Tomo-e Gozen 移動天体検出用機械学習モデルの改修

和田空大, 酒向重行, 紅山仁, 瀧田怜, 倉島啓斗
(東京大学)

2024/05/15 木曽シュミットシンポジウム2024

©東京大学木曽観測所

Outline

1. Introdution
2. Method
3. Result
4. Discussion
5. Prospect

Outline

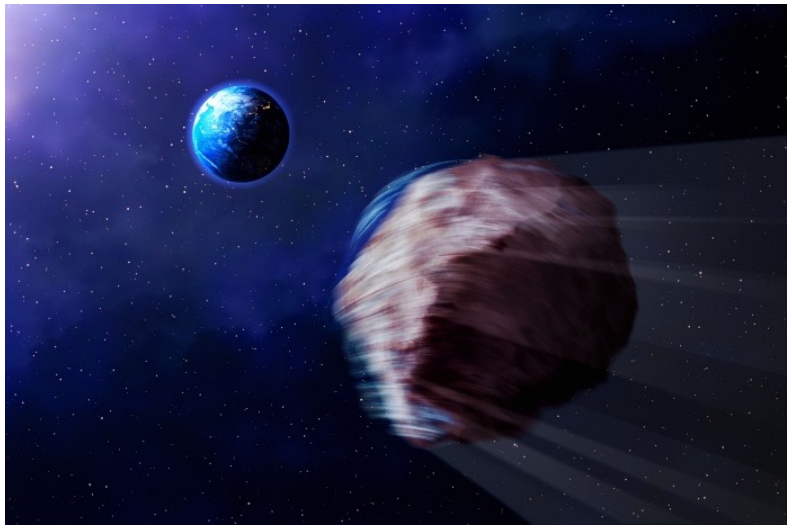
1. **Introdution**
2. Method
3. Result
4. Discussion
5. Prospect

地球接近小惑星（Near-Earth Objects, NEOs）

NEO：地球に近づく軌道を持つ小惑星（近日点距離が1.3 au 未満）

- 多くはメインベルト小惑星から軌道進化
- 地球への物質輸送に関する情報

- 見かけの移動が高速
- 観測好機が短い

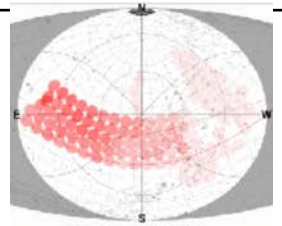


©東京大学木曽観測所

高時間分解能×即時フォローアップがその観測に有効

2019年から Tomo-e Gozen での NEO 探査開始

既存の高速移動天体検出システムとその課題



1. **全天サーベイと移動天体検出**

   検出された天体の特徴量を用いて
   移動天体らしさのスコアづけを行う
   （**機械学習(ランダムフォレスト)**）

2. **高いスコアの検出を選ぶ**

   高いスコアの検出を選ぶ

3. **追跡観測による軌道決定**

   人の目で見て移動天体と判断したものを
   追跡観測する

紅山仁 修士論文

典型的な高速移動天体候補の数

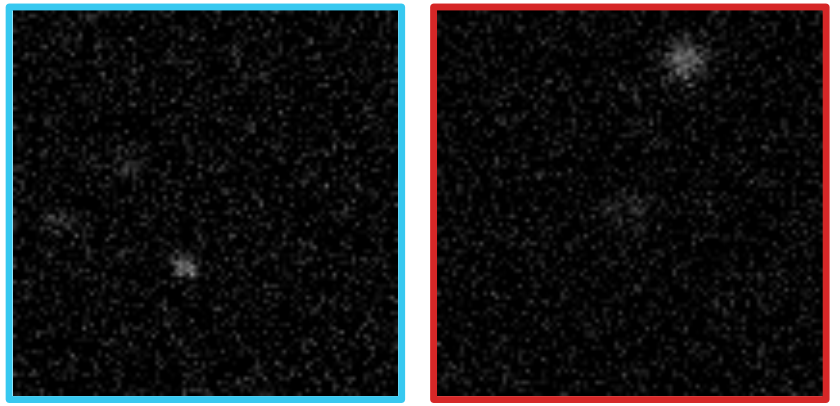
$\sim 10^6$/day

$\sim 10^2$/day

$\sim$ 10/day

現状：誤検出を多く生むため人の目で判断

→ **自動化が求まれる**



移動天体
→ 追跡観測へ

誤検出
→ 追跡しない

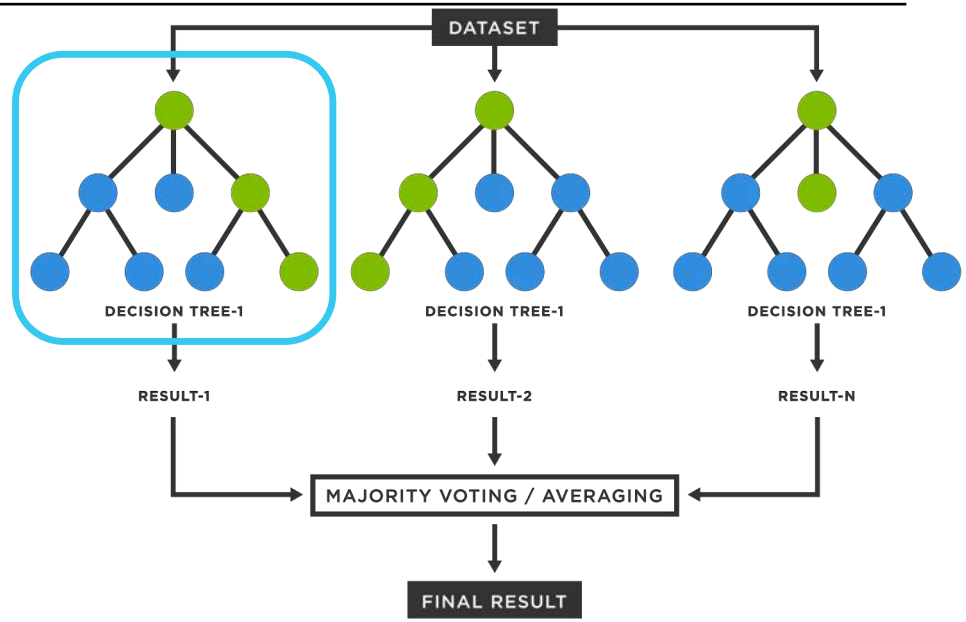
機械学習モデル

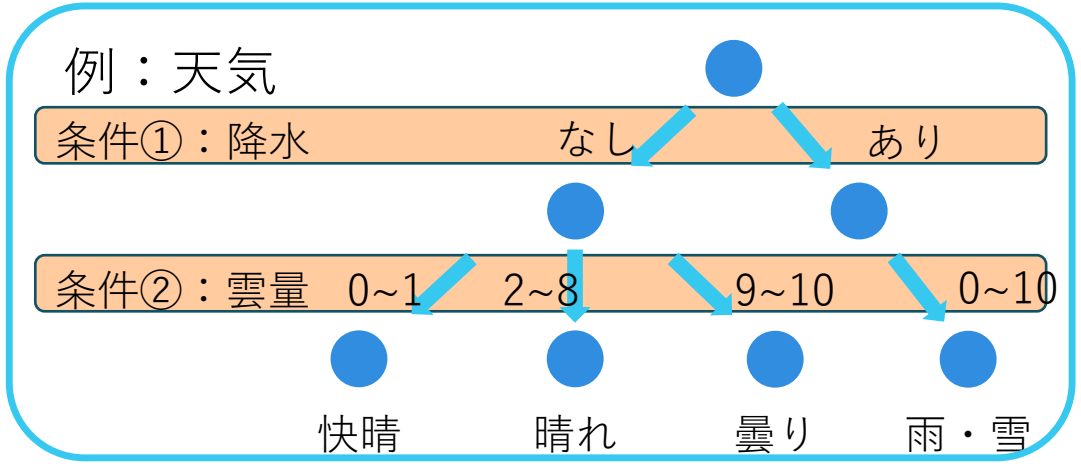
https://www.spotfire.com/glossary/what-is-a-random-forest

手法：**ランダムフォレスト**

- **分類**を得意とする機械学習アルゴリズム
  (決定的識別モデルの1つ)
- 複数の**決定木**の多数決で結果を求める

  決定木：複数条件での区分結果を表したツリー





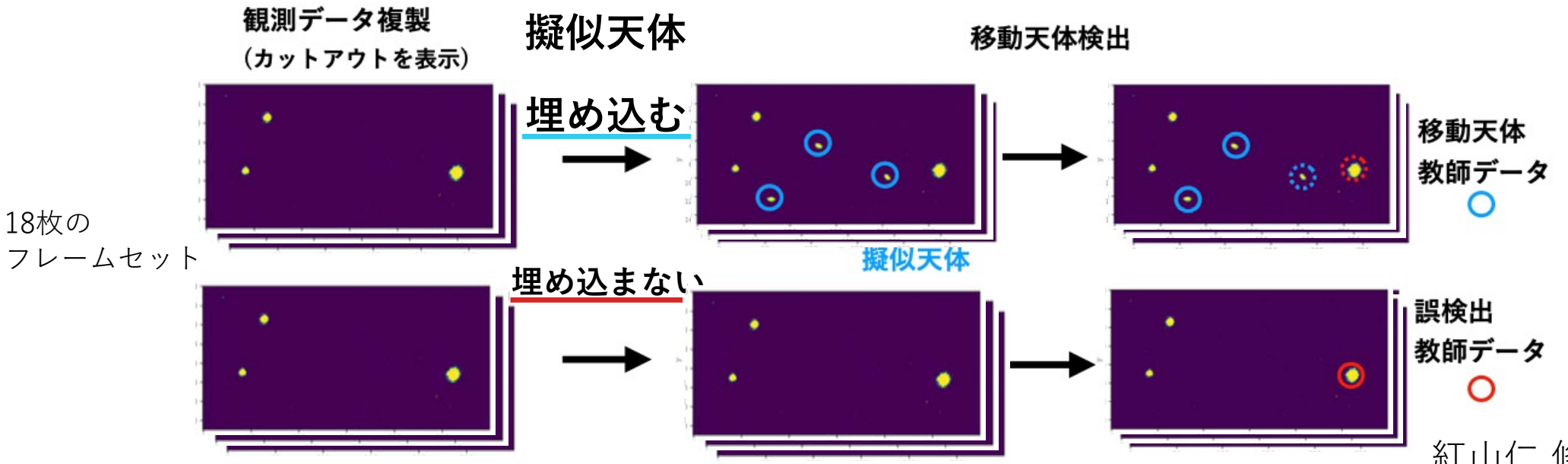
よいモデルの作成に必要なポイント

- **適切な特徴量**の設定
- **多様な学習データ**の準備

学習データ①（既存モデルで使用）

擬似移動天体の埋め込み

- ある観測データ(18枚のフレームセット)に対し擬似的に移動天体を埋め込む
- **埋め込んだ天体**を検出したとき、これを**正しい検出**とラベルをつける
- **埋め込まない**データから検出されたものは、**誤り**とラベルをつける



各6,000件の検出を学習に使用

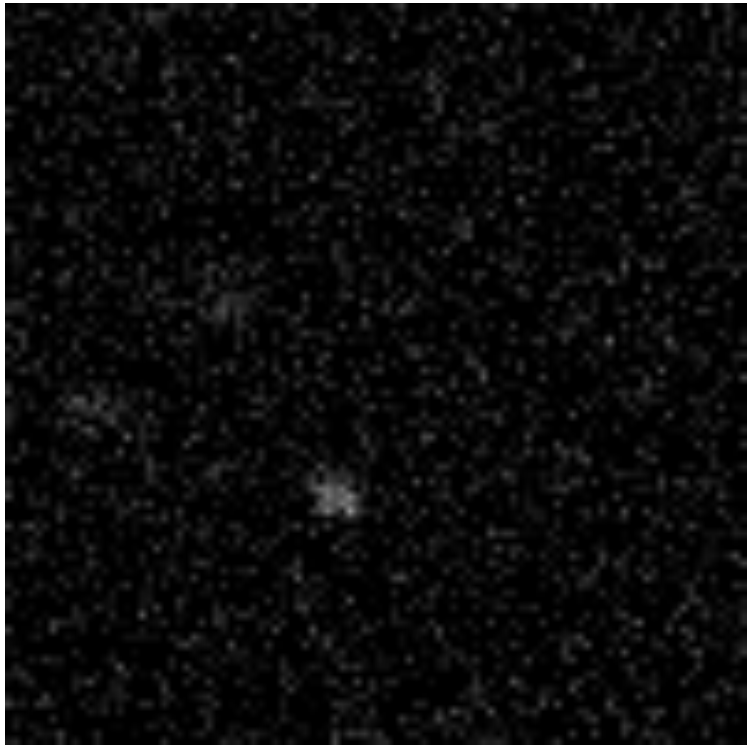
紅山仁 修士論文

特徴量①（既存モデルで使用）

14種類の特徴量を採用

- **静止画的な特徴量**
  - PSF の 2次モーメント(x, y) 2.4, 2.0
  - PSF の 共分散 -0.03
  - PSF の FWHM 3.4
  - 開口内総 flux とピークの比 22.1
  - 天体の明るさのピーク値 146.7
  - 天体の明るさの誤差 318.0
  - 背景光のカウント値 116.7
- **動画的な特徴量**
  - 全フレーム中の検出割合 1.0
  - 速度の大きさ 1.8
  - 速度の標準偏差(x, y) 0.3, 0.3
  - 速度の符号変化の割合(x, y) 0.3, 0.0

高スコア検出の例 (Score=1.0)

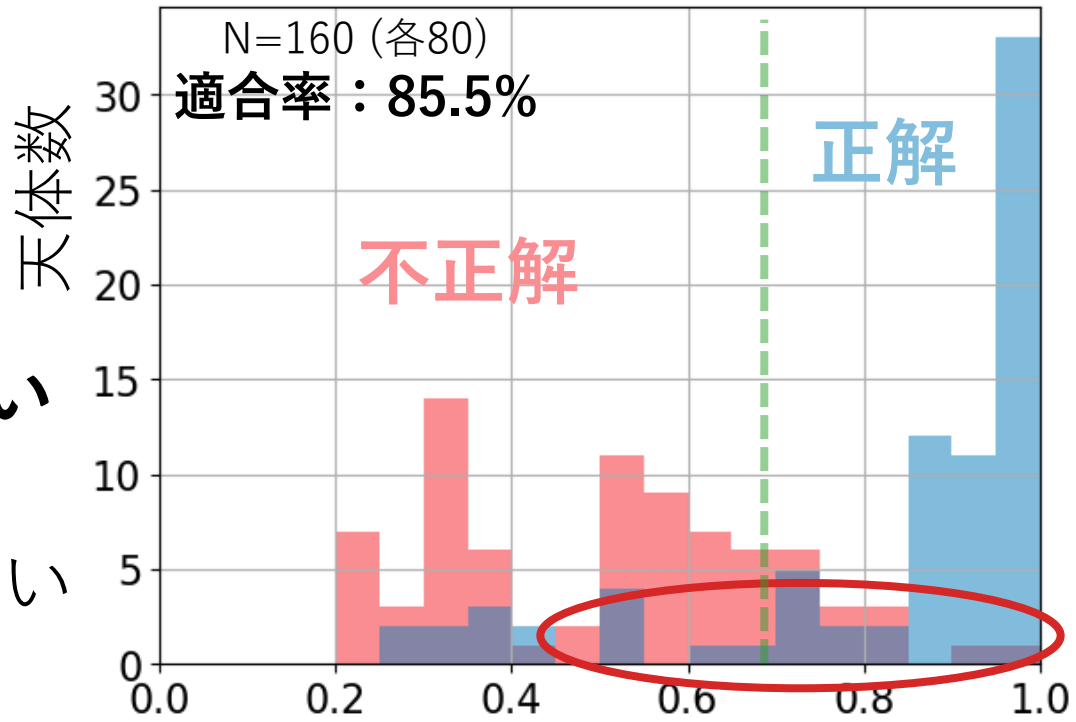


Date/Time: 2022-12-27T13:34:38.211
ID: 22Lbwvqs_05008-02143

既存モデルの課題と本研究の取り組み

- ある程度の分類は可能だが誤検出を多く生む
  - **既存モデルの適合率：85.5%**
    - 適合率の定義：TP/(TP+FP)
      正解と判断したもの(TP+FP)のうち
      正しく正解と分類できた(TP)割合
    - 分類の閾値はスコア0.7に設定
- **自動化の際に誤報率を<10%にしたい**
  → 適合率>90%が必要
- そのため本研究では新たに以下を行い
  モデルの性能向上を図った
  1. 実観測データを学習に使用
  2. 新たな特徴量を定義

※ データ削減のため
スコア0.2未満のものは
データベースに残らない



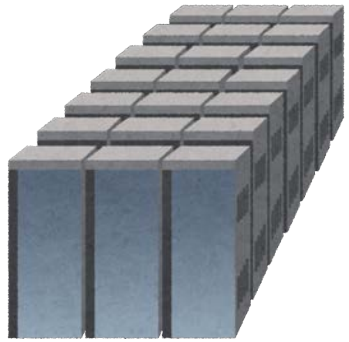
人間がラベルづけしたデータに対する
既存モデルでのスコア

Outline

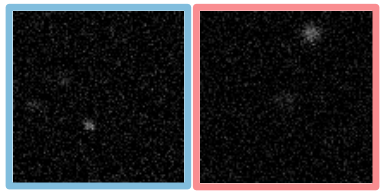
1. Introdution
2. **Method**
3. Result
4. Discussion
5. Prospect

学習データ②　本研究で新たに追加

- 既存モデルを使って**実観測データ**から検出された結果を**目視で確認し、再度ラベルづけ**した。
  これを**学習データ+テストデータとして使用**（計760検出）



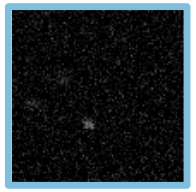
実観測データから検出された結果
正検出・誤検出ともに含む

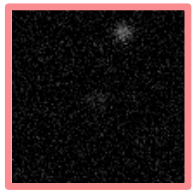


…

正解

不正解





…

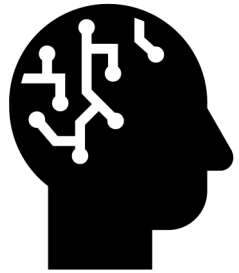
目視で確認
再度、移動天体か
ラベルづけ

600件

160件

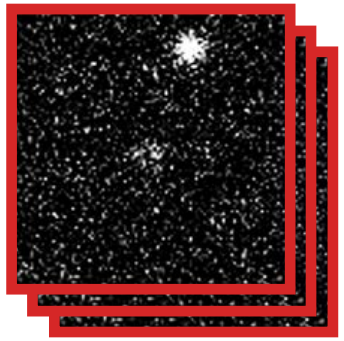
テストデータとする

新モデル

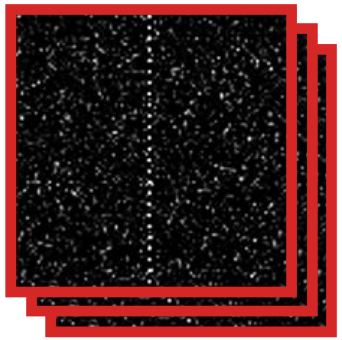


機械学習モデルの
学習データに

特徴量② 本研究で新たに追加



各フレーム間の変位の
大きさと方向が一定でない

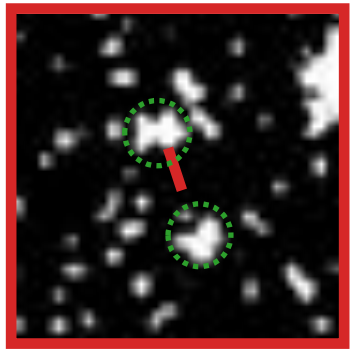


各フレーム間の変位が
特定の方向のみ

全フレームの中で
検出される座標が二極化

誤検出除去のため**新たな特徴量**を定義
特徴量の種類を増やしたモデルを作成

A) 各フレーム間の変位の大きさの分散
B) 各フレーム間の移動方向の分散
　偏角を1/2し cos をとった値を用いた
C) 重心からの最短検出距離 (下記参照)



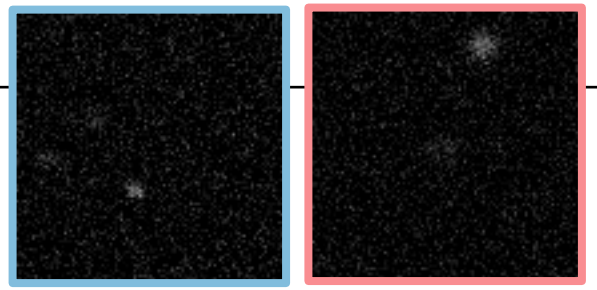
左図の赤線の長さ。計算式は
$\min(|\boldsymbol{w} - \boldsymbol{x_i}|)$
$\boldsymbol{x_i}$: i番目のフレームの検出座標
$\boldsymbol{w}$: 全てのフレームの検出の
重心座標

※2つの恒星を
移動天体として誤検出したもの

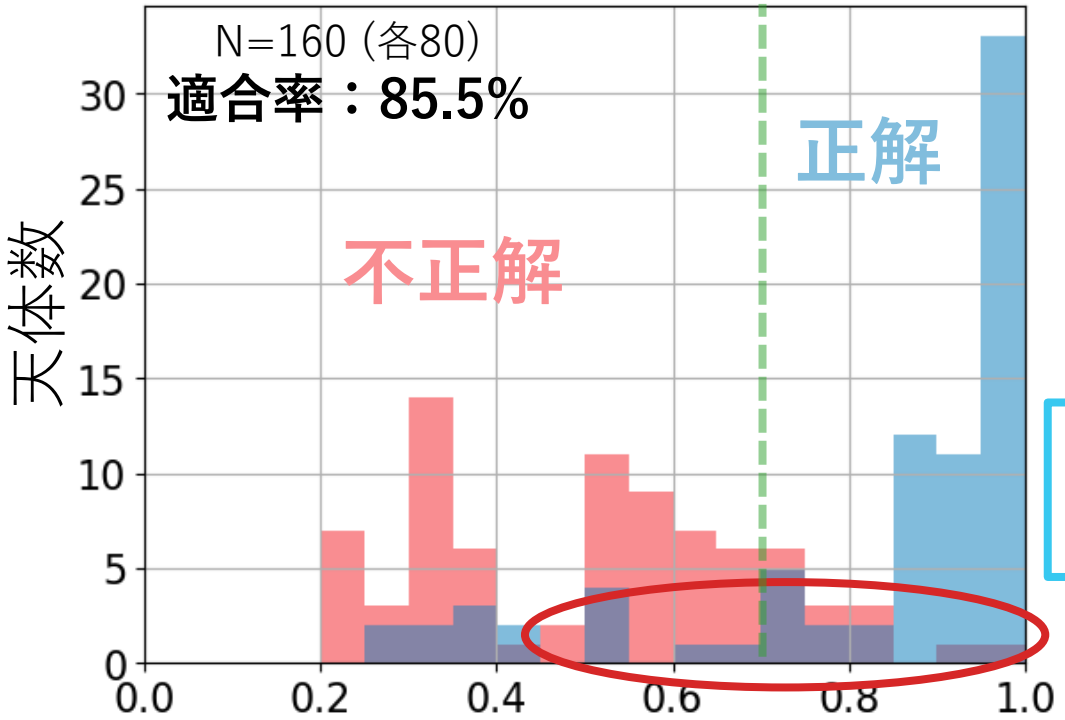
Outline

1. Introdution
2. Method
3. **Result**
4. Discussion
5. Prospect

学習データ②を追加した結果

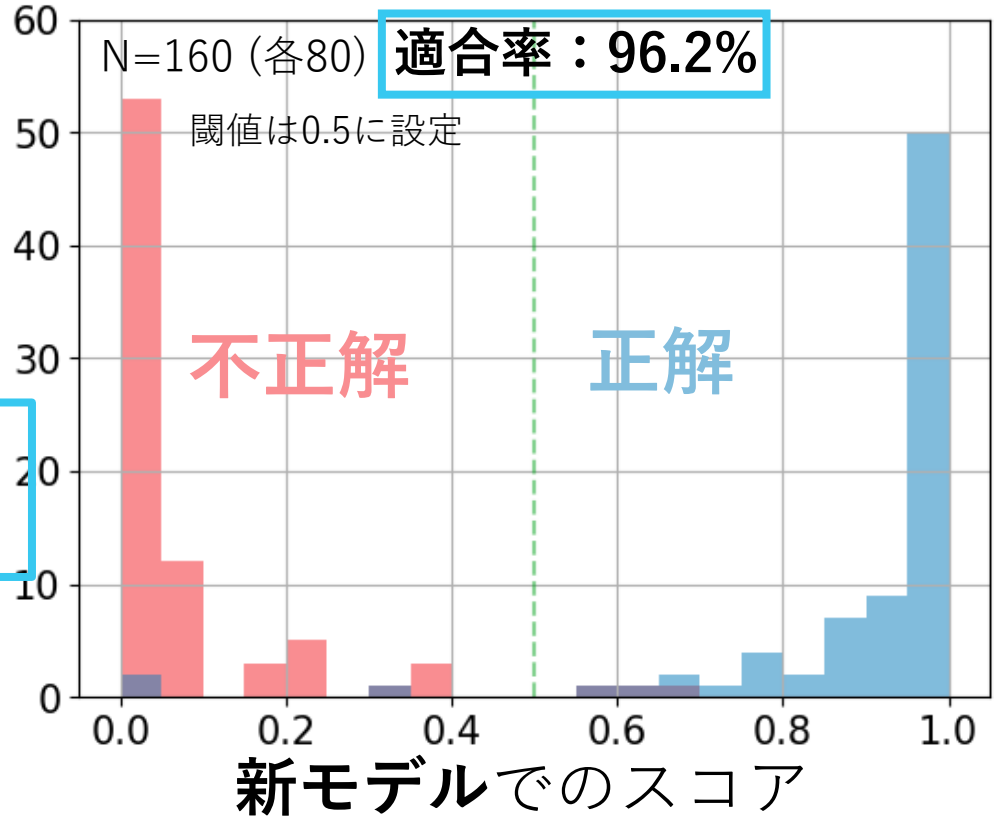


- 実観測データを学習データに追加した結果
  適合率が96.2%に上昇



※ データ削減のため
スコア0.2未満のものは
データベースに残らない

より正確な
分類に成功

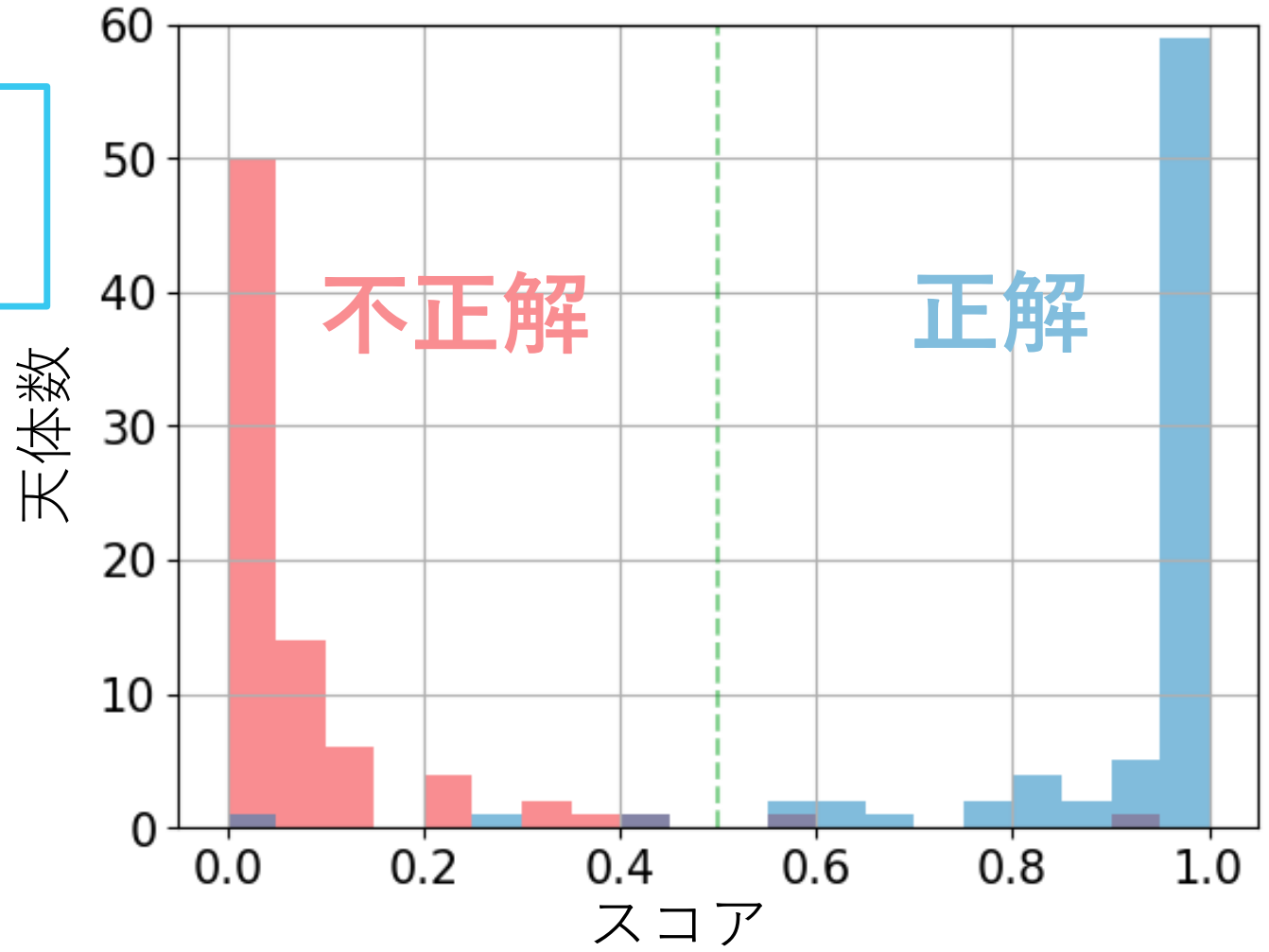


さらに特徴量②を追加した結果

- 3つの特徴量を追加した

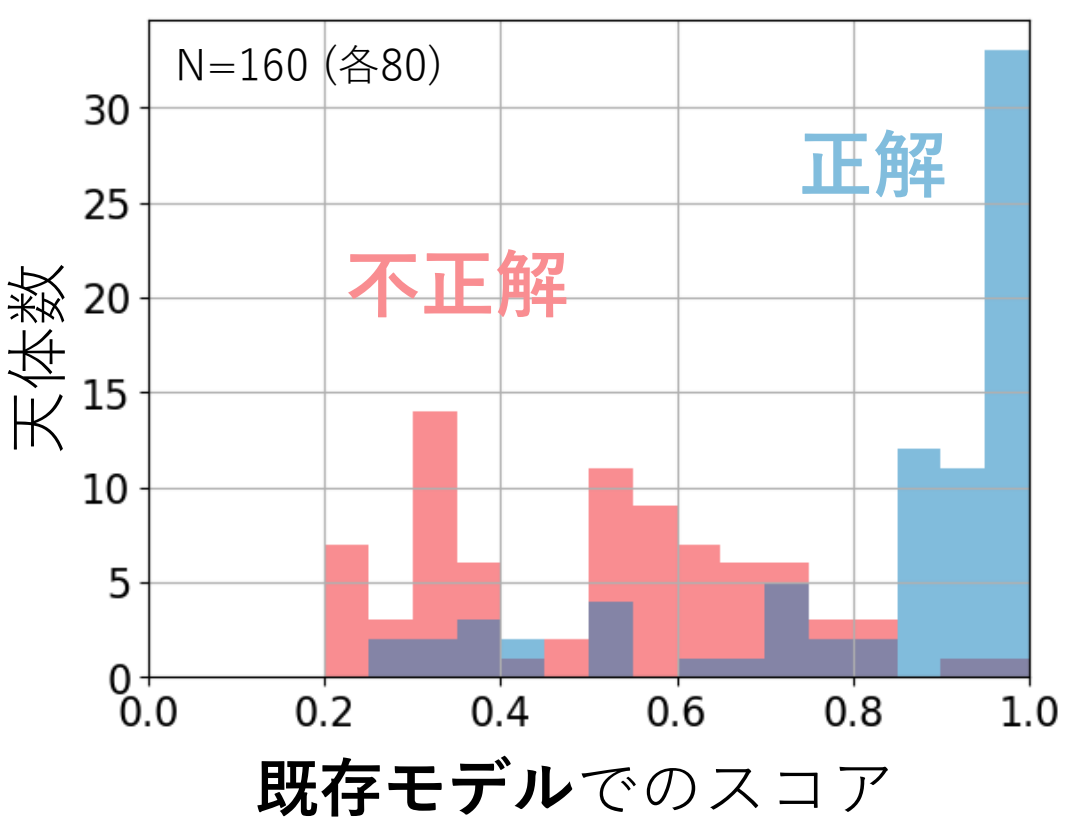
A) 各フレーム間の変位の大きさの分散
B) 各フレーム間の移動方向の分散
偏角を1/2し cos をとった値を用いた
C) 重心からの最短検出距離 (下記参照)

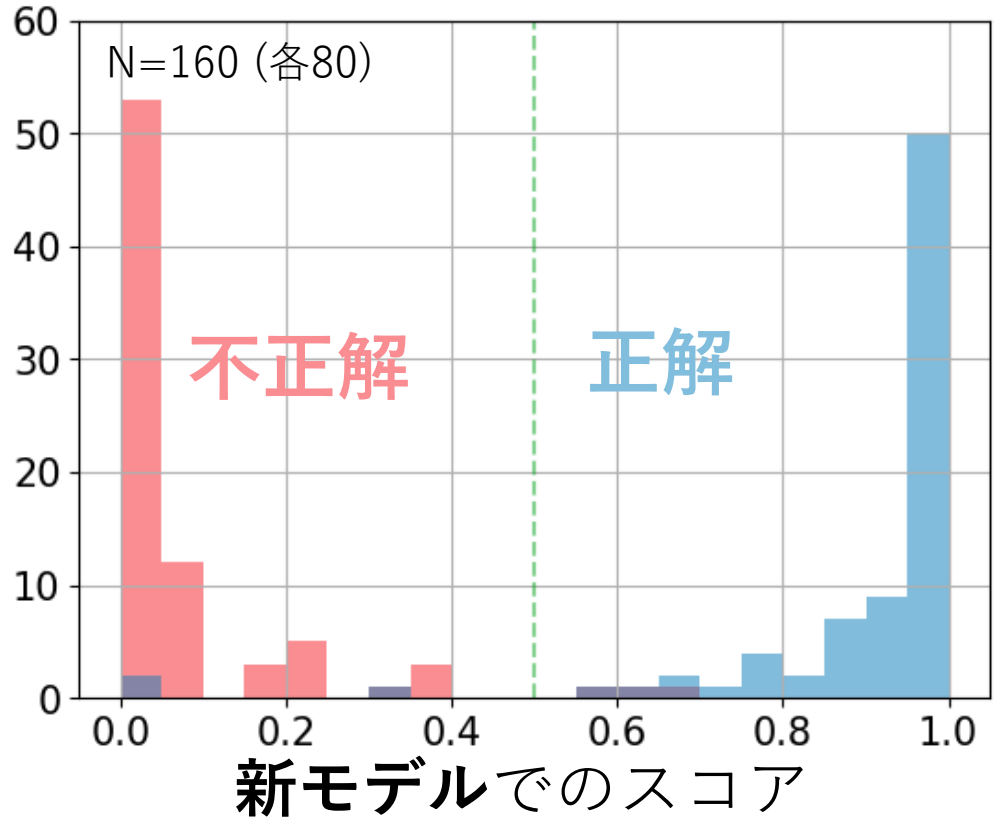
- 適合率は**97.5%**に上昇
(目標を達成できる値)



学習データ・特徴量を追加した結果

- モデルの適合率が **85.5%** → **97.5%** と上昇





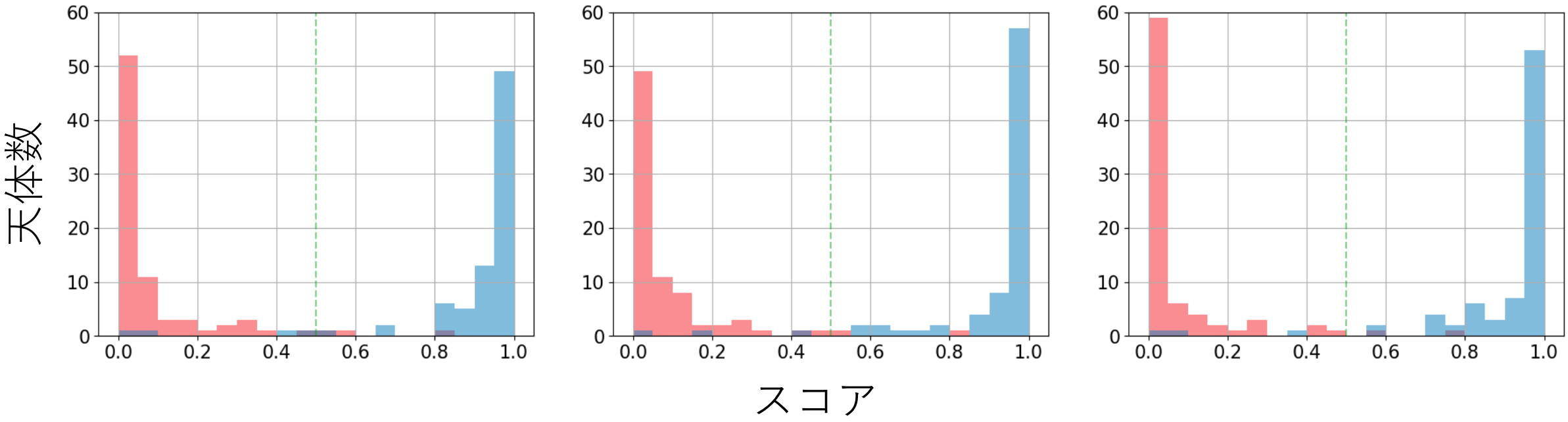
特徴量の種類ごとのモデル性能比較

いずれも効果的であることが確認できたため全て使用した

特徴量Aのみ追加
適合率：96.5%

特徴量Bのみ追加
適合率：97.5%

特徴量Cのみ追加
適合率：97.5%



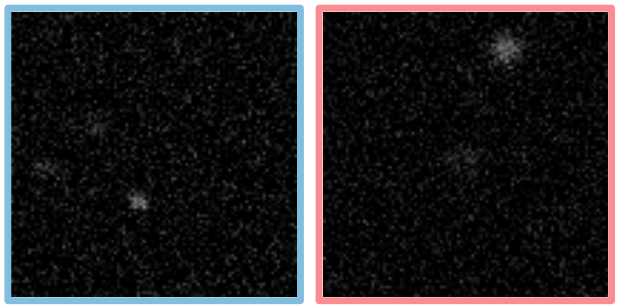
Outline

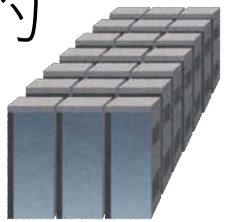
1. Introdution
2. Method
3. Result
4. **Discussion**
5. Prospect

実観測データの寄与が大きい

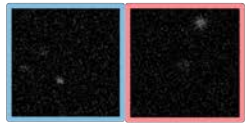
実観測データを学習に使用することがモデルの性能向上に**効果的**

- 既存モデルで用いた擬似移動天体だけでなく
  実際のデータを用いることが重要
- 既存モデルでは**除去困難な誤検出パターンを誤りと学習させる**ことによって
  誤検出のスコアを下げることが可能（次ページ）
- 特徴量の追加でも誤検出を減らすことに成功したが
  学習データの寄与が支配的

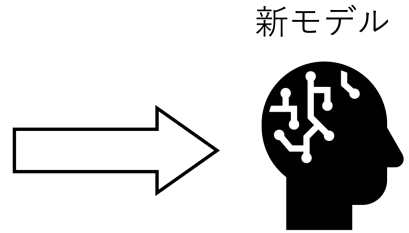




実観測データから検出された結果
正検出・誤検出ともに含む

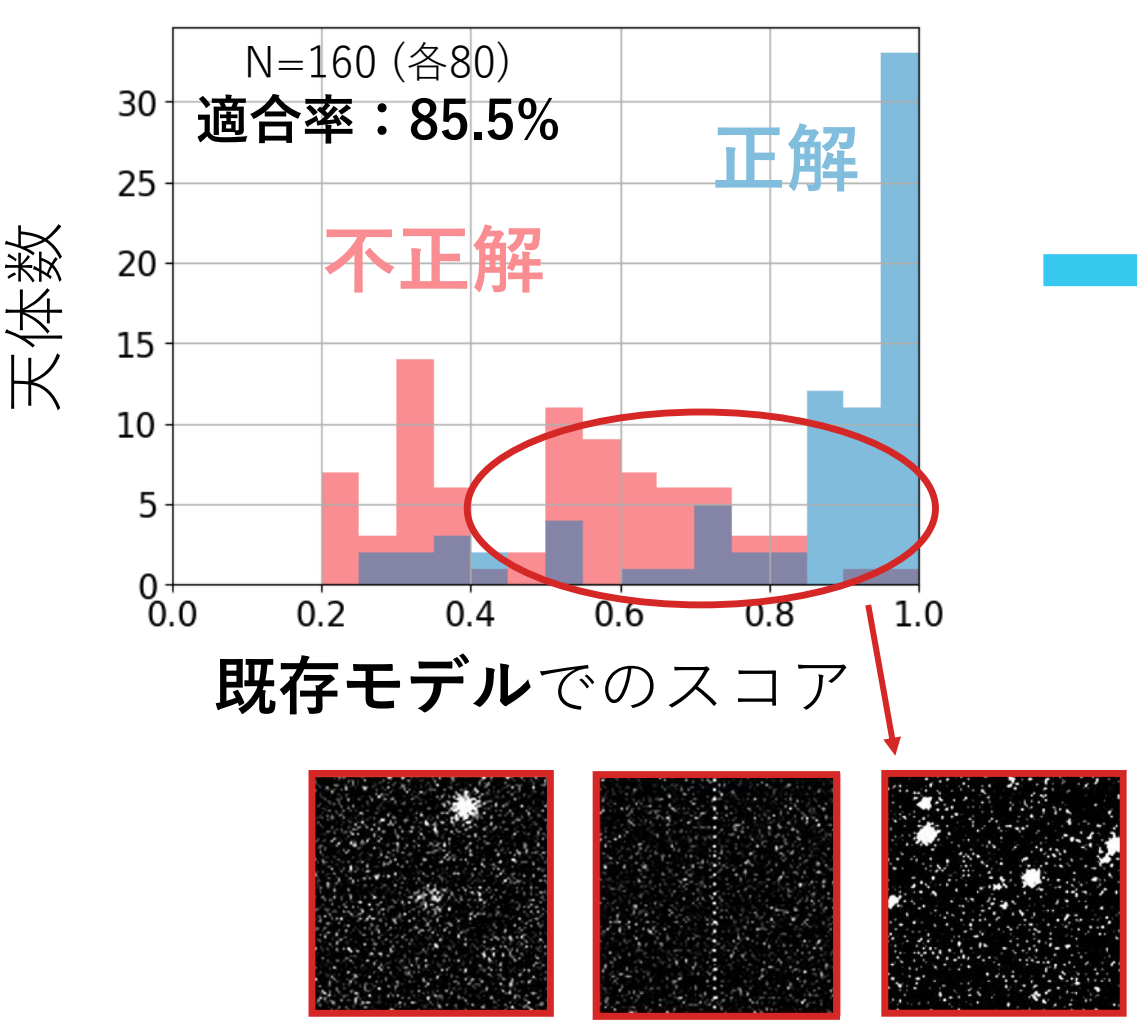


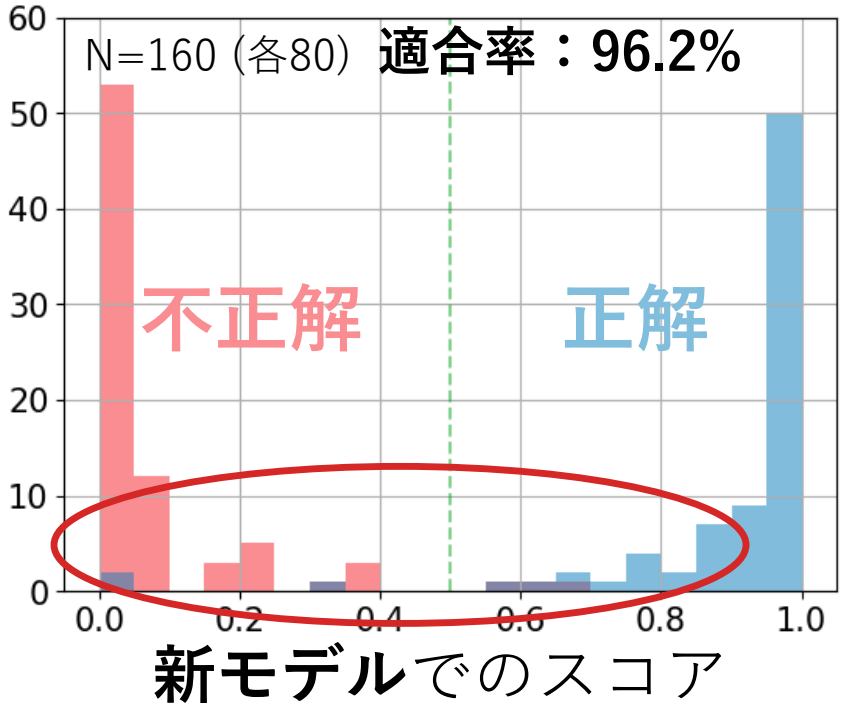
目視で確認
再度、移動天体か
ラベルづけ



機械学習モデルの
学習データに

実観測データの寄与が大きい



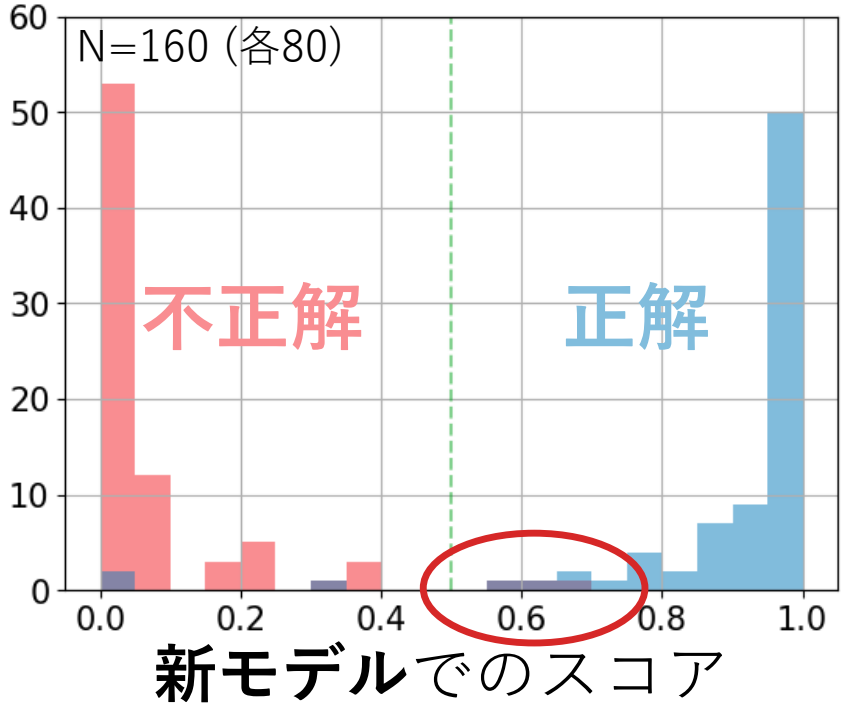


誤検出スコアを大幅に下げられた

残る誤検出

- 残る誤検出は大気の乱れや望遠鏡の揺れによるもの
  - 移動速度の制限により大部分が除去できる(2-5”/sec)
  - 移動速度が遅いものはシーイングダンスなどを考慮し
    さらなる高精度化が要求される
    (e.g. Mitsuda, K. Tomo-e Gozen 2Hz 測光データを用いた
    Seeing Danceの調査. 木曽シュミットシンポジウム2023; May 30, 2023)

- 追跡を行うのは1晩あたり3天体まで
  → 高スコア天体から優先的に追跡すれば
    実用上はほぼ問題ない

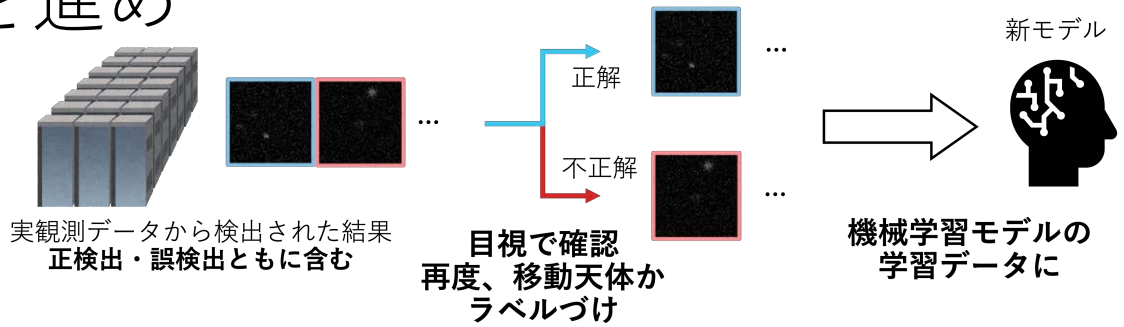


Outline

1. Introdution
2. Method
3. Result
4. Discussion
5. **Prospect**

課題と展望

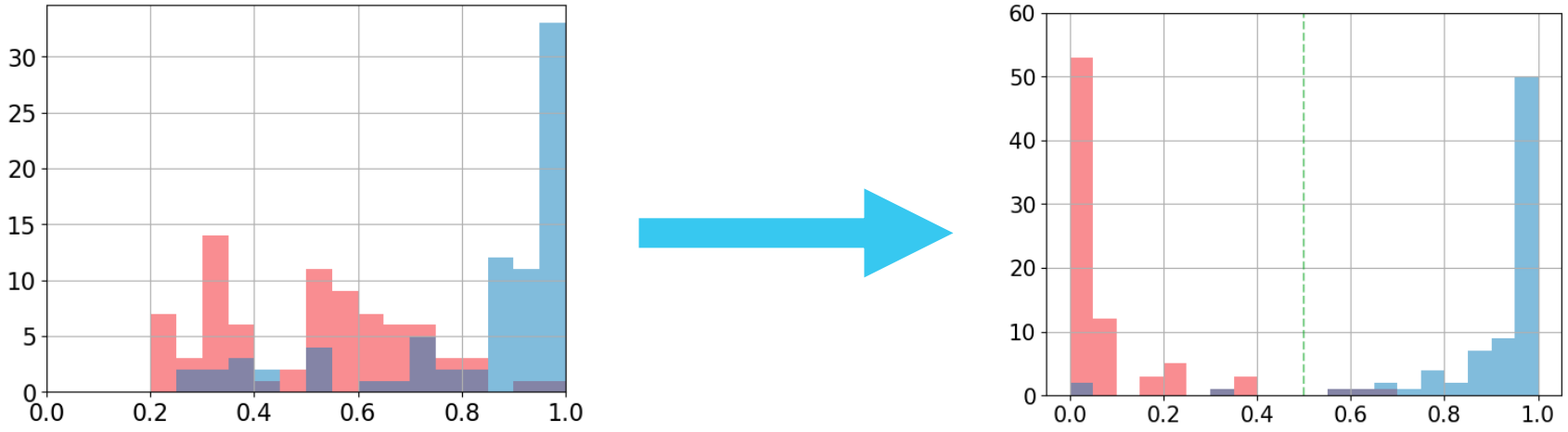
- **検証にはより多くのデータセットが必要**
  → 実観測データへのラベルづけを進め
  データセットを増やしていく



- モデル性能の目標は達成した
  → **今後は自動化のツールを準備する**

機械学習モデルのアップデート
→ 検出の精度改善

フォローアップの自動化
→ 発見レートの向上

まとめ

- Tomo-e Gozen **移動天体検出システム**に利用されてきた機械学習モデルには**誤検出が多かったため新たなモデルを作成**した
- **実観測データ**を学習データに追加し**新たな特徴量を定義**したところ**適合率が 85.5% → 97.5% と改善し、目標の性能を達成した**
- モデルの性能向上には**実観測データの利用が効果的**だと確認した
- 今後は**自動化のツールを準備**していく方針である