XAIを用いた人とAIの協調

2025年12月4日

氏名 芝垣 佳尚
所属 AWL株式会社

目次

1.課題

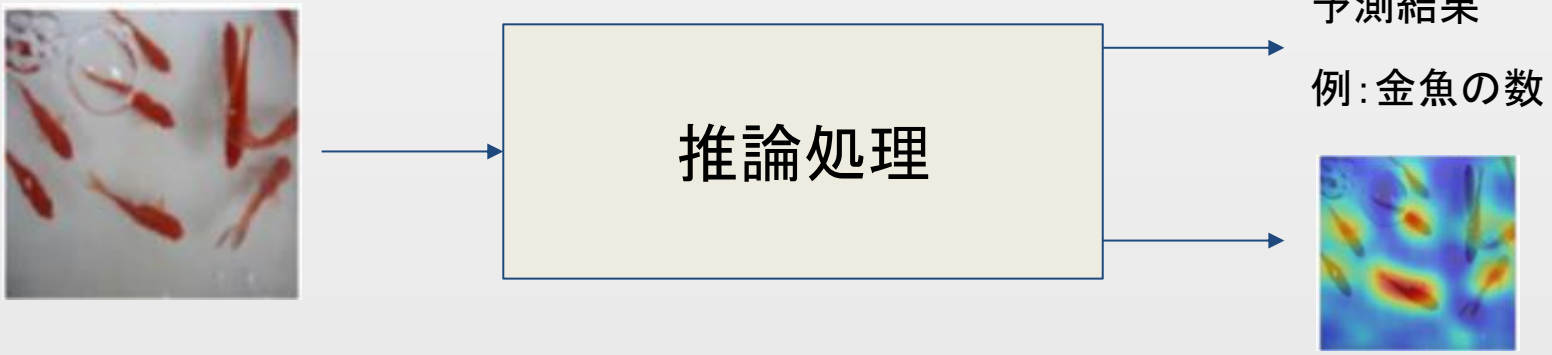
業務における課題

- カメラによる防犯システムは、経験豊富な専門家によって検出が行われている。しかし、この作業は人員確保や教育に時間とコストがかかり、大規模なシステム展開を阻害している
- この課題を解決するため、AIによる自動検出と専門家による最終判断を組み合わせた「人間＋AI協調システム」を導入し、専門家の負担軽減を目指す
- その際、専門家がAIの判断を信頼しやすくするため、AIの判定理由を提示するXAI（Explainable AI）の利用を検討する

- 本発表では、画像認識タスクにおいてXAIによる説明が、人間とAIの協調によるパフォーマンス（意思決定の精度）向上にどの程度寄与するかを調査した結果を報告する

XAIについて

- XAIとは、モデルから追加情報を抽出する技術[1]
  - 追加情報=説明
  - この説明を追加する事で、AIの予測を人間自身の専門知識と照らし合わせて検証が可能となり、人間+AIのパフォーマンスが向上すると考えられる
  - 画像タスクでは、ヒートマップや注釈など視覚的な説明が主流(CAM, Grad-CAM, LIME)[2]



金魚画像は[2]のFig11から抜粋

調査内容

- 画像認識タスクにおける、人間とAIの協調パフォーマンスをXAIの説明を用いて検証した、以下の2つの論文を基に調査を行った

- 文献1: XAIがパフォーマンスを向上させたケース
  - Explainable AI improves task performance in human–AI collaboration[3]
- 文献2: 説明が逆効果となったケース
  - Effects of Presenting Multiple Types of AI Explanations for Visual Task[4]

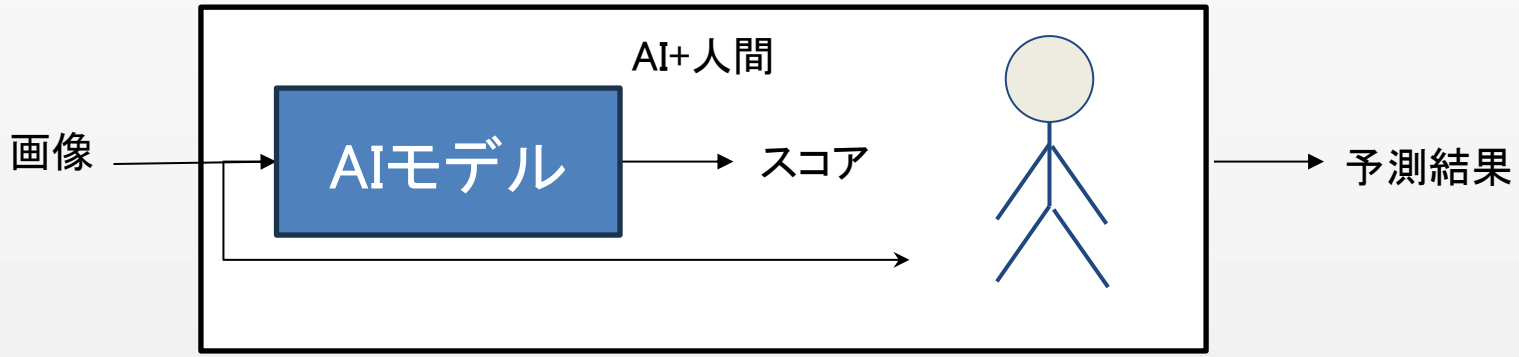
2.事例①：XAIがパフォーマンスを向上させたケース

パフォーマンスを向上させたケース

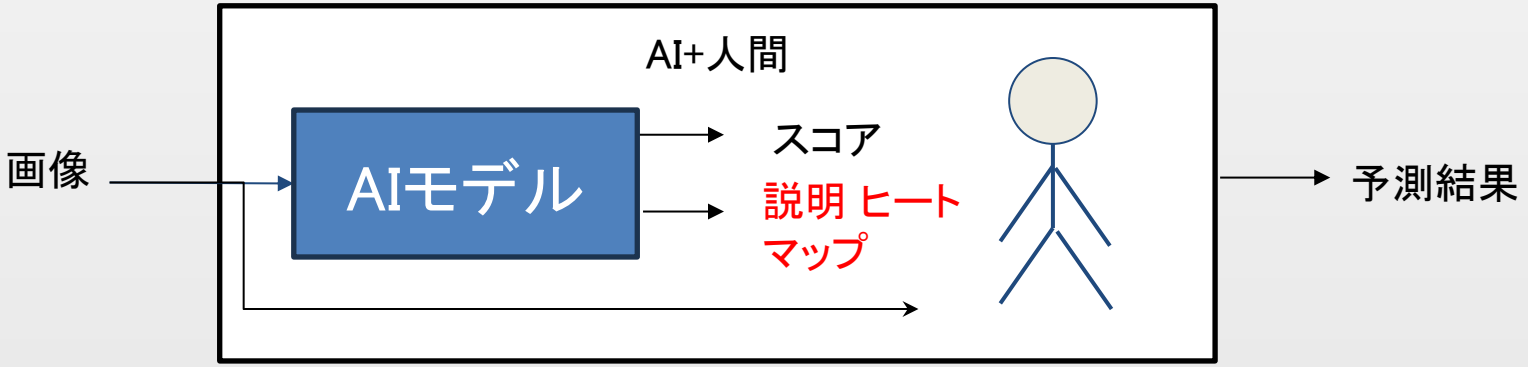
- 文献1
  - Explainable AI improves task performance in human–AI collaboration[3]
  - 人間とAIの協調により、出力のパフォーマンスが向上するか仮説検証を実施
  - 検証実験
    - Study1: 電子部品の欠陥検知
    - Study2: 肺の疾患検知
  - ブラックボックスAIとXAIで評価を実施

実験内容

ブラックボックスAI

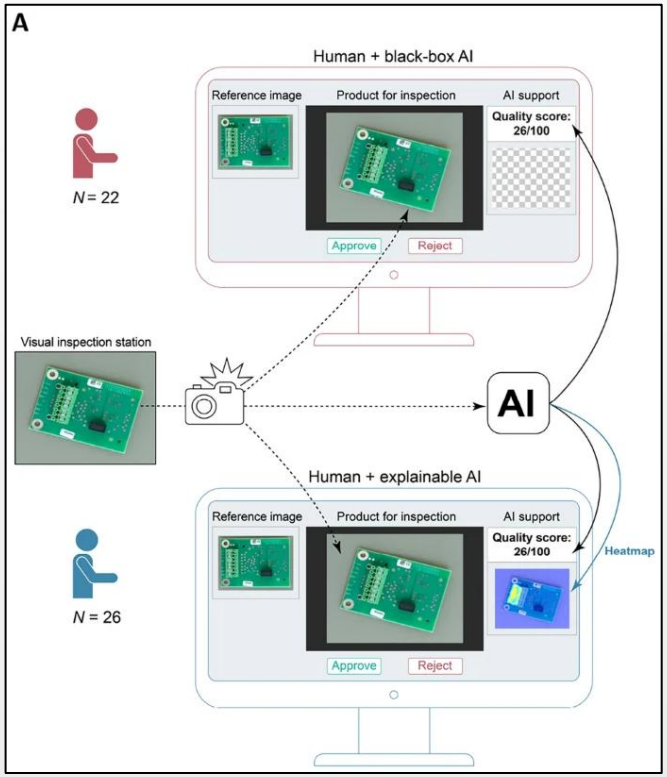


XAI



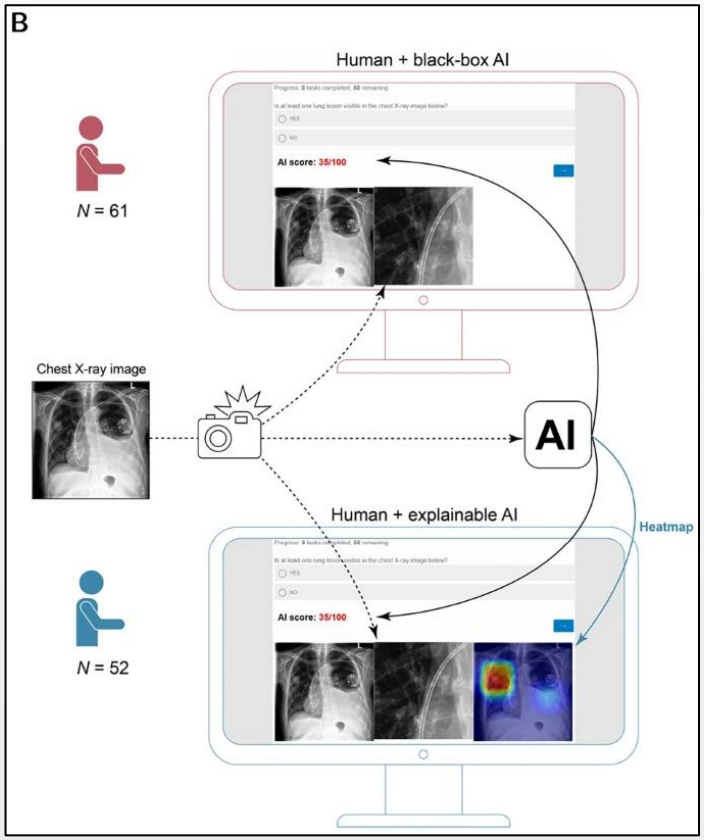
Study1:電子部品の欠陥検知

- 用意した画像: 4種類の電子部品 x 50パターンの画像(7枚が異常)
- 48人(検査員: ブラックボックス:22人、XAI:26人)
- 1人につき200枚(画像)
- 48人 x 200枚 = 9,600試行
  - 事前にAIのスコアが90以下の場合は間違える可能性ありと伝えた



Study2: 肺の疾患検知

- 用意した画像: 50パターンの肺のレントゲン画像(7枚が疾患あり)
- 113人(放射線医: ブラックボックス:61人、XAI:52人)
- 1人につき50枚(画像)
- 113人 x 50枚 = 5,650試行
  - 事前にAIのスコアが90以下の場合間違える可能性ありと伝えた



実験結果 Study1電子部品

- 実験結果の平均値を以下の表にまとめた

シナリオ	Balanced Accuracy	欠陥検出率
AIアルゴリズム単体	95.6%	92.9%
人間 + Blackbox AI	88.6%	82.0%
人間 + XAI	96.3%	93.0%

- 説明可能なAIを使った作業員はブラックボックスAIを使った作業員よりも優れたパフォーマンスを達成したことがわかった
- ブラックボックスAIの作業員は平均88.6%の精度を達成したのに対し、説明可能なAIの作業員は平均96.3%の精度を達成
- 説明可能なAIを利用した作業員は、欠陥検出率においてもブラックボックスAIを利用した作業員を上回り、平均82.0%に対して平均93.0%の精度を達成した

実験結果 Study2肺の疾患

- 実験結果の平均値を以下の表にまとめた

シナリオ	Balanced Accuracy	疾患検出率
AIアルゴリズム単体	82.2%	71.4%
人間 + Blackbox AI	79.1%	90.4%
人間 + XAI	83.8%	90.4%

- 説明可能なAIを利用した放射線医は、ブラックボックスAIを利用した同業の放射線医よりも優れた結果となった
- ブラックボックスAIを利用した放射線医の精度は平均79.1%にとどまったが、説明可能なAIを利用した放射線医は平均83.8%の精度を達成
- 電子部品の実験とは対照的に、疾患検出率は平均90.4%で成績に差は見られなかった
  - 実験前から予想されていた結果で、AIによる説明の有無にかかわらず医師は元々誤検知よりも偽陰性（見逃し）の回避を最優先した結果、たとえ説明を付加しても検出率向上の余地が無かったと考えらえる

実験結果まとめ

- ヒートマップ形式の説明可能AIがブラックボックスAIと比較して、人間とAIの協調におけるパフォーマンスを大幅に向上させることを実証した
  - ヒートマップを用いた説明により、電子部品で7.7ポイント、肺の疾患で4.7ポイントのパフォーマンス向上が確認された結果となった
  - 専門家がAIの正確な予測を受け入れ、誤りを修正する傾向が強まったため

▽

- AIが正しい場合 → 助言を信頼して活用
- AIが誤っている場合 → 人間の専門知識で修正

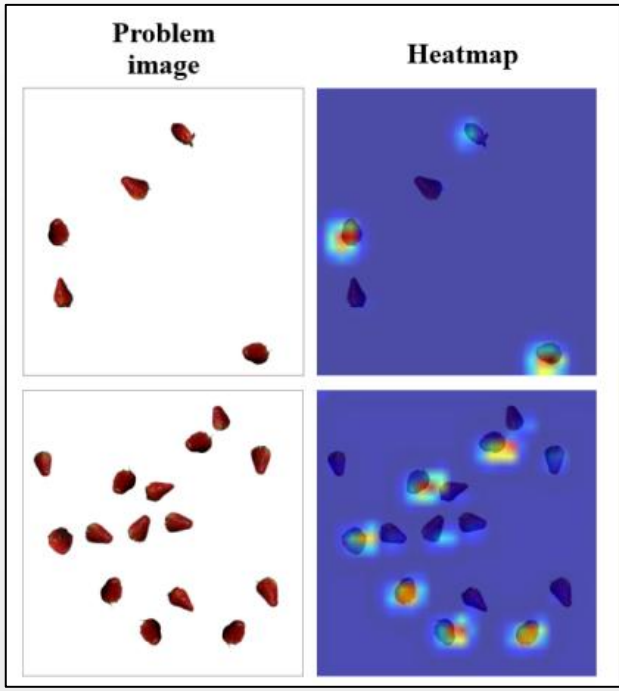
3.事例②：説明が逆効果となったケース

説明が逆効果になったケース

- 文献2
  - Effects of presenting multiple types of AI explanations for visual task[4]
  - 一般ユーザーがAIの説明をどのように受け取り、信頼や判断にどう影響するかを検証した内容
  - 画像認識タスクを用いて、AIの説明がユーザーの信頼やAI結果の受容に与える影響を実験的に調査

実験内容

- タスク：腐ったイチゴを見つける視覚識別課題
  - 5個もしくは15個の画像から識別
- 被験者：Yahoo!クラウドソーシングで募集した146名
- 測定項目：正解率、AI回答の受容率(AIからの回答を受け入れるか？)、AIへの信頼度、表示時間(人間が決定するまでの時間)
- 条件は3種類: 説明なし、ヒートマップ、全説明



実験内容

条件	表示内容
説明なし No explanation	結果表示-腐っているイチゴの有無
AI heatmap	結果＋ヒートマップ
全説明 All explanation	結果＋ヒートマップ AI自身の目的を表示 (this AI is designed to detect specific symbols and assist people) 信頼度を数値で表示(数値は0.6-0.7の間で出るように調整)

Figure 3: All-explanation, AI-heatmap, and no-explanation interface

実験結果

条件	正答率	信頼度	受容率	表示時間
説明なし（AI回答のみ）	最も高い	低い	低い	最も短い
ヒートマップのみ（AI回答＋ヒートマップ）	低い	高い	大差なし	長い
全説明（AI回答＋目的＋信頼性＋ヒートマップ）	低い	高い	最も高い	最も長い

- 文献2中に記載された検定結果を元に「説明なし」条件を基準として、他の条件（全説明・ヒートマップのみ）との統計的な有意差に基づいて「高い」「低い」などの比較評価を記載した
- 説明の種類による影響は、イチゴが5個でも15個でも同様に現れていたため、刺激数ごとの結果は分けずに記載
- 他の2条件よりも有意差がある場合は「最も」と記載
- 赤いセルがポジティブ、青いセルがネガティブな結果

実験結果

条件	正答率	信頼度	受容率	表示時間
説明なし（AI回答のみ）	最も高い	低い	低い	最も短い
ヒートマップのみ（AI回答＋ヒートマップ）	低い	高い	大差なし	長い
全説明（AI回答＋目的＋信頼性＋ヒートマップ）	低い	高い	最も高い	最も長い

- 説明があることで人間はAIをより信頼し回答を受け入れやすくなるが、実際の正答率は向上しない。むしろ下がる
  - この傾向は、タスクの難しさに関わらず見られており、イチゴ5個、15個表示に関わらず一貫して同じ傾向だった
  - 全説明の条件では最も受容が高くなるが、説明を表示することで過信や過度の依存につながり正答率は低下する

- **AIが間違っている場合でも、人間はAIを信じてしまって訂正できなかった**

4. 考察:XAIの効果を左右する要因とその対策

XAIの効果を左右する要因

- 文献1,2の結果から、XAIの効果を左右する要因として、協調する人間の専門知識が挙げられる
  - 専門家：説明によりAI協調の精度向上
    - タスクに対してドメイン知識のある専門家は、AIの説明を活用してAIの正しい判断を確信し、誤った判断を訂正する事ができる(文献1の実験結果より)
  - 非専門家：説明で信頼増すが精度向上せず
    - ドメイン知識の乏しい非専門家は、説明付きAIに対する信頼が高まる一方で、判断精度は改善ない
    - むしろ説明があることでAIを過信し、誤った結果をそのまま受け入れてしまうリスクがある(文献2の実験結果より)
  - 結果、人間のドメイン知識にあわせた説明が必要

ドメイン知識にあわせた説明とは？

- 専門家向け
  - 詳細で深い説明を提供することが有効
    - 例：ヒートマップやスコアなど
    - ある程度情報量が多い説明でも、AIの提案を否決する能力がある
    - ただし、情報過多になった場合、判断までの速度が遅くなりタスクに対する処理速度が低下する可能性がある
      - 迅速な判断を要するタスクにおいては、複数の文章を組み合わせた複雑な説明ではなく、ヒートマップや信頼度スコア表示のように、視覚的にわかりやすい説明が適切である

ドメイン知識にあわせた説明とは？

- 非専門家向け①

1.AIの自信の有無で説明量を変える[3][5]

- 自信が無いときに説明を量を増やすと、むしろ人間はそのAIの予測を信じやすい（過信が起こる）
- 説明量を減らして人間側に判断を促す必要がある
  - ただし、説明を減らしすぎると過小信頼や過小依存が生じる可能性があるので、最低限の情報（例：AIの信頼度スコアや簡単な根拠）を提示する

ドメイン知識にあわせた説明とは？

- 非専門家向け②

2.説明のタイミングを工夫する[5]

- AI予測＋説明の後に人間の判断を実施すると、AIの情報により先入観を生んでしまう
  - 人間の独立した判断を妨げる可能性
- 人間がまず自分で回答を考え、その後AIの予測と説明を確認する

5.まとめ

まとめ①

- 画像認識タスクにおいてXAIによる人間とAIの協調によるパフォーマンス向上の調査を実施した
  - パフォーマンスが向上した事例
  - 説明が逆効果となり向上しなかった事例
- 二つの事例から効果を左右する要因について議論した
  - 専門家：説明によりAI協調の精度向上
    - AIの説明を活用してAIの正しい判断を確信し、誤った判断を訂正する事ができる
  - 非専門家：説明で信頼増すが精度向上せず
    - 説明があることでAIを過信し、誤った結果をそのまま受け入れてしまうリスクがある

まとめ②

- 要因を元に、人間のドメイン知識に合わせたXAIの説明方法について提案
  - 専門家向け:
    - 詳細で深い説明を提供することが有効
  - 非専門家:
    - 1.AIの自信の有無で説明量を変える
    - 2.説明のタイミングを工夫する
- 今回は扱わなかったが、XAIは説明抽出や生成AIへの対応[6]、さらには評価指標(説明の質や有用性をどう測るか)や倫理面(説明の悪用)などさらなる発展の余地があり、今後多面的な研究を通じてより信頼性の高いAIの実現が期待される

ありがとうございました

参考文献

[1]機械学習モデルの判断根拠の説明

- 原聡 産総研知能研究センター【第40回AIセミナー】 https://www.slideshare.net/slideshow/ver2-225753735/225753735#17

[2]A comprehensive review of explainable artificial intelligence (XAI) in computer vision.

- Cheng, Z., Wu, Y., Li, Y., Cai, L., & Ihnaini, B. (2025). Sensors, 25(13), 4166. https://doi.org/10.3390/s25134166

参考文献

[3]Explainable AI improves task performance in human–AI collaboration

- Julian Senoner, Simon Schallmoser, Bernhard Kratzwald, Stefan Feuerriegel, Torbjørn Netland 2024 Scientific Reports https://doi.org/10.1038/s41598-024-82501-9

[4]Effects of presenting multiple types of AI explanations for visual task

- Maehigashi, A., Fukuchi, Y., & Yamada, S. (2024) Proceedings of the 12th International Conference on Human-Agent Interaction, 160–166. https://doi.org/10.1145/3687272.3688328

参考文献

[5]Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance

- Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, Daniel S. Weld 2021 CHI Conference on Human Factors in Computing Systems (CHI ’21), Yokohama, Japan

[6] Explainable Generative AI (GenXAI): a survey, conceptualization, and research agenda

- Johannes Schneider 2024 IEEE Access