IMES DISCUSSION PAPER SERIES

ディープフェイク検知モデルの評価・比較：研究事例を活用する際の留意点と課題

宇根正志（うねまさし）

Discussion Paper No. 2025-J-9

IMES

INSTITUTE FOR MONETARY AND ECONOMIC STUDIES

BANK OF JAPAN

日本銀行金融研究所

〒103-8660 東京都中央区日本橋本石町 2-1-1

IMES Discussion Paper Series 2025-J-9
2025 年 11 月

ディープフェイク検知モデルの評価・比較：研究事例を活用する際の留意点と課題

宇根正志（うねまさし）*

要　　旨

スマートフォンを用いた金融取引では、顧客の本人確認手段として、顔の動画や静止画による生体認証が採用されるケースがある。こうした生体認証への脅威として、偽の動画を提示してなりすましを試みる攻撃が想定される。最近では、AI・機械学習によるディープフェイクを用いた攻撃の可能性を示唆する研究成果が複数発表され、現実的な脅威として認識する必要性が高まっている。対策としては、機械学習モデルによってディープフェイクを検知する手法の研究が活発化しており、複数の検知モデルを横並びで評価・比較した研究成果も発表されはじめている。こうした動向を踏まえると、今後、生体認証を採用している金融機関は、ディープフェイクによるリスクを評価し、必要に応じて対策を検討することが必要となるであろう。検知モデルの採用を検討する際には、想定されるディープフェイクや検知モデルの評価基準をまず設定し、公開されている評価・比較の研究事例を参照しながら、評価基準と合致したものを選択することが望ましい。また、ディープフェイクに関する技術の進歩が非常に速く、対策の有効性やリスクが時間とともに変化するため、技術動向のフォローやリスクの再評価を継続的に行うことも重要である。

キーワード：機械学習、生体認証、セキュリティ、ディープフェイク、
　　リスク管理、AI
JEL classification: G21、O33

*　日本銀行金融研究所参事役（E-mail: masashi.une@boj.or.jp）

本稿は 2025 年 8 月 29 日時点の情報に基づいている。本稿の作成に当たっては、市野将嗣准教授（電気通信大学）から有益なコメントを頂戴した。ここに記して感謝したい。ただし、本稿に示されている意見は、筆者個人に属し、日本銀行の公式見解を示すものではない。また、ありうべき誤りはすべて筆者個人に属する。

目　次

１．はじめに

近年、AI や機械学習の研究が急速に進展し、実際には行われていない人間の動作をあたかも行われていたかのようにみせる動画や音声を容易に生成できるようになっている。こうした技術や生成されたメディアは、深層学習を含む機械学習に基づいて開発されてきたことから、ディープフェイクと呼ばれている。例えば、自分の髪型や洋服を選ぶ際に、自分の動画や静止画に特定の髪型や洋服を合成してフィット感を確かめるといったユースケースがよく知られている。また、与えられたテキストを外国語に翻訳し、それに対応する音声データを生成・出力する自動翻訳も代表例である。最近では、テキストから画像や音声を生成できる生成 AI によるディープフェイクも注目を集めている（笹原［2023］、Gaur [2023]、坂本・宇根［2025］）。

情報セキュリティの観点では、ディープフェイクは、金融サービスにおいてなりすましに悪用される可能性がある。例えば、モバイル・バンキングの使用時にスマートフォン内蔵のカメラで撮影した顔の動画によって本人確認（生体認証）を行う場合、攻撃者が他人の顔の動画を生成し、それをカメラに提示するなどしてなりすましを試みる攻撃が想定される。こうした攻撃は、金融機関の顧客の個人情報が盗まれたり不正な取引が実行されたりするリスクにつながる。金融庁は、AI ディスカッションペーパー（金融庁［2025］）のなかで、特定の人物になりすました偽の動画や音声による詐欺行為などがディープフェイクによって容易になる可能性を指摘するとともに、新たなリスクとして留意することが重要であるとしている[1]。

生体認証の文脈では、人工物をセンサーに提示してなりすましを試みる攻撃シナリオが既によく知られており、さまざまな対策手法が研究されてきた（宇根［2016, 2024］、Khan and Kahn [2025]）。例えば、顔の動画を用いた認証の場合、カメラで取得した動画を分析し、血流による肌の色合いの変化（Li *et al.* [2016]）、顔の表面の微細な凹凸の有無（Xu, Li and Deng [2015]）、目や鼻などの特徴点の三次元構造の有無（Wang *et al.* [2019]）などを手掛りに、対象が人間か否かを判定する対策手法が挙げられる。

こうしたアプローチに加え、顔の動画が与えられたときにそれがディープ

[1] 海外のセキュリティ当局もディープフェイクの脅威に注意を促している。アメリカの国家安全保障局・連邦捜査局・サイバーセキュリティ社会基盤保障庁は、ディープフェイクによって、重要人物へのなりすまし、重要な情報への不正アクセスなどが発生しうるとしたうえで、対策の必要性を指摘している（National Security Agency, Federal Bureau of Investigation, and Cybersecurity and Infrastructure Security Agency [2023]）。また、アメリカの金融犯罪取締ネットワークは、金融機関の顧客の動画などを生成 AI によって偽造し、他人になりすまして口座開設を試みたりフィッシングや詐欺を実行したりするおそれがあるとの注意喚起を行っている（Financial Crimes Enforcement Network [2024]）。

フェイクか否かを機械学習の技術によって判定・出力するアルゴリズム（ディープフェイク検知手法）の研究が近年活発化している。ディープフェイク検知手法に訓練用のデータを適用して検知モデルを生成し、それを生体認証のシステムに組み込んで活用することが考えられる。最近では、大規模言語モデル（LLM: Large Language Model）を活用する手法も相次いで提案されている（Deng *et al.* [2024]、Lin *et al.* [2025]、Sony *et al.* [2025]）。

顔の動画や音声による生体認証を顧客の本人確認の手段として採用している金融機関は、金融サービスのセキュリティを維持していくうえで、ディープフェイクによるなりすましのリスクを適切に評価する必要がある。そのうえで、リスクが許容レベルを超えていた場合、リスクを軽減する方法を検討することが求められる。リスク評価や対策手段の検討を適切に行うためには、ディープフェイクに関する技術・実装の動向をタイムリーにフォローすることが重要である。

本稿では、顔の動画のディープフェイクに焦点を当てて、検知モデルを評価・比較した最近の研究事例を紹介し、検知モデルを選択するために研究事例を活用する際の留意点や課題を考察する。2 節では、顔の動画のディープフェイクの生成手法や、ディープフェイクによるなりすましの可能性を検証する実験の事例を紹介する。3 節では、検知モデルを評価・比較した最近の研究事例を紹介する。4 節では、ユーザの観点から研究事例を活用する際の準備や留意点を考察するとともに、研究コミュニティにおける課題についても考察する。

２．顔の動画を対象とするディープフェイクと実験研究

本節では、人間の顔の動画を対象とするディープフェイクの主な生成手法と、生成されたディープフェイクによるなりすましの可能性を示唆する実験研究を紹介する。

（１）主な生成手法の類型

ディープフェイクの生成手法は、主に次の 3 つに分類されることが多い（例えば、Le *et al.* [2025]、Tariq *et al.* [2025]）[2]。

- なりすましの対象者の顔の静止画を準備し、攻撃者自身の顔の動きに合わせてその静止画を変化させて動画を生成する（リエナクトメント）。
- なりすましの対象者の顔の静止画の一部に、攻撃者の顔の動画の一部[3]を切り取って埋め込むことによって動画を生成する（リプレイスメントま

[2] 個々の手法に関しては、宇根［2024］において紹介されているので参照されたい。

[3] 切取りと埋込みの対象となるのは、一般に、目、鼻、口、耳などの特徴的な部位とそれらの周辺領域である。

たはフェイス・スワップ)。

- なりすまし対象者の特徴を示すテキストなどを生成 AI モデルに入力してなりすまし対象者の動画を生成する(シンセティック・フェイス・ジェネレーション)。

（2）ディープフェイクによるなりすましの実験

顔の静止画や動画による生体認証システムにおいて、ディープフェイクによるなりすましの実現可能性を検証した 3 件の実験の概要を説明する。これらの実験結果を記述するペーパーは、実験の詳細をすべて明記しているわけではないものの、ディープフェイクを人間の顔の静止画や動画と誤って判定する事象を報告している。

イ．独自に構築したシステムにおける実験

（イ）川名らによる実験

川名らは、スマートフォン内蔵のカメラで顔の動画を取得し、別途撮影した運転免許証の静止画と照合する生体認証システムを構築した(川名ほか［2021］)。そのうえで、オープンソースのツールを用いて生成したディープフェイクが誤って受け入れられるか否かを検証した。

実験では、なりすまし対象者の運転免許証を撮影して顔の静止画を取得し、生体認証システムへ送信する。次に、攻撃者の顔の動画をカメラで取得し、それをディープフェイク生成ツールに入力してディープフェイク（なりすまし対象者の顔の静止画に攻撃者の動画の動きを反映させたもの）を生成する。ディープフェイクをディスプレイ上に表示し、それをスマートフォンのカメラで撮影して取得した動画を生体認証システムへ送信する。生体認証システムは、運転免許証の静止画とディープフェイクの動画を照合し、一致するか否かを判定する。

実験は 1 人の実験協力者（なりすまし対象者）に対して行われた。その結果、実験協力者に対するディープフェイクがその静止画と一致すると誤って判定された[4]。

（ロ）Šalko らによる実験

Šalko らは、顔の動画や静止画による生体認証システムを構築し、公開されて

[4] 川名らは、後続の研究として、この生体認証システムにディープフェイクの検知モデル（4 件）を追加し、ディープフェイクを検知できるか否かも検証している（川名ほか［2024］)。検証の結果、リプレイスメントに基づくディープフェイクに対して、ほとんどの検知モデルにおいて 50%以上の確率で検知することができたが、リエナクトメントに基づくディープフェイクに対しては、50%以上の確率で検知できたものは 1 つだけであった。

いる顔の動画のデータセットを用いてディープフェイクによるなりすましの可能性を検証した（Šalko, Firc, and Malinka [2024]）。

実験では、データセット（人間の顔の動画とディープフェイクを格納）から、なるべく正面を向いた個人の顔の動画と、その個人に対応するディープフェイクをそれぞれ選択したうえで、両者を生体認証システムに入力して一致するか否かを判定させるというものである。

実験の結果、一致と誤って判定する確率が、異なる個人の顔の動画を照合した際に一致と誤って判定する確率よりも高いことが判明した。これを踏まえ、Šalkoらは、ディープフェイクを考慮せずに生体認証システムの判定しきい値を設定した場合、ディープフェイクを誤って受け入れる確率が高くなる可能性があると考察している。

ロ．クラウドで提供されている生体認証サービスにおける実験

Liらは、クラウドで提供されている顔の動画による生体認証のサービス（6件）を対象に、ディープフェイクが提示された際にどの受け入れられるかを実験した（Li *et al.* [2022]）。

対象となった生体認証サービスでは、いずれも、各ユーザが認証に用いる顔の動画をスマートフォン内蔵のカメラで撮影し、撮影したデータをクラウド上のサーバに送信する。サーバは、受信した動画を事前に登録されたものと照合し、同一の個人か否かを判定する。顔の動画の形態はサービスによって異なっており、提示された動画が人間を撮影したものであることを確認するために、顔の向きをランダムに変化させるケースや、ランダムに示した数字を発音させるケースが準備されていた。

実験では、事前に登録された顔の動画を模倣するディープフェイクを生成し、サーバに送信して判定させた。その結果、顔の向きをランダムに変化させた動画の場合、誤って一致と判定した確率が約 80%となったディープフェイクが存在した。また、ランダムに示した数字を発音させた動画の場合、誤って一致と判定した確率が約 60%となったディープフェイクが存在した。

この結果は、実運用されている生体認証サービスにおいても、ディープフェイクによるなりすましに対して脆弱なケースがあることを示している。

３．ディープフェイク検知手法の類型と検知モデルの評価・比較

生体認証システムにおけるディープフェイク対策としては、入力データ（顔の動画や静止画）を登録済みのものと照合する処理に加えて、入力データが「人間の顔を撮影したものか否か」を判定する処理を準備するケースが多い。後者の処

理として、生体検知（liveness detection）[5]の手法が広く活用されてきたが、近年では、ディープフェイクの生成手法の進展に伴ってディープフェイクか否かを機械学習モデルによって判定する手法が数多く提案されている。

そこで、以下では、機械学習の技術を用いた手法に焦点を当てて、それらの類型や検知モデルの評価・比較研究の事例を紹介する。

（１）主な検知手法の類型

ディープフェイク検知手法の類型として、まず、最近注目を集めている LLM を用いる手法と用いない手法に大別できる。そのうえで、LLM を用いない手法については、動画を構成する個々のフレーム内の情報に着目する手法、個々のフレーム間の関係性に着目する手法、電子透かしを用いる手法に分けることができる（Zou *et al.* [2025]）。

イ．LLM を用いる手法

この類型は、LLM を用いてディープフェイクか否かを判定する手法である。これらの手法では、テキストに加えて動画も入力することができるマルチモーダル型の LLM（MLLM: multi-modal LLM）[6]が用いられる（Komaty *et al.* [2025]、Song *et al.* [2025]）。MLLM では、一般に、動画とともにその処理方法がテキスト（プロンプト）として入力され、処理結果がテキストとして出力される。ディープフェイク検知の文脈では、判定したい動画と「動画がディープフェイクか否かを判定せよ」との主旨のプロンプトが MLLM に入力され、判定結果が出力される（Gani *et al.* [2025]）。また、プロンプトに「判定の理由を回答せよ」と付け加えることによって、判定の理由も出力させる手法もある（Zou *et al.* [2025]）。ただし、本稿執筆時点では、顔の動画のディープフェイクに焦点を当てて複数の MLLM を評価・比較した研究成果は、筆者が知る限り公開されていないようである[7]。

---

[5] 生体検知は、生体認証システムにおけるセンサーの対象が人間か否かを判定することを目的とする技術であり、主に、センサーに人工物を提示してなりすましを試みる攻撃（presentation attack）への対策として研究されてきた。生体検知はディープフェイクを提示する攻撃に対しても対策となりうるが、2 節（2）ロで紹介した Li らの実験で示されているとおり、生体検知のみではディープフェイクを十分に検知できないケースも想定される。対応として、例えば、ディープフェイク検知モデルと組み合わせるなどが考えられる。

[6] LMM（Large Multi-modal Model）と呼ばれることもある。

[7] 顔の静止画のディープフェイクに関する評価・比較の論文については、Narayan, VS, and Patel [2025]など、いくつか発表されている。

ロ．LLM を用いない手法

（イ）フレーム内の情報に着目する手法

この類型は、動画の各フレームに含まれる情報を手掛りにそれぞれのフレームごとにディープフェイクか否かを判定した後、最後に、各フレームの判定結果を統合して動画全体としてディープフェイクか否かを判定する手法である。フレームごとの判定において、例えば、顔の不自然な色や影、顔の形状の歪み、顔の特徴点（目、鼻、口、耳など）の位置関係の不整合性、頭部の三次元構造の不整合性の有無を検証する手法が知られている（例えば、Bai, Lin, and Cao [2024]）。

（ロ）フレーム間の関係性に着目する手法

この類型は、動画のフレームに含まれる情報の時系列的な推移などの関係性を手掛りとしてディープフェイクか否かを判定する手法である。例えば、顔の形状、特徴点の位置、色などがフレーム間で整合的か否かを検証する手法が挙げられる（He *et al.* [2024]、Yan *et al.* [2024]）。

（ハ）電子透かしを用いる手法

この類型は、動画を作成する際に特定のデータ（電子透かし）を動画の各フレームに埋め込み、ディープフェイクと区別できるようにする手法である。ディープフェイクを検知する際には、対象となっている動画から電子透かしを抽出できるか否かを検証する。電子透かしを抽出できた場合、その動画を「ディープフェイクでない」と判定する（Luo *et al.* [2023]、Zhang *et al.* [2023]）。

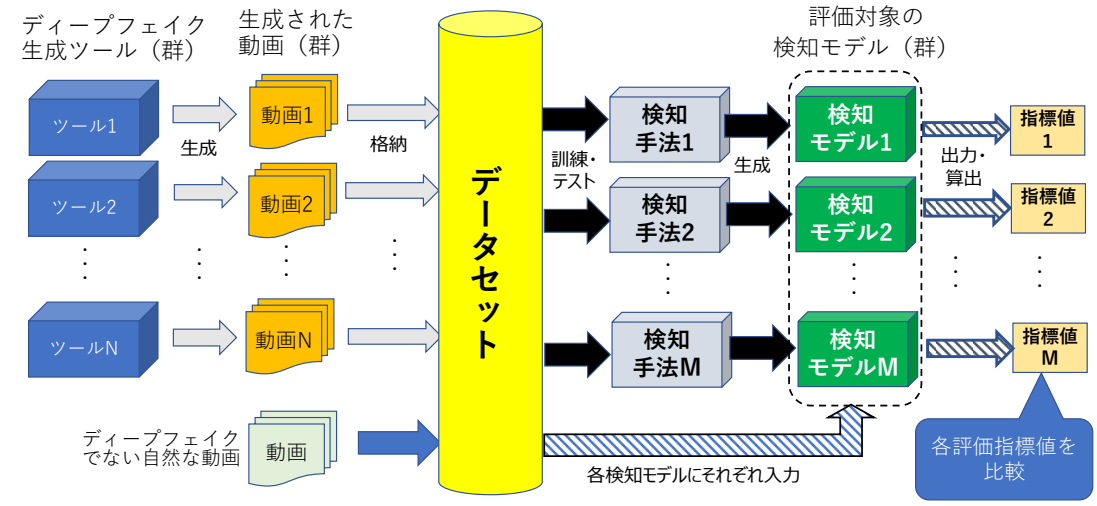
（２）検知モデルの評価・比較の方法

本節 (1) で紹介した各検知手法のうち、電子透かしを用いる手法については、複数の手法を対象に評価・比較した研究成果が少ないことから対象外とする。検知モデルの評価・比較は一般的には次の流れで実施される（図 1 を参照）。

① データセットを準備する。
- データセットには、ディープフェイクと現実のデータの両方が含まれる。ディープフェイクについては、例えば、既存のディープフェイクの生成ツールを用いてディープフェイクを生成したり、インターネット上に存在するディープフェイクを収集したりする。

② 評価対象となる検知モデル（複数）を生成する。
- 検知モデルの生成に用いるデータ（訓練用データやテスト用データ）は、データセットから抽出する。検知モデルが生成済みの場合、改めてモデルを生成するのではなく、既存のものを使用するケースがある。

図 1 ディープフェイク検知モデルの評価・比較の流れ



③ 評価指標値を算出する。

- 各検知モデルに評価用データ（データセットの一部）を入力し、判定結果を得るとともに、評価指標値を算出する。生体認証の観点では、ディープフェイクをなるべく見逃さないことが重要であるため、「ディープフェイクが入力されたときにそれを正しく判定した割合」（再現率）が高いことが望ましい。また、アプリケーションの可用性の観点からは、「現実のデータが入力されたときに誤ってディープフェイクと判定した割合」（偽陽性率）が低いことが望ましい。そこで、再現率と偽陽性率の関係を表す ROC（Receiver Operating Characteristic Curve、図 2 参照）や AUC（Area Under the ROC Curve、図 2 参照）[8]が評価指標として採用される場合がある。
- 偽陽性率の代わりに、「ディープフェイクと判定されたケースのうち、その判定が正しかったものの割合」（適合率）を採用し、適合率と再現率が同時に高くなるケースが望ましいとする見方もある。この場合、適合率と再現率との関係を表す F 値（F score、図 3 参照）[9]が評価指標として用いられる。
- このほか、正解率[10]を評価指標とするケースもある。

---

[8] X 軸に偽陽性率をとり、Y 軸に再現率をとって、両者の関係をグラフで表したときの曲線が ROC である。また、ROC の積分値が AUC である。AUC の値は 0 から 1 の間の値をとり、値が大きいほど判定精度が高いと評価される。

[9] F 値は、適合率と再現率の調和平均として定義され、0 から 1 の間の値をとる。適合率と再現率が同時に高い値を示すと F 値も大きな値となり、判定精度が高いと評価される。

[10] 正解率は全試行に対する正しい回答（ディープフェイクが入力された場合にはディープフェ

図 2 ROC と AUC

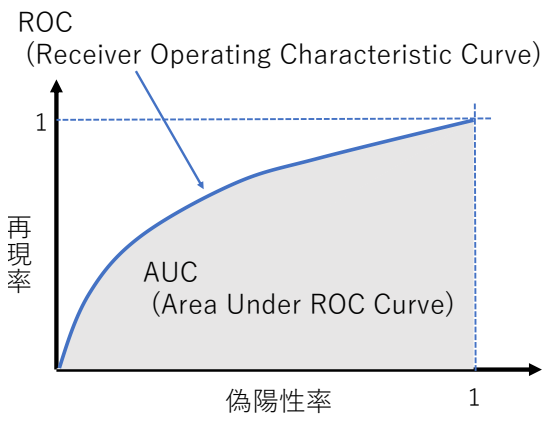
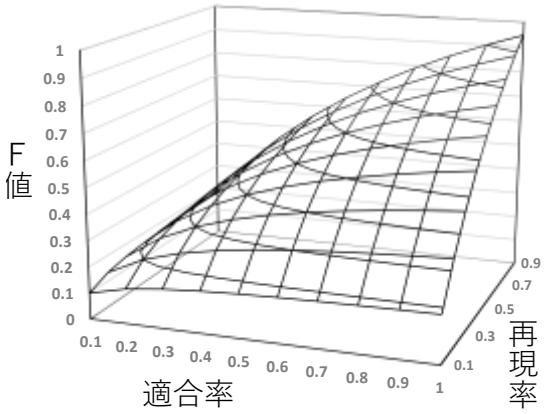


図 3　適合率・再現率と F 値



④　検知モデル間で評価指標値を比較する。

（３）主な研究事例

以下では、顔の動画や静止画のディープフェイクに対する検知モデルを評価・比較した最近の主な研究事例 4 件（LLM を用いないもの 3 件、LLM を用いるもの 1 件）を紹介し、最後に結果を概観する（表 1 を参照）。ここでは、第三者による評価・比較の論文に絞り、特定の手法の提案やその優位性を主張する論文を対象外とした。また、MLLM の評価・比較については、論文執筆時点において顔の動画のディープフェイクを対象とするものが見当たらなかったため、顔の静止画のディープフェイクを対象とするものを取り上げる。

イ．Yan らの研究

（イ）概要

Yan らは、顔の動画のディープフェイクを対象とする検知モデル 15 件[11]を対象に評価・比較した（Yan *et al.* [2023]）。具体的には、①AI・機械学習分野におけるトップクラスの学会や論文誌で発表されている、②顔の静止画のディープフェイクの検知にも使用できるように、動画のフレーム内の情報を手掛りとする、③（検知モデルの基になる）検知手法がオープンソースとして公開されているという条件を満たす検知モデルが対象として選ばれた。

評価には、公開されているディープフェイクのデータセット 9 件（ディープ

イクであると判定。ディープフェイクでない入力の場合にはディープフェイクでないと判定）の数の割合であり、0 と 1 の間の値となる。値が大きいほど望ましい。

[11] 対象となった検知モデルは、MesoNet、MesoInception、CNN-Aug、EfficientNet-B4、Xceptionnet、Capsule Forensics、RECCE、Face X-ray のほか、Xceptionnet をベースとした 7 件（FFD、CORE、DSP-FWA、UCF、F3Net、SPSL、SRM）である。

表１ ディープフェイク検知モデルの評価・比較に関する主な研究事例

研究事例（文献）	データセットとディープフェイク生成手法	評価対象の検知モデル	主な評価指標値	評価指標値が高い検知モデル
Yan *et al.* [2023]	約 13 万件の動画（公開済みのデータセットに含まれているものを使用）	15 件 ・LLM を用いない ・独自にモデルを生成	AUC 再現率 適合率 正解率	SPSL（フレーム内の情報に基づく手法）
Bei *et al.* [2024]	約 42 万件の動画（独自に生成） ― 生成手法：16 件	7 件 ・LLM を用いない ・独自にモデルを生成	AUC	Exposing（フレーム間の関係性に基づく手法）
Le *et al.* [2025]	・約 200 件の動画（独自に生成） ― 生成手法：7 件 ・約 1400 件の動画（動画プラットフォームから抽出） ― 生成手法：不明	16 件 ・LLM を用いない ・既存のモデルを使用	AUC F 値 再現率 適合率 正解率	FTCN AltFreezing （フレーム間の関係性に基づく手法）
Narayan, VS, and Patel [2025]	・300 件のプロンプト（独自に生成） ・300 件の静止画（公開済みのデータセットから抽出） ― 生成手法：不明	28 件（MLLM）	正解率	GPT-4o

（備考）Narayan, VS, and Patel [2025] は顔の静止画のディープフェイク検知を対象としている。

フェイクは合計約 13 万件）[12]が用いられ、これらに含まれる動画の顔の部分を抽出し、統一的なフォーマットに加工・修正したものが使用された。ディープフェイクの生成手法は、主に、リエナクトメントまたはリプレイスメントに分類されるものである。

各検知モデルは、9 件のデータセットのうちの 1 件に含まれる動画（群）によって生成された。

（ロ）評価・比較

Yan らは、検知モデルの訓練に使用したデータセットを除く 8 件のデータセットに含まれるデータをそれぞれ各検知モデルに入力し、評価指標値（AUC）を測

[12] 具体的には、FaceForensics++、CelebDF-v1、CelebDF-v2、DeepFake Detection、DeepFake Detection Challenge Preview、DeepFake Detection Challenge、UADFV、FaceShifter、DeeperForensics-1.0 である。これらのうち、FaceForrensics++、DeepFake Detection、DeepFake Detection Preview、DeepFake Detection Challenge には、複数の異なる手法によって生成されたディープフェイクが混在している。他のデータセットのディープフェイクについては、それぞれ同一の手法によって生成されている。

定した。その結果、いずれの検知モデルにおいても AUC の値がデータセットによって大きく変化し[13]、すべてのデータセットに対して偏りなく高い判定精度を示す検知モデルは存在しなかった。

また、Yan らは、すべてのデータセットのディープフェイクを生成手法によって 4 つのグループに分けたうえで、そのうちの 1 つのグループに属するディープフェイクのみを用いて各検知モデルを生成し、残りの 3 つのグループのディープフェイクを各検知モデルにそれぞれ入力して評価指標値を測定した（クロス・バリデーション）。その結果、いずれの検知モデルにおいても、入力したディープフェイクのグループによって AUC の値が大きく変化した[14]。

Yan らは、今回対象とした検知モデルでは、訓練用データに含まれていないディープフェイクを高い確率で検知することが難しい[15]との見方を示した。

ロ．Bei らの研究

（イ）概要

Bei らは、顔の動画や静止画のディープフェイクのデータセットを構築するとともに、顔の動画や静止画のディープフェイクを対象とする検知モデルを評価・比較した（Bei *et al.* [2024]）。動画のディープフェイクを対象とする検知モデルとして、提案論文などから比較的高い性能を期待できる 7 件を対象とした[16]。

データセットは、顔の動画・静止画のディープフェイクを生成する手法 34 件によって独自に生成されたディープフェイク（約 35 万件の静止画、約 42 万件の動画）などから構成されている。動画のディープフェイクは 16 件の手法によって生成され、リエナクトメントおよびリプレイスメントに属するものに加え、シンセティック・フェイス・ジェネレーションに属する手法も用いられている[17]。

---

13 例えば、検知モデル SPSL（Spatial-Phase Shallow Learning）は、AUC の平均値（約 0.79）が最も高かったものの、データセットによって AUC の値が 0.64 から 0.95 の間で変化した。SPSL は、動画のフレーム内の情報を手掛りとする手法の 1 つであり、顔の形状、特徴点の位置などの情報に加え、画素値の系列を周波数領域の系列に変換したデータを用いて検知する（Liu *et al.* [2021]）。

14 各検知モデルにおける AUC の値は 0.50 から 0.80 の間の値であった。

15 ディープフェイク検知の研究では、検知モデルの訓練用データに含まれていないディープフェイクであっても高い確率で検知できるという性質（汎用性）の実現が研究目標の 1 つとなっている。そのため、評価ではクロス・バリデーションが実施されることが多い。

16 対象となった検知モデルは、MesoNet、EfficientNet-B0、Xceptionnet、F3-Net、CViT、SLADD、Exposing である。

17 リエナクトメントに属する生成手法が 4 件（ATVG-Net、Motion-cos、Talking Head Video、FOMM）、リプレイスメントに属する手法が 7 件（MMReplacement、DSS、DeepFakes、FaceShifter、BlendFace、SimSwap、FSGAN）、シンセティック・フェイス・ジェネレーションに属する手法が 5 件（AnimateDiff-Lightning、Hotshot、Zeroscope、MagitTime、AnimateLCM）である。

（ロ）評価・比較

Bei らは、ある 1 つの手法によって生成されたディープフェイクを用いて各検知モデルを生成し、他の手法によるディープフェイクを「未知のディープフェイク」として各検知モデルに入力して評価指標値（AUC）を測定した。その結果、いずれのモデルにおいても、入力されたディープフェイクの種類によって AUC の値が大きく変化した[18]。

また、Bei らは、各検知手法において、シンセティック・フェイス・ジェネレーションによるディープフェイクで生成した検知モデル（モデル S）と、リプレイスメントやリエナクトメントによるディープフェイクで生成した検知モデル（モデル R）をそれぞれ準備し、モデル S に対してリプレイスメントやリエナクトメントによるディープフェイクを入力した際の判定精度と、モデル R に対してシンセティック・フェイス・ジェネレーションによるディープフェイクを入力した際の判定精度をそれぞれ測定・比較した。その結果、モデル S の判定精度がモデル R の判定精度よりも低くなる傾向がみられた[19]。

今回対象とした検知モデルについて、Bei らは、さまざまなタイプのディープフェイクを偏りなく高い確率で検知することが難しく、検知モデルの訓練用データに含まれていないディープフェイクが入力されると判定精度が低下する可能性があるとの見方を示している。そのうえで、実際の生体認証に適用する際には、想定されるディープフェイクの種類を考慮して検知モデルを選択する必要がある旨を指摘している。

---

[18] 例えば、AUC の値が比較的高かった検知モデル Exposing は、モデル生成に用いられなかったタイプのディープフェイクを入力すると、AUC の値が 0.59 から 0.90 の間で変化した。Exposing は、各フレーム内の特徴点の周辺領域の情報を手掛りにディープフェイクか否かを判定する手法である（Ba *et al.* [2024]）。具体的には、顔の左右の対称性（形状、位置、皮膚の色など）、目・口・歯の輪郭の明確性、これらのフレーム間の整合性などをチェックする（複数のフレーム間の関係性を手掛かりにする手法の類型に属する）。

[19] この傾向について、Bei らは、モデル R がフレーム内の特徴点に着目して判定するのに対して、モデル S は特徴点だけでなく、それ以外の顔や背景の領域の情報も考慮して判定しているためと考察している。すなわち、モデル S は、リプレイスメントやリエナクトメントによるディープフェイク（特徴点周辺のみが加工・合成対象）をチェックする際に、加工・合成されていない領域（特徴点周辺以外の顔の部分や背景）もチェック対象とすることから、ディープフェイクであると判定しづらくなる。一方、モデル R は、シンセティック・フェイス・ジェネレーションによるディープフェイクをチェックする際に、特徴点周辺のみをチェック対象とし、加工・合成されていない領域をチェックすることがない。Bei らは、こうした加工・合成の状況の違いが判定精度の差異につながったとの見方を示している。

ハ．Le らの研究

（イ）概要

Le らは、顔の動画のディープフェイクのデータセットを構築するとともに、検知モデル 16 件[20]を評価・比較した（Le *et al.* [2025]）。対象とする検知モデルは、2019 年から 2023 年にかけて主要な学会・論文誌において提案論文が発表されている 51 件のなかから、複数のタイプのディープフェイクを検知可能である旨が提案論文で主張されているとともに、検知モデルが生成済みのものを選択した（既存の検知モデルを使用して評価を実施）。なお、各検知モデルの訓練用データは統一されていなかった。

Le らは、7 種類の手法[21]によってそれぞれ生成したディープフェイク（約 200 件）、および、インターネット上の動画プラットフォーム[22]で公開されているディープフェイク（約 1,400 件）を含むデータセットを構築した。動画プラットフォームのディープフェイクの生成手法は不明であった。

（ロ）評価・比較

Le らは、独自に生成したディープフェイクをそれぞれ各検知モデルに入力したところ、2 件の検知モデル（FTCN、AltFreezing）[23]の判定精度が、いずれのタイプのディープフェイクに対しても比較的高い値となった[24]。Le らは、これらの検知モデルがいずれも複数のフレーム間における顔の形状などの時系列的な整合性を手掛りとしていることから、今回評価に使用したディープフェイクに対して、複数のフレーム間の関係性を手掛りとする手法が有効であるとの見方を示している。

動画プラットフォームから入手したディープフェイクを各検知モデルに入力したところ、ディープフェイクをまったく検知できなかった検知モデルが 2 件

---

20 検知モデルは、Xceptionnet、FTCN、MAT、SBIs、CADDM、AltFreezing、LGrad、CCViT、Capsule Forensics、LRNet、CLRNet、ICT、MCX-API、LipForensics、EfficientNet-B4、ADD である。

21 生成手法はオープンソースとして提供されているものであり、DeepFaceLab、Faceswap、Dfaker、LightWeight、FOM-Animation、FOM-Faceswap、FSGAN であった。いずれもリエナクトメントまたはリプレイスメントに分類される手法によるものであった。

22 使用された動画プラットフォームは、Reddit、YouTube、Bilibili、TikTok であった。これらにおいて公開されている動画のうち、AI によって生成された旨のラベルが付与されているものがディープフェイクとみなされて使用された。

23 FTCN（Fully Temporal Convolution Network）は、複数のフレームにおいて顔の形状や特徴点の位置などが整合的か否かを重視する検知モデルである（Zheng *et al.* [2021]）。AltFreezing は、動画の各フレーム内の情報、および、複数のフレーム間の関係性に基づいて別々のニューラルネットワークを生成し、それらを組み合わせた検知モデルである（Wang *et al.* [2023]）。

24 FTCN と AltFreezing の評価指標値は、平均すると、AUC の値がともに約 0.98、F 値がそれぞれ約 0.93、約 0.94 であった。また、再現率はそれぞれ約 0.95、約 0.94 であった。

存在していたほか、その他の検知モデルにおいても判定精度が低下した[25]。判定精度の低下について、Le らは、各検知モデルの訓練用データに含まれていなかったタイプのディープフェイクが動画プラットフォーム上に存在し、それらを検知することができなかった可能性があるとしている。

ニ．Narayan らの研究

（イ）概要

Narayan らは、顔の静止画を対象とする各種タスクを実行する MLLM を評価・比較するためのデータセット（FaceXBench）を構築するとともに、それぞれのタスクに関して複数の MLLM[26]を評価・比較した結果を発表した（Narayan, VS, and Patel [2025]）。このタスクには顔の静止画のディープフェイク検知も含まれており、ここではディープフェイク検知に関する結果に絞って紹介する。

データセットとして、公開されているデータセット 2 件[27]から選択・抽出されたディープフェイク 300 件と、ディープフェイク検知用のプロンプト（独自に作成）300 件が準備された。300 件のプロンプトには以下の 3 種類が含まれていた[28]。

- 複数のディープフェイクを提示しつつ、タスクの内容を説明しないで、「これらのうち、どの静止画がディープフェイクですか？」という質問のみを含むプロンプト（ゼロショット・タスク）
- ディープフェイク検知というタスクの内容を説明したうえで、複数のディープフェイクを提示しつつ、「どの静止画がディープフェイクですか？」という質問を行うプロンプト（インコンテキスト・タスク）
- ディープフェイクを提示しつつ、タスクの説明に加えて、「質問への回答をステップ・バイ・ステップで考えてください。そして、回答を出力するだけでなく、その理由や考え方も示してください」という指示を含むプロンプト（チェーンオブソート・プロンプティング）

評価対象とした MLLM については、オープンソースとして公開されているも

---

25 各検知モデルにおける AUC の値は 0.39 から 0.69 の間であった。比較的高い判定精度が測定された FTCN と AltFreezing においても、AUC の値がそれぞれ約 0.58、約 0.60 となった。

26 ここでの MLLM は、テキストと静止画がプロンプトとして与えられ、テキストによって示されるタスクの回答をテキストで出力するものである。

27 データセットとして CelebDF と FaceForensics++が選択された。

28 300 件のプロンプトのうち、3 種類のプロンプトがそれぞれいくつ含まれていたかについては、論文に記載されていない。

の 26 件[29]と、商用サービスとして提供されているもの 2 件（GPT-4o、GeminiPro 1.5）が選択された。

（ロ）評価・比較結果

Narayan らは、各 MLLM に対してプロンプトをそれぞれ入力して正解率を測定した。すべてのプロンプトを対象とした場合の正解率は、いずれの MLLM においても低い値となった[30]。また、3 種類の MLLM を対象に、プロンプトの種類別に正解率を計測・比較したところ、正解率はほとんど変化しなかった[31]。

この結果を踏まえ、Narayan らは、今回対象とした MLLM における判定精度が低く、現時点ではディープフェイク検知に有効な手段とは言い難いほか、プロンプトの種類を変更しても判定精度の向上を期待しづらいとの見方を示した。

（４）小括

本節（3）で紹介した研究事例をまとめると以下のとおりである。

- Yan らの事例、Bei らの事例、Le らの事例は、顔の動画のディープフェイクを対象とする検知モデルを評価している（いずれも LLM を用いない）。評価指標として AUC などを採用している点で共通しているが、データセット、ディープフェイク生成手法、検知モデルの種類や準備方法（独自生成、または、既存のものを使用）などが区々であり、事例間の比較が困難である。また、検知モデルの訓練用データに含まれていないディープフェイクに対して、いずれの検知モデルも判定精度が低下するなどの傾向がみられた。
- 上記の 3 つの事例において評価指標値が相対的に高かった検知モデルは区々であり、高い判定精度を期待できる検知手法のタイプや特徴を特定するには至っていない。

---

[29] 対象となったオープンソースの MLLM は、Chameleon-7b、CogVLM2-19b、Eagle-X4-8B-Plus、Idefics-9b-Instruct、Idefics2-8b、Idefics-80b-Instruct、InternVL2-8b、InternVL-Chat-v1.5、InternVL2-76b 、 LLaVA-OneVision-0.5b-OV 、 LLaVA-OneVision-7b-OV 、LLaVA-OneVision-7b-SI 、 LLaVA-OneVision-72b-OV、LLaVA-NeXT-Interleave-7b、LLaVA-v1.5-7b、LLaVA-v1.5-13b、Mantis-SIGLIP-8b、MiniCPM-Llama3-v2.5、Monkey-Chat、PaliGemma、Phi-3.5-Vision、Qwen2-VL-7b-Instruct、Qwen2-VL-72b-Instruct、VILA 1.5-13b、VILA 1.5-3b、VILA 1.5-40b である。

[30] 正解率はいずれの MLLM においても約 0.05 から約 0.30 の間の値となった。最も高い正解率（約 0.30）となったのは GPT-4o であった。

[31] 対象となった MLLM は Phi-3.5-VISION、Qwen2-VL-7B-Instruct、InternVL2-8B であり、ゼロショット・タスク、インコンテキスト・タスク、チェーンオブソート・プロンプティングの場合の正解率が、それぞれ、約 0.27 から約 0.30 の間の値、約 0.27 から約 0.32 の間の値、約 0.24 から約 0.32 の間の値となった。

- Narayan らの事例は、顔の静止画のディープフェイクを対象とする検知モデル（LLM を用いる）を評価している。評価指標として正解率が採用されており、いずれの検知モデルにおいても正解率が低く、現時点では現実的な対策として有効とはいえない。

４．考察

本節では、ディープフェイク検知モデルの活用を検討するユーザ（リスク・ベースでの意思決定を志向）の観点から、検知モデルの評価・比較の研究事例を活用する際にどのような対応が望ましいかを考察する。また、評価・比較を実施する研究コミュニティ側での主な課題についても考察する。

（１）ユーザ側の準備：検知モデルに関する要求事項

検知モデルの評価・比較は、「顔の動画を用いた生体認証において、ディープフェイクの提示による攻撃のリスクを軽減するために有効な検知モデルを選択する」ために行われる。検知モデルを選択するうえで、ユーザは、生体認証システムにおいて悪用される可能性があるディープフェイクの生成手法をリストアップするとともに、検知モデルの評価基準を設定しておくことが望ましい。

イ．想定されるディープフェイク生成手法のリストアップ

なりすましに悪用される可能性があるディープフェイクの生成手法として、公開済みの手法すべてが候補となる。もっとも、すべてをリストアップして検討することは現実的とはいえないため、対象を絞り込んだうえで優先順位を付けることが考えられる。例えば、以下のようなタイプの生成手法を対象とすることが考えられる。

- オープンソースとして公開されており容易にモデルを入手・使用できるなど、ディープフェイク生成に要する特別なノウハウやコストが不要な手法
- ディープフェイクによるなりすましの成功確率が比較的高いと評価されている手法
- ディープフェイクを高速・リアルタイムで生成できる手法

こうした考慮点を列挙し、より多くの考慮点に合致するものに高い優先順位を付けてリストアップすることが考えられる。

また、ディープフェイク生成手法は今後も進化する可能性があるため、生成手法のリストを継続的にメンテナンスすることが望ましい。想定される生成手法のリストを作成しても、時間の経過とともに既存手法の改良や新手法の出現に

よってリストは陳腐化する。ディープフェイク生成手法の研究動向をフォローし、新手法の追加や陳腐化した手法の削除などを定期的に実施することが重要である。

ロ．ディープフェイクの検知モデルの評価基準の設定

（イ）許容できるなりすまし成功確率の上限

ユーザは、生体認証によって保護するアプリケーションにおいて、ディープフェイクによるなりすましのリスクを見積ったうえで、なりすましの成功確率がどの程度であれば許容できるかを明らかにしておくことが望ましい。例えば、以下のステップでこれを行うことが考えられる（宇根［2024］）。

① 一定期間におけるなりすましによる被害額（X）をリスクとして見積もる。
- X は、一定期間におけるなりすましの試行回数の上限[32]と、1 回のなりすまし成功によって生じうる被害額の上限[33]をそれぞれ見積り、両者を掛け合わせて算出する。

② 許容できる被害額の上限（Y）をそれぞれ見積もる。
③ Y÷X＝Z を計算し、許容できるなりすまし成功確率の上限とする。

なりすまし成功確率は、ディープフェイクが入力として与えられた際に正規のユーザと誤って判定する確率であり、その確率は「1－再現率」に相当する。したがって、上記の手順で Z の値が決定されると、再現率が「1－Z」以上の値となる検知モデルが採用の候補となる。「1－Z」の値は、許容できる再現率の下限といえる。

（ロ）許容できる偽陽性率の上限

再現率などの評価指標値は、検知モデルの判定しきい値次第で変動する。例えば、ある判定しきい値のもとで、再現率が「1－Z」の値よりも大きくなったとしても、偽陽性率も高い値となる可能性がある[34]。偽陽性率が高いと、正規のユーザが生体認証をパスできずアプリケーションを使用できないケースが頻発しう

---

32 なりすましの試行回数の上限として、例えば、生体認証システムのユーザ数に、一定期間中に一人のユーザが生体認証システムにアクセスできる回数の上限を乗じて算出することが考えられる。また、既に生体認証システムが運用されており、なりすましなどの不正行為が観察されている場合、一定期間内の不正行為の回数を「なりすまし試行回数」とすることも考えられる。

33 被害額の上限に関しては、1 回のなりすましによって実行可能な取引額の上限や、不正に入手した情報資産の価値の上限によって見積もることが考えられる。

34 一般的に、高い再現率を達成するために判定しきい値を低く設定すると、偽陽性率も高くなる傾向がある。

る。ユーザは、ビジネス上、許容できる偽陽性率の上限を設定しておくことが望ましい。

許容できる偽陽性率の上限が決まれば、ユーザは検知モデルの評価結果（ROC）を参照し、偽陽性率と再現率の条件が同時に達成される判定しきい値が存在するか否かを確認する（図 4 参照）。そうした判定しきい値が存在すれば、採用の候補となる。

（２）評価・比較の研究事例を活用する際の留意点

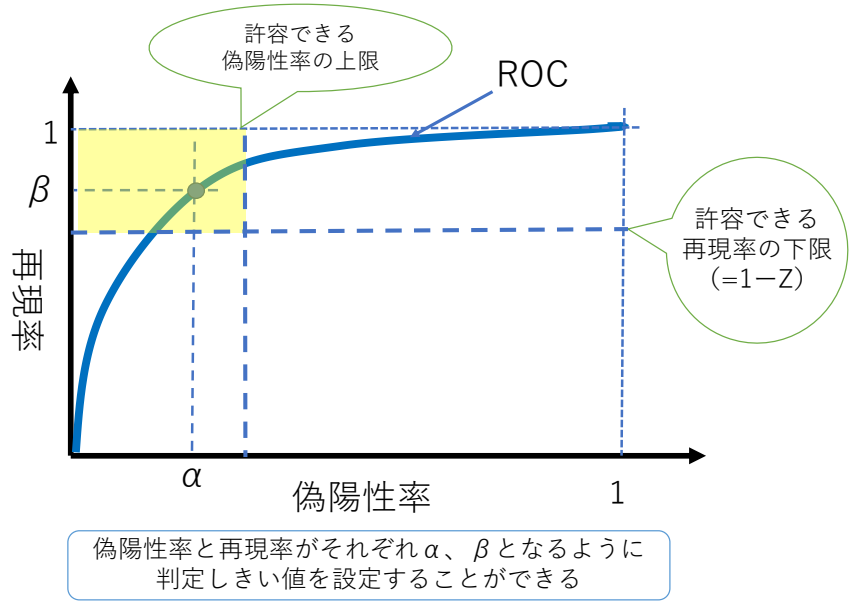
想定されるディープフェイク生成手法のリストアップ・優先順位付け、および、検知モデルの評価基準の設定を行った後、検知モデルの評価・比較の研究事例を活用して検知モデルを選択する際の留意点を考察する。

イ．ディープフェイク生成手法

まず、ユーザは、評価・比較の研究事例で使用されているデータセットに含まれているディープフェイクの生成手法を確認し、それらのなかにリストアップしたものが含まれているか否かを確認することが望ましい。含まれている生成手法が存在し、それらによって生成されたディープフェイクが検知モデルに入力されるなどして評価に活用されているならば、評価・比較の結果を参照する意義がある。

3 節（3）で紹介した Bei らの事例と Le らの事例では、データセットを構成するディープフェイクが独自に生成され、その際に用いられた生成手法が公開されていることから、準備したリストと照らし合わせることができる。Yan らの事

図 4 評価基準を満たす再現率と偽陽性率



例では、ディープフェイクの生成に使用されたデータセットが公開されていることから、各データセットの情報を入手してディープフェイクの生成手法を確認し、準備したリストと照合することが可能である。

一方、Le らの事例のうち、動画プラットフォームから抽出したディープフェイクによる評価に関しては、ディープフェイクの生成手法が不明であり、準備したリストと照合することができず、評価・比較の結果の活用が難しい[35]。

ロ．検知モデルの生成に用いられたデータ

評価対象の検知モデルの生成にどのようなデータが用いられたかを確認しておくことは、評価の妥当性を検討するうえで重要である。また、こうした情報は、評価・比較の結果として選択した検知モデルを生体認証システムに実装する際に、評価対象であった検証モデルを再現する（または検証モデルを改善する）うえで有用である。確認事項として主に次の点に留意することが望ましい。

- リストアップした生成手法によるディープフェイクが検知モデルの生成に使用されているか？
  - 使用されていないとすれば、リストアップした生成手法によるディープフェイクが検知モデルに入力された際に、判定精度が低下する可能性がある。
- 検知モデルの訓練用データにおけるディープフェイクとそれ以外のデータの量やバランスが妥当か？
  - 訓練用データの量がなるべく多い方が望ましい。ディープフェイクの量が少ないと、本来の判定精度が発揮されず低い精度に止まる可能性があることから、評価・比較の事例として参考にできない場合がある。
  - ディープフェイクと現実の動画のデータとの間の比率に偏りがあると、モデルの出力にも偏りが生じる可能性がある[36]。

---

[35] こうしたディープフェイクであっても、各検知モデルに入力して判定精度を計測・比較することは可能である。もっとも、生成手法が不明であるため、測定した評価指標がどのようなタイプのディープフェイクに対する判定精度を表しているかを特定できない。検知してほしいタイプのディープフェイクに対する効果を見極めるうえで参考にすることができるかも不明である。活用方法として、例えば、リストアップしたディープフェイクに関する評価結果が同一となる検知モデルが複数存在し、それらの優劣を決定する必要がある場合に、生成手法が不明なディープフェイクに関する評価結果を参照するという方法がありうる。

[36] Layton らは、訓練用データやテスト用データに含まれるディープフェイクと現実の動画のデータの比率を変化させながら、それぞれ検知モデルを生成して判定精度を測定したところ、比率の変化によって判定精度が有意に変化したと報告している（Layton *et al.* [2024]）。そのうえで、Layton らは、検知モデルの運用時における入力全体に占めるディープフェイクの割合を推定し、

- 複数の検知モデルを評価・比較している場合、それらの訓練用データが統一されていたか？
  - 各検知モデルの生成において異なる訓練用データが用いられていた場合、判定精度を同一条件で比較することが困難となり、評価指標値を比較する意義が失われる可能性がある。

3 節（3）における Yan らの事例と Bei らの事例では、検知モデルが独自に生成されており、モデルの訓練用のデータの詳細をそれぞれの論文執筆者に問い合わせることができる。一方、Le らの事例では、既に生成されていた検知モデルが評価対象となっており、各モデルの訓練用のデータを把握するためには、各モデル生成者に問い合わせる必要があると考えられる。また、訓練用のデータが各モデルで異なり、統一されていない可能性が高いとみられる。

Narayan らの事例の対象となっている MLLM の場合には、LLM をベースにそれぞれ生成されており、モデルの訓練用データの詳細を把握することは難しい。そのため、MLLM をブラックボックスとして評価結果を解釈する必要がある。

ハ．評価指標

研究事例で示される評価指標値には、再現率や偽陽性率が含まれていることが望ましい。これらが示されているならば、ユーザは許容できるなりすまし成功確率との関係を考慮し、生体認証システムへの実装の対象となりうるか否かを検討することができる。

3 節（2）で紹介した Yan らの事例、Bei らの事例、Le らの事例では、いずれも AUC が評価指標となっており、AUC の値を算出する際に用いられた ROC から再現率と偽陽性率の関係を知ることができる[37]。一方、Narayan らの事例では、正解率のみが示されており、AUC が示されていない。したがって、これらの事例では、再現率や偽陽性率の関係を確認することができない可能性がある。

（３）研究コミュニティにおける課題

評価・比較の研究事例を踏まえると、研究コミュニティにおける主な課題として、研究事例間で評価結果を比較することが難しいことが挙げられる。3 節（3）で紹介したように、評価の前提・内容・指標が研究事例の間で区々であり、第三

それに合致するように検知モデルの訓練用データに占めるディープフェイクの割合を決定する必要があるとしている。

[37] 研究事例の論文に AUC の記載があるものの、ROC が記載されていないというケースもありうるが、論文執筆者に ROC のデータを問い合わせることによって再現率と偽陽性率の関係を確認することができると考えられる。

者が複数の研究事例の結果を比較することは困難である。個々の研究事例の中で、それぞれが対象としている検知モデルを評価・比較することができるに過ぎない。

対応として、研究者が、評価・比較研究の成果を公開するにあたり、データセットを構成するディープフェイクの生成手法、検知モデルの生成に用いられた訓練用データなどを公開し、別の研究者もそれらを可能な限り参照・使用できるようにすることが望ましい[38]。異なる研究者が同じ前提や内容で評価を実施できるようになれば、評価・比較の先行研究の追試も可能となり、結果の信頼性の向上にもつながると期待できる。

また、3 節（2）において説明したように、ユーザがなりすましによるリスクを考慮して検知モデルの評価基準（許容できる再現率の下限や偽陽性率の上限）を設定するとすれば、判定精度として、再現率と偽陽性率の関係（ROC）が公表されることが望ましい。この点についても各研究事例において統一されれば、研究事例間での比較可能性が高まると期待できる。

５．おわりに

本稿では、顔の動画を用いた生体認証への脅威としてディープフェイクによるなりすましを想定し、対策手段の 1 つであるディープフェイク検知に焦点を当てて、主な検知手法の類型、検知モデルの評価・比較に関する主な研究事例を紹介した。また、ユーザが研究事例を活用する際の留意点、研究コミュニティにおける主な課題について考察した。

顔の動画を用いた生体認証は、パソコンやスマートフォンに標準装備されているカメラや画像処理のソフトウェアを用いて実施可能である。そのため、生体認証を実施するうえで改めてカメラやソフトウェアを準備する必要がない（実装のためのコストが低い）ほか、パスワードを覚えたり認証のための特別なデバイス（ワンタイムパスワードトークンなど）を安全に保持したりする必要がない（ユーザの利便性が高い）。こうしたメリットを享受するという観点からは、ディープフェイクによるなりすましのリスクを他の技術を用いてどのように軽減するかが重要となる。

---

38 こうした対応は既に一部の学会で始まっている。例えば、セキュリティ分野のトップカンファレンスの 1 つである USENIX Security Symposium では、投稿論文の査読の公正性を担保することに加え、研究成果の再現性（replicability）と再生産性（reproducibility）を促進するために（Open Science Policy）、投稿論文の執筆者に対して、執筆や研究に使用したデータやツールを投稿時点で公開するとともに、それらのリストや格納場所を投稿論文に明記することを求めている（2026 年の Call for Papers。アクセス先 : https://www.usenix.org/conference/usenixsecurity26/call-for-papers。アクセス日：2025 年 8 月 29 日）。

ディープフェイク対策としてディープフェイクの検知モデルを採用する際に、ユーザは、リスク管理の観点から、なりすましに悪用されうるディープフェイクの生成手法をリストアップし、検知モデルの評価基準（許容できる再現率の下限や偽陽性率の上限）を設定することが望ましい。そのうえで、公開されている研究事例を参照しつつ、評価基準に合致するものを検討することが考えられる。なりすましのリスクの評価などが困難な場合には、音声による認証など、他の認証手段と組み合わせるマルチモーダル認証を採用する（多重防御）、または、認証成功によって実行できる取引の金額の上限を低く設定するといった対応を検討することが有用である。

生体認証を活用している金融機関は、リスク管理の一環としてディープフェイクに関する技術動向をフォローし、現在採用しているなりすましへの対策手段の有効性やリスクのレベルを確認することが望ましい。ディープフェイクに関する研究は非常に活発であり、新しい生成技術や検知技術が次々に提案されている。既存の対策手段が陳腐化するスピードも速いことから、技術動向のフォロー、リスクの再評価、対策手段の見直しを継続的に行うことが重要である。

以　上

【参考文献】

宇根正志、「生体認証システムにおける人工物を用いた攻撃に対するセキュリティ評価手法の確立に向けて」、『金融研究』第 35 巻第 4 号、日本銀行金融研究所、2016 年、55～90 頁

――――、「スマートフォンによる顔認証のセキュリティ：ディープフェイクによる脅威と対策」、『金融研究』第 43 巻第 4 号、日本銀行金融研究所、2024 年、95～122 頁

川名のん・大島敬志・鈴木　茜・吉野雅之、「Deepfake 動画を用いた eKYC に対するなりすまし攻撃の検知手法の検討」、『第 38 回人工知能学会全国大会論文集』、4M3-GS-10-03、人工知能学会、2024 年、1～4 頁

――――・長沼　健・吉野雅之・太田原千秋・冨樫由美子・笹　晋也・山本恭平、「Deepfake を用いた e-KYC に対するなりすまし攻撃と対策の検討」、『第 35 回人工知能学会全国大会論文集』、1F2-GS-10a-02、人工知能学会、2021 年、1～4 頁

金融庁、「AI ディスカッションペーパー（第 1.0 版）－金融分野における AI の健全な利活用の促進に向けた初期的な論点整理－」、金融庁、2025 年（https://www.fsa.go.jp/news/r6/sonota/20250304/aidp.pdf、2025 年 7 月 7 日）

坂本静生・宇根正志、『AI・量子コンピュータにかかわるリスク管理』、オーム社、2025 年

笹原和俊、『ディープフェイクの衝撃：AI 技術がもたらす破壊と創造』、PHP 研究所、2023 年

Ba, Zhongjie, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren, "Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection," *Proceedings of the 38th AAAI Conference on Artificial Intelligence* (2024): 719-28. https://doi.org/10.1609/aaai.v38i2.27829.

Bai, Jianfa, Man Lin, and Gang Cao, "AI-Generated Video Detection via Spatio-Temporal Anomaly Learning," arXiv: 2403.16638 (2024). https://arxiv.org/pdf/2403.16638.

Bei, Yijun, Hengrui Lou, Jinsong Geng, Erteng Liu, Lechao Cheng, Jie Song, Mingli Song, and Zunlei Feng, "A Large-Scale Universal Evaluation Benchmark for Face Forgery Detection," arXiv: 2406.09181v2 (2024). https://arxiv.org/pdf/2406.09181.

Deng, Jingyi, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Qian Wang, and Chao Shen, "A Survey of Defenses against AI-Generated Visual Media: Detection, Disruption, and Authentication," arXiv: 2407.10575v1 (2024). https://arxiv.org/pdf/2407.10575.

Financial Crimes Enforcement Network. 2024. "FinCEN Alert on Fraud Schemes Involving Deepfake Media Targeting Financial Institutions." Effective July 7, 2025. https://www.fincen.gov/sites/default/files/shared/FinCEN-Alert-DeepFakes-

Alert508FINAL.pdf.

Gani, Hanan, Rohit Bharadwaj, Muzammal Naseer, Fahad Shahbaz Khan, and Salman Khan, "VANE-Bench: Video Anomaly Evaluation Benchmark for Conversational LMMs," arXiv: 2406.10326v2 (2025). https://arxiv.org/pdf/2406.10326.

Gaur, Loveleen, *DeepFakes: Creation, Detection, and Impact*, CRC Press, 2023.

He, Peisong, Leyao Xhu, Jiaxing Li, Shiqi Wang, and Haoliang Li, "Exposing AI-Generated Videos: A Benchmark Dataset and Local-and-Global Temporal Defect Based Detection Method," arXiv: 2405.04133 (2024). https://arxiv.org/pdf/2405.04133.

Khan, Farrukh Aslam, and Muhammad Khurram Khan, "Generative AI and Deepfake Detection in Biometric Systems," *Cognitive Computation* 17, no. 112 (2025): 1-21. https://link.springer.com/article/10.1007/s12559-025-10469-3.

Komaty, Alain, Hatef Otroshi Shahreza, Anjith George, and Sébastien Marcel, "Exploring ChatGPT for Face Presentation Attack Detection in Zero and Few-Shot in-Context Learning," arXiv: 2501.08799v1 (2025). https://arxiv.org/pdf/2501.08799.

Layton, Seth, Tyler Tucker, Daniel Olszewsky, Kevin Warren, Kevin Butler, and Patrick Traynor, "SoK: The Good, The Bad, and The Unbalanced: Measuring Structural Limitations of Deepfake Media Datasets," *Proceedings of the 33rd USENIX Security Symposium* (2024): 1027-44. https://www.usenix.org/system/files/usenixsecurity24-layton.pdf.

Le, Binh M., Jiwon Kim, Simon S. Woo, Kristen Moore, Alsharif Abuadbba, and Shahroz Tariq, "SoK: Systematization and Benchmarking of Deepfake Detectors in a Unified Framework," arXiv: 2401.04364v4 (2025). https://arxiv.org/pdf/2401.04364.

Li, Xiaobai, Jukka Komulainen, Guoying Zhao, Pong-Chi Yuen, and Matti Pietikäinen, "Generalized Face Anti-Spoofing by Detecting Pulse from Face Videos," *Proceedings of the 2016 23rd International Conference on Pattern Recognition* (2016): 4239-44. https://ieeexplore.ieee.org/document/7900300.

Li, Changjiang, Li Wang, Shouling Ji, Xuhong Zhang, Zhaohan Xi, Shanqing Guo, and Ting Wang, "Seeing is Living? Rethinking the Security of Facial Liveness Verification in the Deepfake Era," *Proceedings of the 31st USENIX Security Symposium* (2022): 2673-90. https://www.usenix.org/system/files/sec22-li-changjiang.pdf.

Lin, Li, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu, "Detecting Multimedia Generated by Large AI Models: Survey," arXiv: 2402.00045v6 (2025). https://arxiv.org/pdf/2402.00045.

Liu, Honggu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu, "Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain," *Proceedings of the 2021 IEEE/CVE Conference on Computer*

Vision and Pattern Recognition (2021): 772-81. https://openaccess.thecvf.com/content/CVPR2021/papers/Liu_Spatial-Phase_Shallow_Learning_Rethinking_Face_Forgery_Detection_in_Frequency_Domain_CVPR_2021_paper.pdf.

Luo, Xiyang, Yinxiao Li, Huiwen Chang, Ce Liu, Peyman Milanfar, and Feng Yang, "Dvmark: A Deep Multiscale Framework for Video Watermarking," *IEEE Transactions on Image Processing* 34 (2023): 4371-85. https://ieeexplore.ieee.org/document/10086041.

Narayan, Kartik, Vibashan VS, and Vishal M. Patel, "FaceXBench: Evaluating Multimodal LLMs on Face Understanding," arXiv: 2501.10360v2 (2025). https://arxiv.org/pdf/2501.10360.

National Security Agency, Federal Bureau of Investigation, and Cybersecurity and Infrastructure Security Agency. 2023. "Contextualizing Deepfake Threats to Organization." Effective July 7, 2025. https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF.

Šalko, Milan, Anton Firc, and Kamil Malinka, "Security Implications of Deepfakes in Face Authentication," *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing* (2024): 1376-84. https://dl.acm.org/doi/pdf/10.1145/3605098.3635953.

Song, Xiufeng, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu, "On Learning Multi-Modal Forgery Representation for Diffusion Generated Video Detection," arXiv: 2410.23623v3 (2025). https://arxiv.org/pdf/2410.23623.

Sony, Redwan, Parisa Farmanifard, Hamzeh Alzwairy, Nitish Shukla, and Arun Ross, "Benchmarking Foundation Models for Zero-Shot Biometric Tasks," arXiv: 2505.24214v1 (2025). https://arxiv.org/pdf/2505.24214.

Tariq, Shahroz, David Nguyen, M.A.P. Chamikara, Tingmin Wu, Alsharif Abuadbba, and Kristen Moore, "LLMs Are Not Yet Ready for Deepfake Image Detection," arXiv: 2506.10474v1 (2025). https://arxiv.org/pdf/2506.10474.

Wang, Zhendong, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li, "AltFreezing for More General Video Face Forgery Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023): 4129-38. https://openaccess.thecvf.com/content/CVPR2023/papers/Wang_AltFeezing_for_More_General_Video_Face_Forgery_Detection_CVPR_2023_paper.pdf.

Wang, Zezheng, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, Guojun Qi, Jun Wan, and Zhen Lei, "Exploiting Temporal and Depth Information for Multi-Frame Face Anti-Spoofing," arXiv:1811.05118v3 (2019). https://arxiv.org/pdf/1811.05118.

Xu, Zhenqi, Shan Li, and Weihong Deng, "Learning Temporal Features Using LSTM-CNN

Architecture for Face Anti-Spoofing," *Proceedings of 2015 3rd IAPR Asian Conference on Pattern Recognition* (2015): 141-5. https://ieeexplore.ieee.org/document/7486482.

Yan, Zhiyuan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu, "DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection," *Proceedings of the 37th International Conference on Neural Information Processing Systems* 201 (2023): 4534-65. https://papers.nips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Dtasets_and_Benchmarks.pdf.

————, Yandan Zhao, Shen Chen, Xinghe Fu, Taiping Yao, Shouhong Ding, and Li Yuan, "Generalizing Deepfake Video Detection with Plug-and-Play: Video-Level Blending and Spatiotemporal Adapter Tuning," arXiv:2408.17065 (2024). https://arxiv.org/pdf/2408.17065.

Zhang, Yulin, Jiangqun Ni, Wenkang Su, and Zin Liao, "A Novel Deep Video Watermarking Framework with Enhanced Robustness to H.264/AVC Compression," *Proceedings of the 31st ACM International Conference on Multimedia* (2023): 8095-104. https://dl.acm.org/doi/abs/10.1145/3581783.3612270.

Zheng, Yinglin, Jinmain Bao, Dong Chen, Ming Zeng, and Fang Wen, "Exploring Temporal Coherence for More General Video Face Forgery Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021): 15044-54. https://openaccess.thecvf.com/content/ICCV2021/papers/Zheng_Exploring_Temporal_Coherence_for_More_General_Video_Face_Forgery_Detection_ICCV_2021_paper.pdf.

Zou, Yueying, Pieipei Li, Zekun Li, Huaibo Huang, Xing Cui, Xuannan Liu, Chenghanyu Zhang, and Ran He, "Survey on AI-Generated Media Detection: From Non-MLLM to MLLM," arXiv: 2502.05240v2 (2025). https://arxiv.org/pdf/2502.05240.