機関である「認証局（CA: Certificate Authority)」によるデジタル署名が付与されている. クライアント側には, あらかじめ信頼できる認証局の公開鍵（ルート証明書）がインストールされており, これを用いて提示された証明書の署名を数学的に検証する. この「信頼の連鎖」に基づく検証プロセスにより, サーバの身元が保証される.

#### 2.2.2 証明書における主要な不備の種別

本研究において分析対象とする, SSL/TLS サーバ証明書における主要な 3 つの不備について述べる.

(1) ホスト名と SAN の不一致
サーバ証明書には, その証明書がどの通信相手（ドメイン名）に対して有効であるかを示す共通名（CN:Common Name）とサブジェクトの代替名（SAN:Subject Alternative Name）が含まれている. 現代の Web ブラウザは, アクセスした URL のドメイン名が, 提示された証明書の SAN（または CN）のリストに含まれているかを厳格に照合する. この際, 照合に失敗した状態は「SAN の不一致（Certificate SAN mismatch)」と判定される. これは、正しい通信相手を保証できないことを意味し, ブラウザ上で接続の中断を促す警告が表示される原因となる.

(2) 有効期限切れ
サーバ証明書の信頼性は, 第三者機関である認証局によって期間限定で保証される. 有効期限が切れた証明書（Certificate expired）は, そのドメインの所有権や公開鍵の正当性がもはや保証されていない状態を指す. これを放置することは, 組織の管理不備を露呈させるだけでなく, 攻撃者による中間者攻撃を許す隙を与える深刻なリスクとなる. 特に、大規模な組織においては資産の把握漏れにより, 期限切れのまま放置されるドメインが蓄積しやすい傾向にある.

(3) 自己署名証明書
自己署名証明書（Self-signed certificate）とは, 公的な認証局による審査や署名を経ず, サーバ管理者が自ら署名して発行した証明書である. この形式の証明書は, 内部ネットワークや開発段階のテスト環境での利用には適しているが, インターネットに公開されたサービスで使用する場合, 第三者による身元保証が一切存在しない. そのため、ブラウザは「信頼できない発行者」として警告を発し, 利用者に強い注意を喚起する.

上述した証明書の不備（SAN の不一致・有効期限切れ・自己署名）は, 証明書の重要な役割の一つである「身元保証」が機能しない状態を作り, データの盗聴や改ざんを行う中間者攻撃（Man-In-The-Middle attack）の隙を与える要因となる.

また, 不備のあるサイトが長期間放置されることは「警告の状態化（Warning Fatigue)」を招き, システム全体のセキュリティ耐性を心理面から脆弱化させる深刻な二次的リスクである.

このように証明書不備は, 技術的・心理的な両面で重大なセキュリティリスクを引き起こす.

### 2.3 大規模言語モデルについて

大規模言語モデル（Large Language Models: LLM）とは, 膨大なデータセットを用いて学習されたディープラーニングモデルの一種であり, テキストの認識や生成, 自然言語処理タスクを高度に実行できる人工知能プログラムである [1], [6], [8].

技術的基盤には「トランスフォーマー（Transformer)」と呼ばれるニューラルネットワーク・アーキテクチャが採用されている. これは単語のシーケンスを処理し, テキスト内の複雑なパターンや文脈を捉えることに長けている [1], [5]. この仕組みにより, LLM は従来のルールベースの手法では困難であった「サイトの運用実態」といった抽象的なコンテクストの推論において, 柔軟かつ高度な判定を可能にしている.

## 3. ASM ツールのデータ分析

本節では, CyCaft 社が提供する ASM ツールである XCockpit EASM を使用し, 組織の攻撃対象領域におけるリスク要因の推移を観測した手法について述べる.

### 3.1 観測対象

調査対象として, 九州大学が保有し, 使用しているドメイン「kyushu-u.ac.jp」およびその配下のサブドメインを設定した. 観測期間は, 2025 年 8 月 31 日から 2025 年 12 月 14 日までの約 3.5ヶ月間とした. 8 月から 10 月までの期間

は, 2 週間に一度データを取得した. 11 月と 12 月は, より詳細な動向を調査するため 1 週間ごとデータにを取得し, 検出されるリスク項目の推移を記録した.

XCockpit EASM では, 検出されたリスクに対して, その深刻度を示すセベリティスコアが割り当てられる. 本分析では, 攻撃者に悪用される可能性が高く, プラットフォーム側で「早急な確認が推奨される」と定義されているセベリティスコアが 7 以上のリスク要因を持つドメインに焦点を絞り, 調査を行った.

時間経過によって組織の公開資産のリスク状態がどのように変化するかを明らかにするため, 以下の 2 つの観点から動向を調査した.

- リスク要因の種類（脆弱性の種類）
  検出されたセベリティスコアが 7 以上のリスクが, 具体的にどのような技術的要因（例：SSL/TLS サーバ証明書の有効期限切れ, 強度の弱い暗号スイート, 設定不備等）に基づいているかを分類し, その出現頻度を観察した.

- 検出されたドメインの動向
  セベリティスコアが 7 以上のリスク要因が検出されたドメインに対して, リスク要因ごとに該当するドメインの顔ぶれに変化がないかを時系列で観察した. 具体的には, 同一のリスク要因に分類されたドメイン群において, 期間中にドメインの増加や減少の詳細を観察した.

### 3.2 分析結果：リスク要因の種類

表 1, 表 2 に, 2025 年 8 月 31 日から 12 月 14 日までのリスク要因ごとの検出件数を示す. 観察期間におけるセベリティスコアが 7 以上のリスク要因の推移を分析した結果, いくつかの増減が確認されたものの, 検出されるリスク要因の種類には大きな変化は見られなかった.

これより, 本調査対象においてリスク要因の種類については, 検出されるリスク要因の種類そのものに時間経過における劇的な変化はみられないことが分かった.

**表 1** 検出されたリスク要因の推移

| リスク要因 | 8/31 | 9/10 | 9/14 | 9/28 | 10/13 | 10/26 |
|---|---|---|---|---|---|---|
| Set-Cookie missing Secure attribute | 6 | 5 | 6 | 1 | 1 | 1 |
| WWW-Authenticate Uses Unsafe Authentication | 0 | 1 | 1 | 1 | 3 | 3 |
| The web service is prone to SSTI attack | 0 | 0 | 0 | 0 | 1 | 1 |
| Certificate expired | 11 | 9 | 13 | 13 | 22 | 29 |
| sp tag in DMARC record at subdomain | 3 | 3 | 2 | 2 | 2 | 2 |
| Set-Cookie doesn't contain secure attributes | 29 | 28 | 32 | 14 | 15 | 17 |
| Self-signed certificate | 2 | 2 | 2 | 2 | 5 | 9 |
| DMARC record sp value set to none | 3 | 3 | 2 | 2 | 2 | 2 |
| Certificate without SAN | 0 | 0 | 0 | 0 | 1 | 1 |
| Certificate SAN mismatch | 12 | 12 | 18 | 18 | 85 | 131 |
| Badly configured localhost records | 2 | 2 | 2 | 2 | 3 | 3 |
| **総件数** | **68** | **65** | **78** | **55** | **140** | **199** |

**表 2** 検出されたリスク要因の推移（2025 年 11 月 2 日–12 月 14 日）

| リスク要因 | 11/2 | 11/9 | 11/16 | 11/23 | 11/30 | 12/7 | 12/14 |
|---|---|---|---|---|---|---|---|
| Set-Cookie missing Secure attribute | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| WWW-Authenticate Uses Unsafe Authentication | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| The web service is prone to SSTI attack | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Certificate expired | 31 | 31 | 31 | 31 | 31 | 28 | 28 |
| sp tag in DMARC record at subdomain | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| Set-Cookie doesn't contain secure attributes | 18 | 18 | 18 | 19 | 19 | 19 | 19 |
| Self-signed certificate | 11 | 11 | 9 | 9 | 9 | 9 | 9 |
| DMARC record sp value set to none | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| Certificate without SAN | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Certificate SAN mismatch | 133 | 133 | 128 | 129 | 129 | 129 | 129 |
| Badly configured localhost records | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| **総件数** | **207** | **206** | **195** | **197** | **197** | **194** | **194** |

### 3.3 分析結果：検出されたドメインの動向

本節では, セベリティスコアが 7 以上のリスクが検出されたドメインの時系列的な変化について, 量的・質的な両面から詳述する.

観測期間を通じて, 深刻なリスクを抱えるドメインの総数は顕著な増加傾向を示した. 特に証明書に関連するリスク要因を持つドメインの増加が全体の件数を大きく押し上げる要因となっており, 具体的なリスク要因別の動向は以下の通りである.

#### 3.3.1 証明書関連のリスク（蓄積・増加傾向）

証明書不備に関するドメインは, 期間を通じて解消されることが少なく, 新規発生し蓄積していく傾向が顕著であった. 以下に示す証明書不備に関するリスク要因の検出数の推移を, 図 2 に示す.

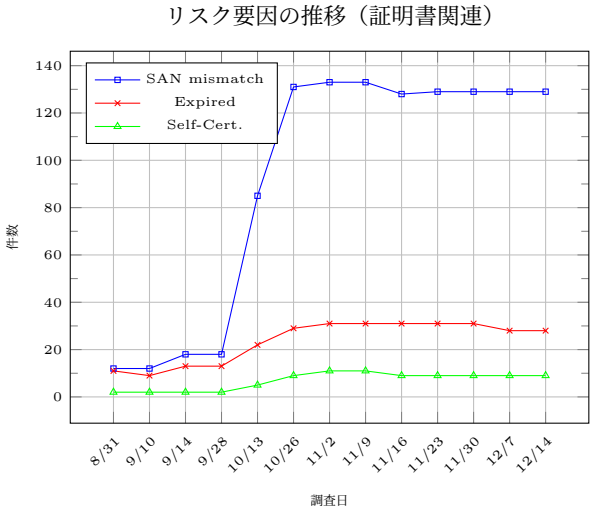


**図 2** 主要なリスク要因の時系列変化

SAN の不一致（SAN mismatch）では, 観測開始時（8 月 31 日）の 12 件から, 11 月 2 日時点には 133 件へと約 11 倍に急増した. 特に 10 月 13 日の観測時に 18 件から 85 件へと急増した. この増加は, cpcalendars や cpcontacts といった特定の接頭辞を持つサブドメイン群が大量に出現したことが原因であった. また, 8 月 31 日から 11 月 2 日において, 証明書有効期限切れ（Expired）は 11 件から 31 件

となり, 自己署名証明書（Self-Cert.）に関しては, 2 件から 11 件へ増加した.

#### 3.3.2 その他のリスク要因

証明書以外のリスク要因では, 減少傾向と恒常的な傾向の二つの傾向が観察された.

リスク要因「SSTI」, 「Certificate without SAN」, 「Badly configured localhost records」, 「WWW-Authenticate Uses Unsafe Authentication」の 4 つは観測期間において多少の増減はあったが, 恒常的な傾向で, かつ件数も 3 件以下であり少なかった.

「DMARC」関連と「cookie の secure 属性」のリスク要因は減少傾向であった. 特に「DMARC」関連のリスク要因は, 11 月 16 日以降の観測では, 観測されなくなった.

検出件数が恒常傾向であったリスク要因と減少傾向であったリスク要因の動向を, それぞれ以下図 3, 図 4 に示す.

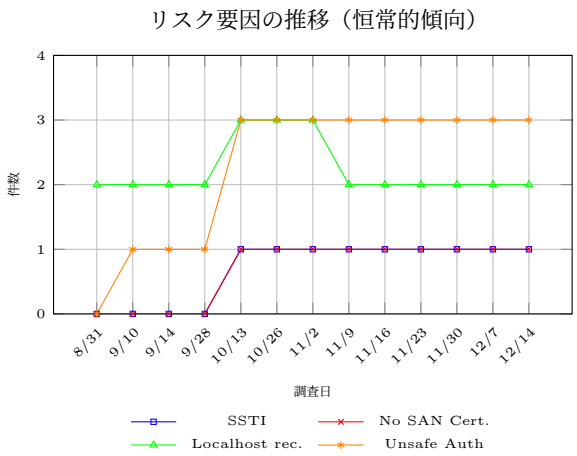


**図 3** リスク要因の時系列変化（恒常的傾向）

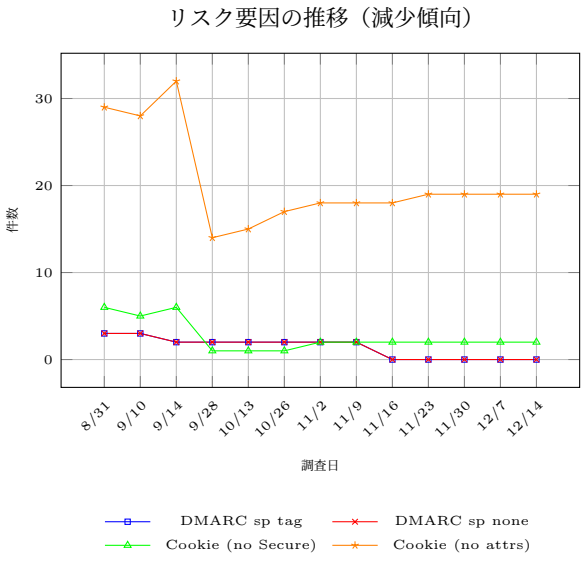


**図 4** リスク要因の時系列変化（減少傾向）

### 3.4 分析のまとめ

以上の結果から, 組織内において管理不備のある証明書を保持するサブドメインが, 時間経過とともに増加・蓄積し続けている現状が浮き彫りとなった. 特に 10 月中旬に見られた SAN の不一致や期限切れ証明書の爆発的な増加は, 管理者の把握が及んでいない資産や, 共通の設定不備を持つサービス群の存在を強く示唆している.

本研究ではリスク要因の増加が顕著であり, 他のリスク要因と比較して最も件数が多かった「証明書に問題があるドメイン」に注目して分類を行うこととした. 次章では, 証明書に問題があるドメインに関する自動的な分類に関して, 分類の基準や自動化の詳細を説明する.

## 4. 証明書に問題があるドメインの自動的分類

本章では, 膨大な IT 資産に対応する「規模」と情報の変化に即座に対応する「リアルタイム性」を両立するため, python および LLM (Gemini API) を用いたドメインの自動的分類の手法を示す.

### 4.1 分類と自動化の必要性

#### 4.1.1 分類の必要性

現状で使用した ASM ツールにおいては, 同一のリスク要因（例：Certificate expired）に対しては一律に同じセベリティスコアが付与される仕様となっている. しかし実際の運用においては, 同一のリスク内容であっても資産の運用状況や背景により, リスクの重要度は異なる. これは, 証明書不備のリスク要因に対しても同様である.

証明書に問題があるドメインにおいて最も深刻な懸念は, 通信が安全に行われない危険性があい, 利用者の情報が盗聴や改ざんの危険にさらされることである. しかし, 検出された膨大なドメイン群に対して, 一律のリスク評価を下すことは適切ではない. 同じ「期限切れ」であっても, 個人情報を扱う入力フォームが存在するサイトと, 既に公開を停止している古いテストページやコンテンツのないサイトでは, リスクの重大度や対処方法が大きく異なってくる.

#### 4.1.2 自動化の必要性

ASM ツールで検知された資産に対して, 資産の運用状況やリスクの重要度を目視で確認することは, 非効率であり, かつ人的リソースの観点から事実上不可能である. 今回の調査対象は一つの大学組織に限定されており, 実際に調査した IT 資産（ドメイン）数は 129 件であった. この規模であっても, 実際に一つひとつのドメインに対してブラウザでのアクセス確認や証明書情報の詳細確認を行い, 分類作業を行うには多大な時間を要した. 一般企業の IT 資産, あるいはより大規模な組織体においては, 管理すべき膨大

な資産に対し, 手動で調査を行うのは困難である.

また, ウェブサイトの構成や証明書の設定状況は常に変化する. たがって, 手動調査では調査結果を優先度に反映させている間に, すでに実際の状況が変化している可能性がある. 変化を即座に捉えて優先順位を更新し続けなければならないが, 手動調査では防御側に求められるリアルタイム性を担保できない.

### 4.2 分類の基準

各ドメインの「実際のサイト状況」に基づいた真のリスクを評価するため, サイトへアクセスした際の画面状況から, ユーザーが受ける印象や利用可能な機能に基づき 4 つの区分に分類した. 証明書不備における最大の問題は「通信内容の安全性を担保できない」点にあるが, サイトの用途によってその被害の深刻度や管理者が取るべき対応方針が異なるため, 以下の基準を設けた.

また, 以下の定義を簡潔にまとめたものを表 3 に示す.

A) ユーザがログイン機能を使用できる
- 基準：
  ログイン画面や, 認証機能画面が存在し, ID/PWD などをユーザが入力可能である.
- 分類の意図：
  認証情報の送受信が発生するため, 通信の傍受や中間者攻撃による被害が最も直接的かつ深刻である.
- 対応方針
  最優先かつ早急な対応を要する. 管理者は直ちに, 証明書の更新をしなければならない. 更新が完了するまでサービスを一時停止させる等の強い措置も検討対象となる.

B) ユーザーが稼働中と判断する
- 基準：
  ログイン機能はないが, コンテンツが正常に提供されており, 概ね 2021 年以降の更新履歴や最新のお知らせ等が確認できるもの.
- 分類の意図：
  アクティブな公開サイトであり, 組織の信頼性に直結する. 警告画面が表示されることで利用者に不安を与え, 組織のブランドイメージを損なうリスクがある.
- 対応方針
  すみやかに管理者に連絡し, 証明書の更新を促す.

C) ユーザーが稼働中でないと判断する
- 基準：
  テスト用ページ, Web サーバのデフォルト画面, 数年間更新が途絶えている放置された資産. （2021 年以降の更新がない）
- 分類の意図：
  管理者が把握していない忘れられた資産や放置された資産などが該当する. 攻撃者にとっての足掛かり（踏み台）になる危険性がある一方, サービスとしての重要度は低い.
- 対応方針
  管理者に対し, 証明書の更新, あるいは不要なサイトの削除・閉鎖を強く促す.

D) その他
- 基準：
  他のページへのリダイレクトが設定されているものや, 上記 A〜C のいずれにも明確に分類できない特殊な挙動を示すもの.
- 分類の意図：
  評価対象から除外するか, あるいは別途詳細な調査が必要なドメインとして切り分ける.
- 対応方針
  状況に応じた個別判断を行う.

**表 3** ユーザー視点に基づくサイト状況の分類定義

| 区分 | 定義・基準 |
|---|---|
| A: ログイン機能 | 認証機能が存在し, ID/パスワード等を入力できるもの |
| B: 稼働中 | コンテンツが提供されており, 2021 年以降の更新があるもの |
| C: 非稼働 | テスト・デフォルトページ, または 2021 年以降更新がないもの |
| D: その他 | リダイレクト設定, または上記に分類不能なもの |

### 4.3 自動化の手順

本節では, 本研究で提案する自動分類システムの全体的なアルゴリズムについて述べる.

以下に処理全体の流れを示す.

(1) サイトの運用状況の取得
  証明書不備を有するドメインに関して, サイトの運用状況を判断するため, curl コマンドから得られる HTTP レスポンスのヘッダおよびボディテキス部分を取得する.

(2) HTTP レスポンスヘッダ情報等を投入
  ステップ 1 で得た, curl コマンドによる情報を Python コードのユーザプロンプト部分に渡す.

(3) Gemini による分類
  Gemini は, ヘッダ情報や HTML 情報等からサイトの画面状況やユーザ視点での稼働状況を推測し, システムプロンプト部分で指定された分類基準によって分類を行う.

### 4.4 実行結果

本節では, 12 月 15 日時点の固定データに含まれるドメイン 115 件に対し, 自動分類を実行した結果について, 手動調査の結果を正解データ（基準）とした評価結果を述べる. なお, 12 月 15 日時点の固定データ 115 件は, 全て証明書に問題が存在するドメインである.

また, 生成 AI による判定の再現性を確認するため, 10 回ずつの反復判定を実施した. ただし, 同一の HTML コンテンツ（curl 実行結果）を持つドメインに対しては, 代表的なサンプルに対して試行を行い, 効率的な検証を行った. 例としては, 多数存在する特定のウェブホスティングサービスに関連するサブドメインのように内容が完全に同一であるものについては, 同一の内容をもつドメインの判定回数の合計値が 10 回に達したところで再現性の確認が取れたとみなした.

#### 4.4.1 分類精度

反復判定を含む合計 210 回の試行結果に基づき, 各区分における分類精度を表 4 に示す. なお, 以下の表における「ドメイン数」は, 同一の HTML コンテンツ（curl 実行結果）を持つドメインは 1 つとして数えている.

手動による調査との一致率は全体平均として 99.5 %であった.

**表 4** 自動分類におけるクラス別精度

| 区分 | ドメイン数 (件) | 一致率 (%) | 判定回数 (正解 / 合計) |
|---|---|---|---|
| 分類 A | 5 | 100.0% | 50 / 50 |
| 分類 B | 6 | 100.0% | 60 / 60 |
| 分類 C | 7 | 98.6% | 69 / 70 |
| 分類 D | 3 | 100.0% | 30 / 30 |
| **全体平均** | | **99.5%** | **209 / 210** |

#### 4.4.2 C の誤判定に関して

「ユーザが稼働中でないと判断する（C）」においては, 手動による調査と比較すると一致率は 98.6 %であった.

C に分類されるドメインは全部で 13 件であった. このうち, 同一 HTML コンテンツをもつテスト用ドメイン群が 7 件存在し, これらを 1 つと数えて 7 件を自動分類した.

誤判定が発生した原因は, サイトの「コンテンツの有無」に対する LLM の解釈に揺らぎがあったためである. 当該ドメインは, アクセス時に「It Works!」とだけ表示されるデフォルトページであり, 定義上は区分 C（非稼働）に該当する. 実際に, 10 回の試行のうち 9 回は正しく区分 C と判定された.

しかし, 誤判定が発生した 1 回においては, LLM が「It Works!」という文字列を単なる初期設定画面としてではなく, ユーザーに対して「サーバーが正常に稼働し, 意図したメッセージを伝達している」という能動的なコンテンツ提供であると推論した. その結果, ユーザーが「このサイトは意図を持って運用されている（稼働中）」と受け取ると解釈し, 区分 B へ分類された.

以下に実際に誤判定された際の出力結果を表示する.

**表 5** Gemini API による分類と手動調査が一致しなかった事例

**対象:** example.kyushu-u.ac.jp

**Gemini 判定:** クラス B

**判定根拠:**

HTTP/1.1 200 OK で応答し、HTML に`<h1>It works!</h1>`が含まれる。ユーザー視点では、このメッセージによりサーバーが正常稼働中と判断できるため「稼働中」に該当する。定義上、デフォルトページはクラス C に含まれるが、ユーザーは「稼働しているがコンテンツ未設定」と解釈するのが自然であるため B とした。

### 4.5 自動分類の有効性と実用性の評価

本節では, 第 4.3 節で示した実行結果に基づき, 本研究で提案した生成 AI による自動分類手法の有効性について, 精度, リスク管理の最適化, および運用効率の観点から評価を行う.

#### 4.5.1 分類精度の信頼性

手動調査の結果を正解データとした場合の一致率は全体平均で 99.5%という極めて高い数値を示した. 特筆すべきは, セキュリティ上の重要度が最も高い「区分 A：ユーザがログイン機能を使用できる」において 100%の精度を達成した点である. これは, 本システムが認証情報を扱う高リスクなサイトを確実かつ正確に特定できることを示しており, 実運用における信頼性は非常に高い.

また, 唯一の不一致事例となった「It works!」の判定についても, 当該ページを単なる技術的なデフォルトページとして切り捨てるのではなく, ユーザ視点を優先した解釈を行った. これは, 単純なキーワードマッチングではなく, 柔軟な文脈理解に基づいてサイト状況を把握できていることを示しており, 主観的な基準に基づく分類手法としての妥当性を裏付けている.

#### 4.5.2 運用背景によるリスク評価の最適化

表 6 のクロス集計結果は, ASM ツールが提供する技術的データに「運用背景」を加えることのリスクの重要性, 対応優先度を明確に示している.

例えば, 本調査で最も多く検出された「SAN の不一致（不一致のみ）」の 97 件のうち, 83 件が早急に対処すべき「区分 A」に属している一方で, 6 件はサービスとしての A と比較して重要度が低い「区分 C」に分類されている. 本システムを用いることで, 管理者は「ログイン機能を持つアクティブなサイト」にリソースを集中させることが可能となる. このように, 技術的なリスク要因の情報とサイト

の運用実態を組み合わせることで, 資産に対する対応の優先順位付けといものが最適化が可能であることがわかった.

**表 6** リスク要因と自動分類結果のクロス集計

| リスク要因の組み合わせ | A | B | C | D | 合計 |
|---|---|---|---|---|---|
| 期限切れ + 不一致 + 自己署名 | 0 | 0 | 7 | 0 | 7 |
| 期限切れ + 自己署名 | 2 | 0 | 0 | 0 | 2 |
| 期限切れ + 不一致 | 4 | 0 | 0 | 0 | 4 |
| 自己署名 + 不一致 | 0 | 0 | 0 | 0 | 0 |
| 期限切れ（のみ） | 4 | 0 | 0 | 1 | 5 |
| 自己署名（のみ） | 0 | 0 | 0 | 0 | 0 |
| 不一致（のみ） | 83 | 6 | 6 | 2 | 97 |
| 計 | 93 | 6 | 13 | 3 | 115 |

#### 4.5.3 運用効率の向上

手動調査には「調査規模の限界」と「リアルタイム性の欠如」という大きな課題が存在していた．手動調査ではASM ツールによって検出されたドメインの分類に対して多大な時間を要し, かつ情報の変化に即座に対応することが困難であった.

これに対し, 本手法ではドメインの分類をプログラムによって短時間で完了させている．これにより, 大規模な IT 資産を抱える組織においても, 管理者の負担を最小限に抑えつつ, 継続的な監視が期待できる.

## 5. 結論

本研究では，ASM ツールが検出する膨大なリスク要因に対し，大規模言語モデル（LLM）である Gemini API を活用してサイトの運用背景を解析し，リスクの重要度に基づき自動分類する手法を提案した.

近年の IT 資産拡大に伴い，継続的な監視の重要性が増しているが，既存ツールでは技術的不備に対して一律の深刻度が割り振られるため，資産の運用状況を考慮した重要度の把握が困難であった．また，管理者が目視で行う手動調査には多大な人的コストを要し，リアルタイム性の確保にも限界があった.

本研究で構築した自動分類システムを検証した結果，手動調査を正解とした場合の一致率は全体平均で 99.5%という極めて高い精度を示した．特に，認証情報の漏洩リスクがあり最も重要度が高い「区分 A：ユーザがログイン機能を使用できる」においては 100%の精度で特定が可能であり，実運用における信頼性の高さが実証された．以上の結果から，LLM を用いたコンテクスト解析は，技術的スキャンの迅速性を維持しつつ，人間のような柔軟な文脈理解に基づいたリスクの自動選別を実現する極めて有効な手法であるといえる.

今後の課題として，まずは判定の再現性を確保するために，人間の介在の必要性が挙げられる．人間が固定的な分類定義を決定し，LLM の推論を規定の枠組み内に誘導しなければならない．次に，今回の実験で確認された「デフォルトページ」に対する LLM 独自の推論など，主観的基準と技術的実態の整合性を高めるためのシステムプロンプトの改善が求められる.

さらに，分析対象を証明書以外の脆弱性へ拡大することや，一般企業や大規模組織における有効性の検証も重要な課題である．加えて，JavaScript 実行後の画面変化を正確に捉えるなど，SPA を含む動的コンテンツへの対応力を強化することで，実用性をさらに高めていくことが期待される.

### 参考文献

[1] Cloudflare: 大規模言語モデル (LLM) とは. (参照: 2026-01-26).
[2] CyCraft Technology: 2024 Taiwan Cybersecurity Exposure Inventory, White paper, CyCraft Technology and National Institute of Cyber Security (NICS) (2024). Accessed: 2026-01-24.
[3] Gelernter, N., Schulmann, H. and Waidner, M.: External Attack-Surface of Modern Organizations, *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, ASIA CCS '24, Association for Computing Machinery, p. 589–604 (online), DOI: 10.1145/3634737.3656295 (2024).
[4] IBM: アタックサーフェス・マネジメント（ASM），`https://www.ibm.com/jp-ja/think/topics/attack-surface-management`. n.d. 参照日: 2025-10-29.
[5] NEC ソリューションイノベータ: LLM（大規模言語モデル）とは？生成 AI との違いや仕組みを解説 (2024). 更新: 2024-07-09, (参照: 2026-01-26).
[6] NTT ドコモビジネス: 大規模言語モデル（LLM）とは？意味・定義 — IT 用語集. (参照: 2026-01-26).
[7] Palo Alto Networks: What is External Attack Surface Management (EASM)?, `https://www.paloaltonetworks.com/cyberpedia/easm-external-attack-surface-management`. n.d. 参照日: 2025-10-29.
[8] Stryker, C.: What are large language models (LLM)? (参照: 2026-01-26).
[9] Tariq, S., Baruwal Chhetri, M., Nepal, S. and Paris, C.: Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities, *ACM Comput. Surv.*, Vol. 57, No. 9 (online), DOI: 10.1145/3723158 (2025).
[10] Trend Micro: 外部アタックサーフェス管理（EASM）とは，`https://www.trendmicro.com/ja_jp/what-is/attack-surface/external-attack-surface-management.html`. n.d. 参照日: 2025-10-29.