

Improvement of the deep-learning real/bogus classifier in the Tomo-e Gozen transient survey

高橋一郎

東北大学

2021年10月5日

目次

酒向さん発表より

あの頃、われわれは何を考えていたのか？ (つづき)

シュミットシンポジウム2019 シュミットシンポジウム2021

装置

- SINETへの接続構想, データによる産学連携と教育普及 (酒向)
- ライトカーブの異常検知法 (山本)

全天サーベイ

- サーベイ経路の最適化 (諸隈, 池田, Pedro) : いずれ雲を避けたい
- CNN機械学習による超新星の検出 (富永, 浜崎) : 誤検出をあと1桁落としたい
- 重力波追観測の進捗 (新納) : 6件を実施

新規提案 (連携)

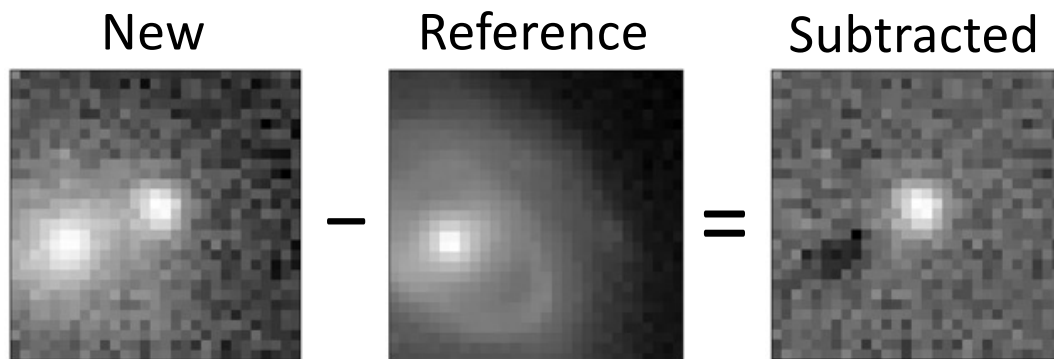
まげろ (MAG) との超新星観測の連携 (酒向)

The diagram shows a list of research topics on the left and names in blue boxes on the right. Lines connect the topics to the names: 'SINETへの接続構想...' connects to '満田'; 'ライトカーブの異常検知法...' connects to '瀧田'; 'サーベイ経路の最適化...' connects to '山本'; 'CNN機械学習による超新星の検出...' connects to '津々木'; '重力波追観測の進捗...' connects to '高橋'; and 'まげろ (MAG) との超新星観測の連携...' connects to '富永'.

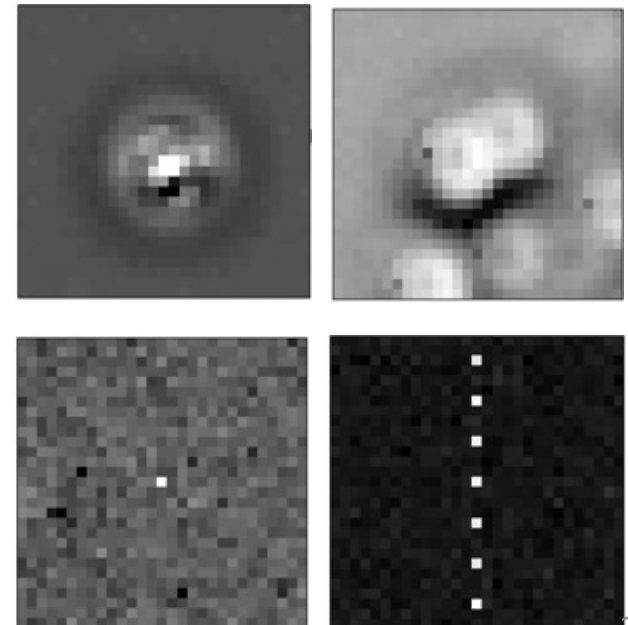
- Introduction
- Cleaning mislabel in training data
- Evaluation with actual test dataset
- Implementation of new model

Transient detection

- 差分を取ることでtransientを検出
- ReferenceにはPS1の画像を使用
 - 現在Tomo-e referenceへ切り替え中
- ゴミ (Bogus) が多い ($10^6/\text{night}$) ので機械学習で分類
 - Real or Bogus



Bogus examples



Previous Real/Bogus Classification

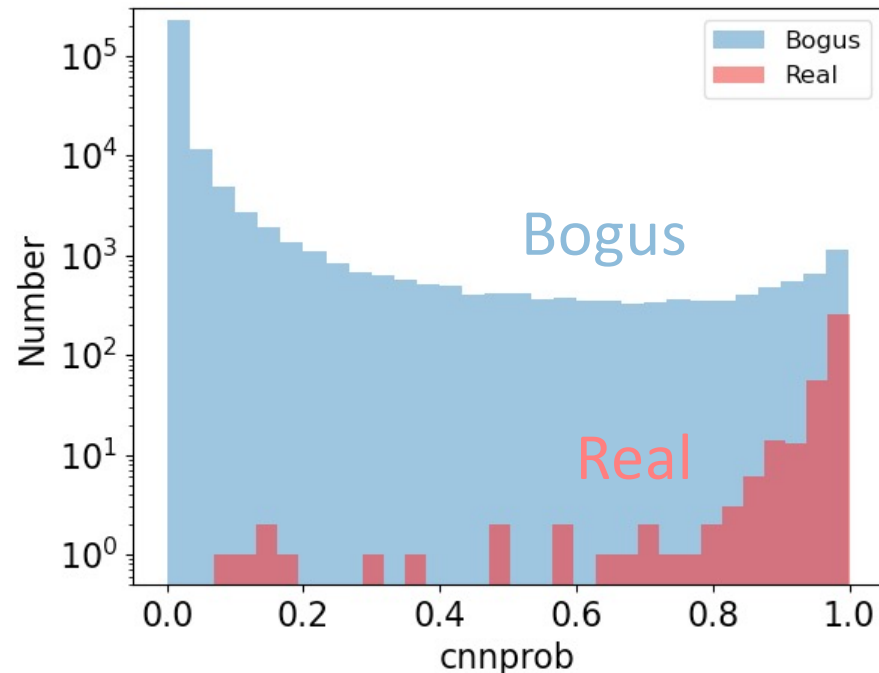
- モデルはシンプルなCNN
 - 入力: 29x29 image x 3 (ref., new and sub.)
- 学習データ
 - Real: シミュレートした星を埋め込み
 - Bogus: 実際のもの
 - 数百万サンプル

従来CNNモデルの構造 (浜崎修論)

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 25, 25, 32)	2432
max_pooling2d_1 (MaxPooling2D)	(None, 12, 12, 32)	0
conv2d_2 (Conv2D)	(None, 10, 10, 64)	18496
max_pooling2d_2 (MaxPooling2D)	(None, 5, 5, 64)	0
dropout_1 (Dropout)	(None, 5, 5, 64)	0
flatten_1 (Flatten)	(None, 1600)	0
dense_1 (Dense)	(None, 128)	204928
dense_2 (Dense)	(None, 64)	8256
dense_3 (Dense)	(None, 1)	65

Total params: 234,177
Trainable params: 234,177
Non-trainable params: 0

従来CNNの出力分布

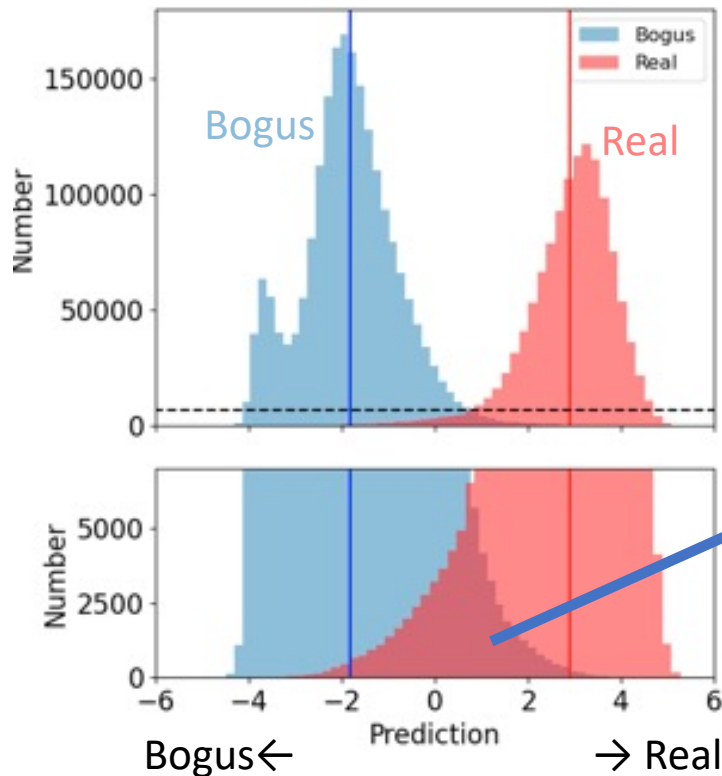


- 分類成績が頭打ちに
- RealがBogusに埋もれる
- 追観測のターゲット選定の障害

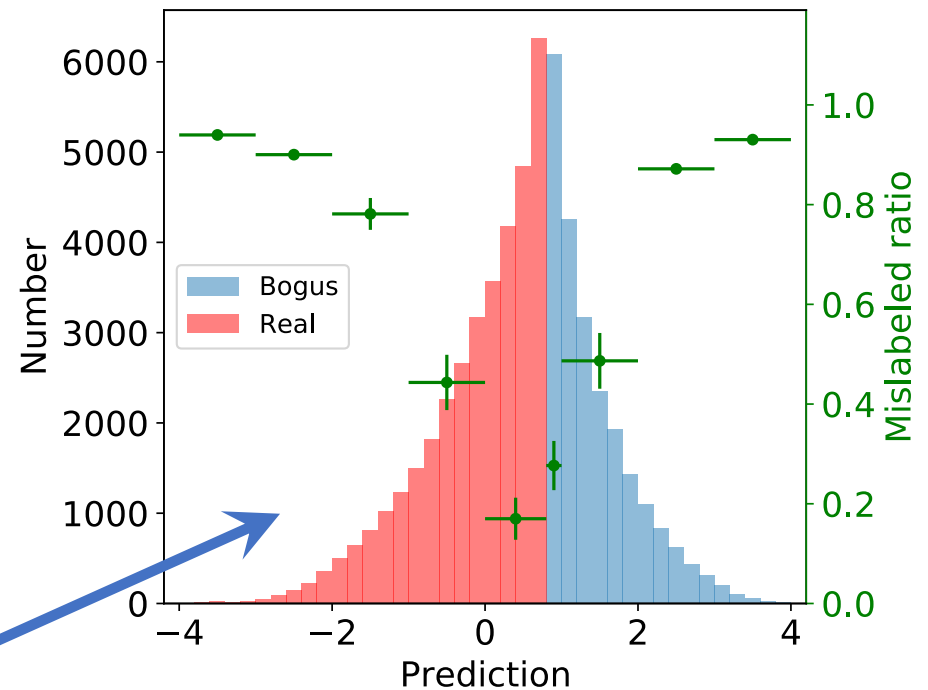
Cleaning mislabel in training data

- 学習データにラベル間違いが存在
- 学習データ自体をCNNで分類
 - 誤分類しているものを標本調査
 - ラベル間違いの比率を計算

学習データの出力分布




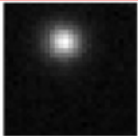


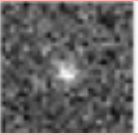
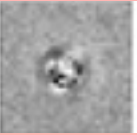
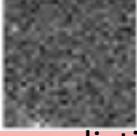





後分類の分布と誤ラベルの比率

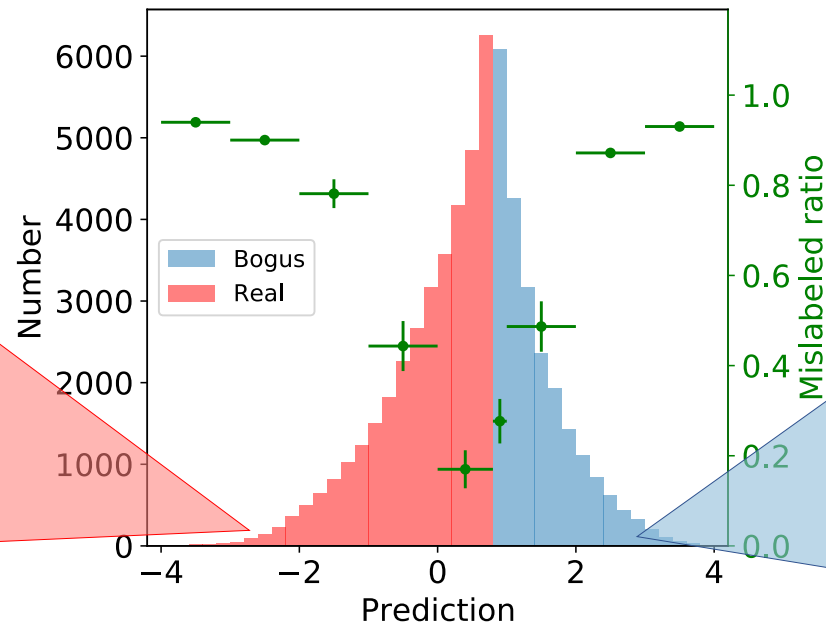


裾野に行けば行くほど誤ラベルが増える

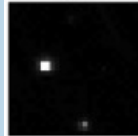




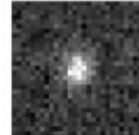

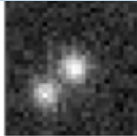
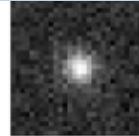
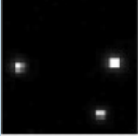


Examples of mislabel

Label : Real

Ref	New	Sub
		
prediction: -2.19		
		
prediction: -3.20		
		
prediction: -3.32		
		
prediction: -3.81		



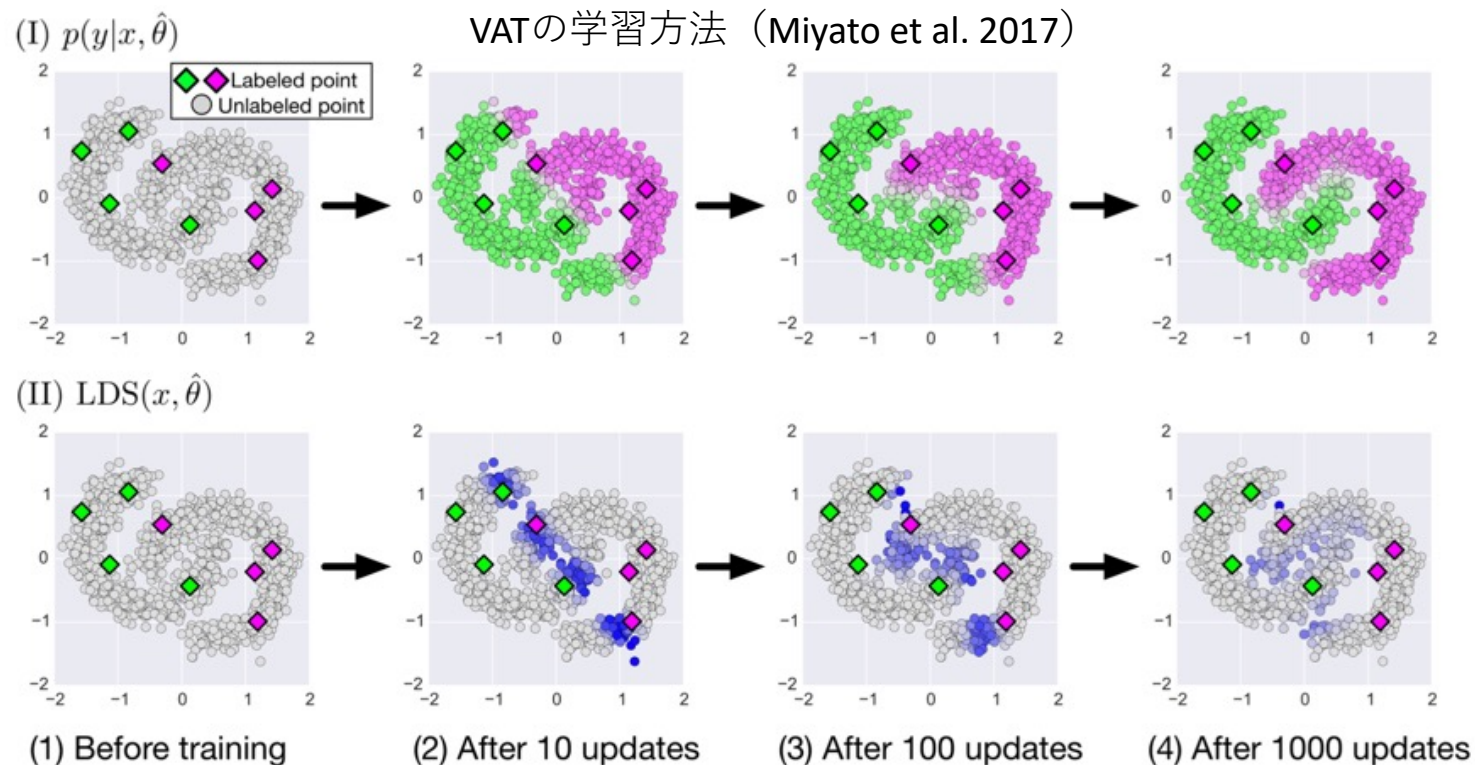
Label : Bogus

Ref	New	Sub
		
prediction: 2.46		
		
prediction: 2.95		
		
prediction: 3.43		
		
prediction: 3.93		

- 機械は正しく分類している
- ラベル間違いの混入率
 - Bogusの0.6%
 - Realの1.0%
- CNNで誤分類しているものを「ラベルなし」にして半教師あり学習を行う
 - Virtual Adversarial Training (VAT, Miyato et al. 2016)

Virtual Adversarial Training (VAT)

- 敵対的学習の一種
 - 分類器が最も誤りやすいノイズを付加したデータで学習 -> ノイズ耐性
- 予測分布が最も大きく変わる方向へノイズを加え、ノイズの付加前後で予測分布の差 (LDS) が小さくなるように学習
 - Local Distributional Smoothness (LDS)
 - LDS 小 -> 境界がデータ集合から遠くなる-> よりロバスト
 - LDSの計算には正解ラベル情報がいらぬ -> 半教師あり学習が可能
- 詳しくはMiyato et al. 2016, 2017を参照

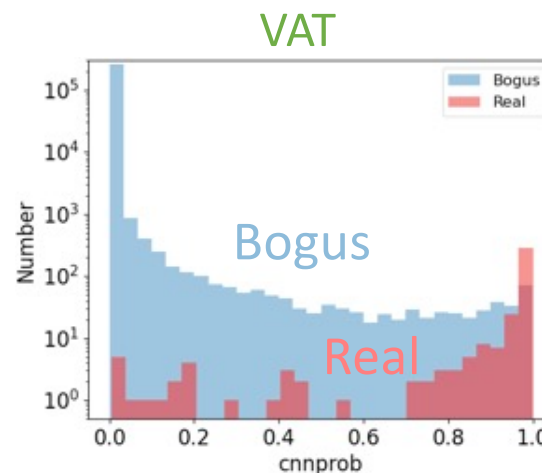
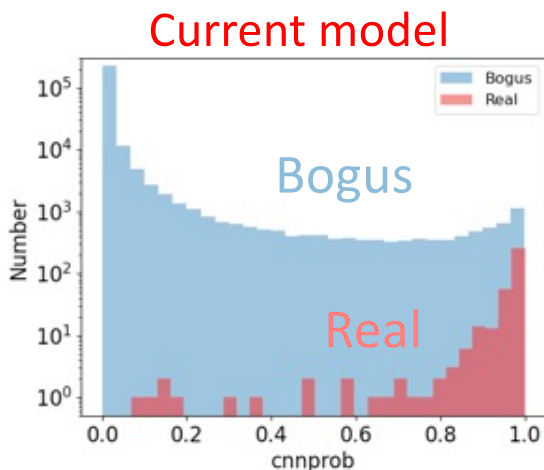


Evaluation with actual test dataset

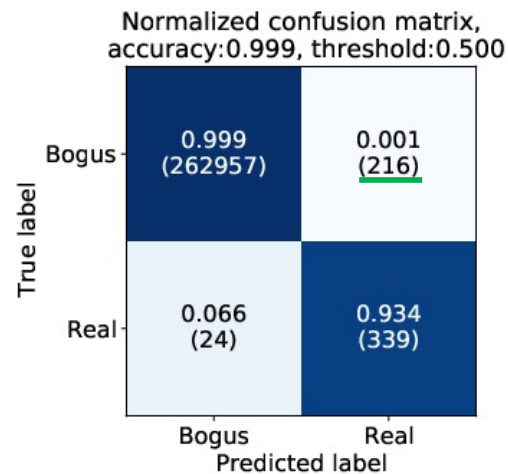
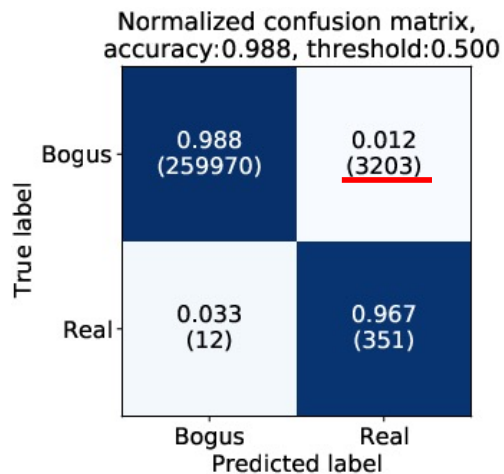
※しきい値: probability=0.5

	accuracy	precision	recall	f1 score
Current model	0.974	0.050	0.981	0.095
VAT	0.998	0.436	0.942	0.596

出力分布

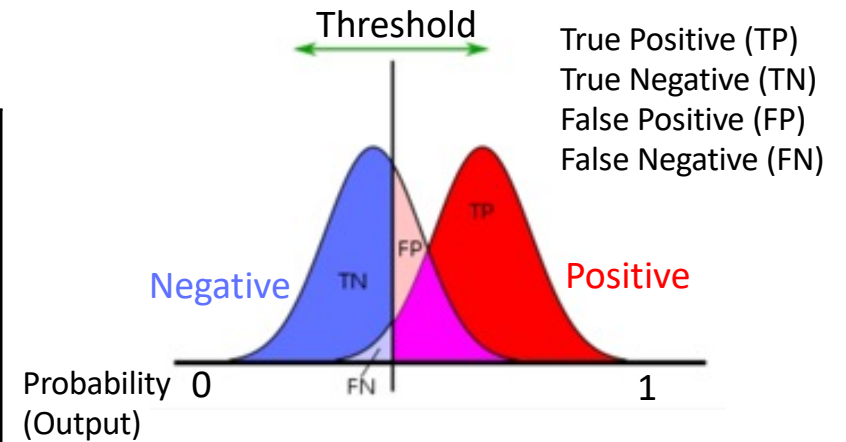
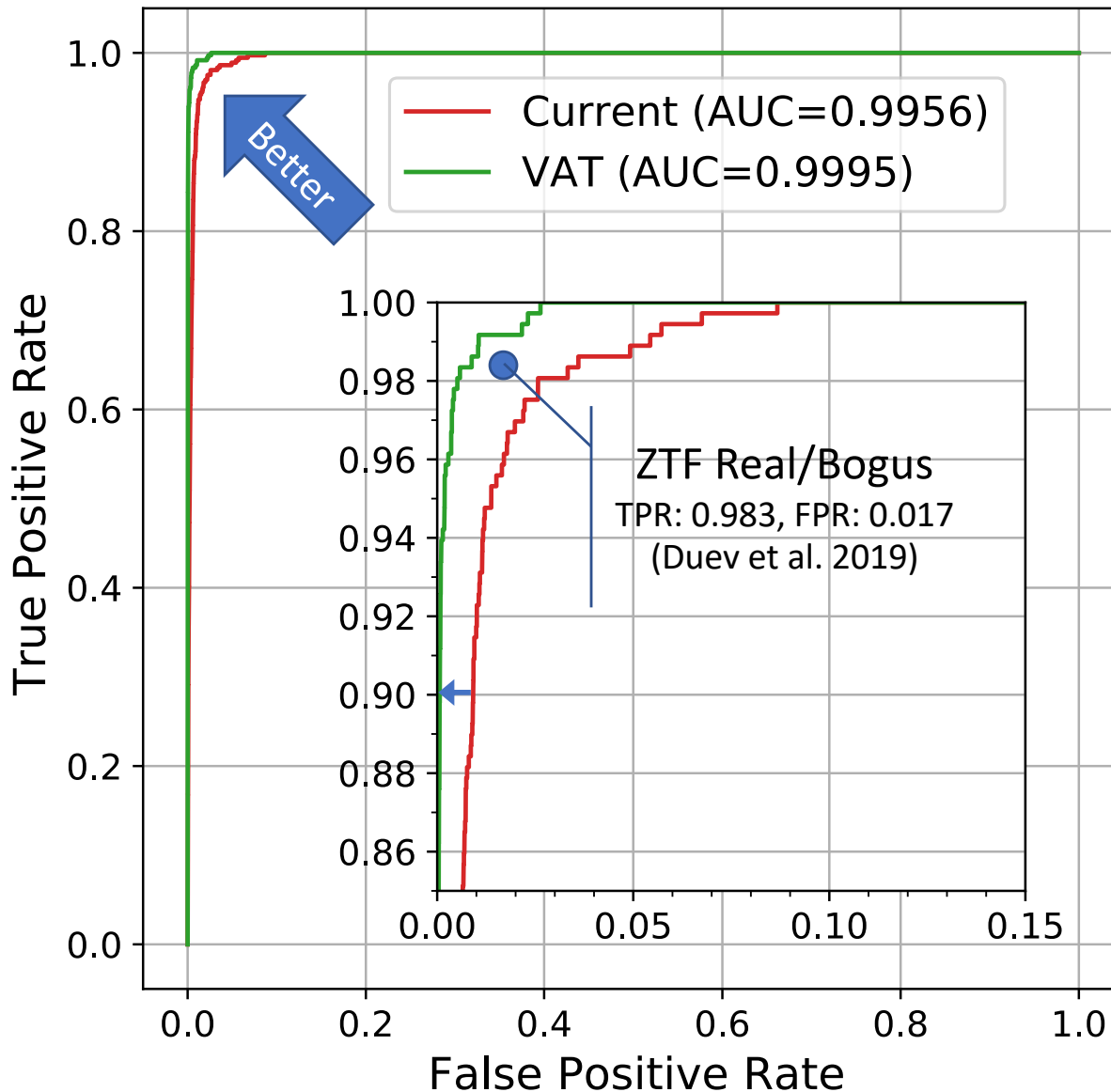


Confusion matrix



Evaluation with actual test dataset

ROC curve



Confusion matrix

	Bogus	Real
Actual Bogus	TN	FP
Actual Real	FN	TP
	Bogus	Real
	Predicted	

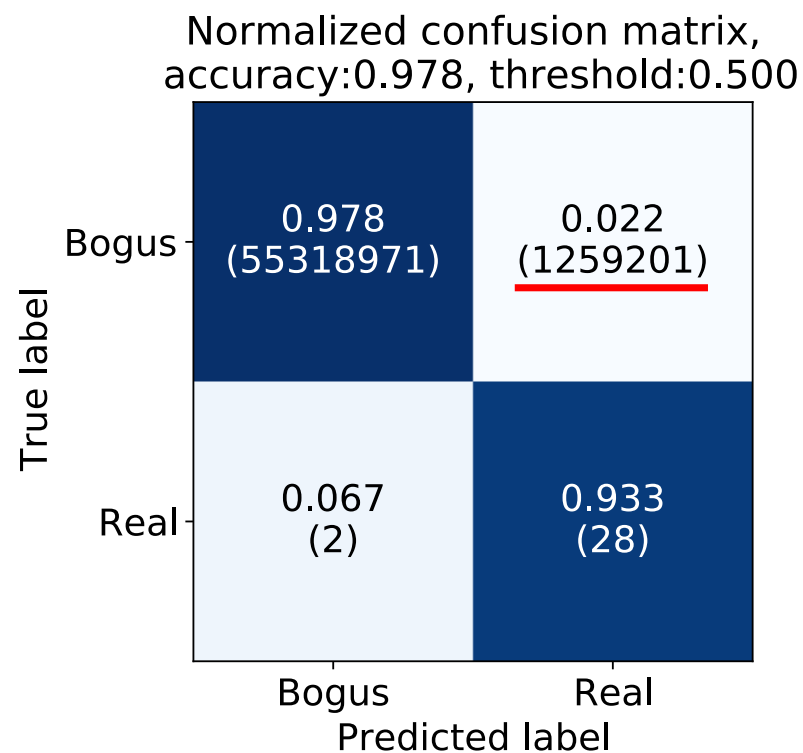
$TPR = TP / (TP + FN)$
 $FPR = FP / (TN + FP)$

model	AUC(ROC)	FPR@TPR=0.9
Current	0.9956	0.00916
VAT	0.9995	0.00065

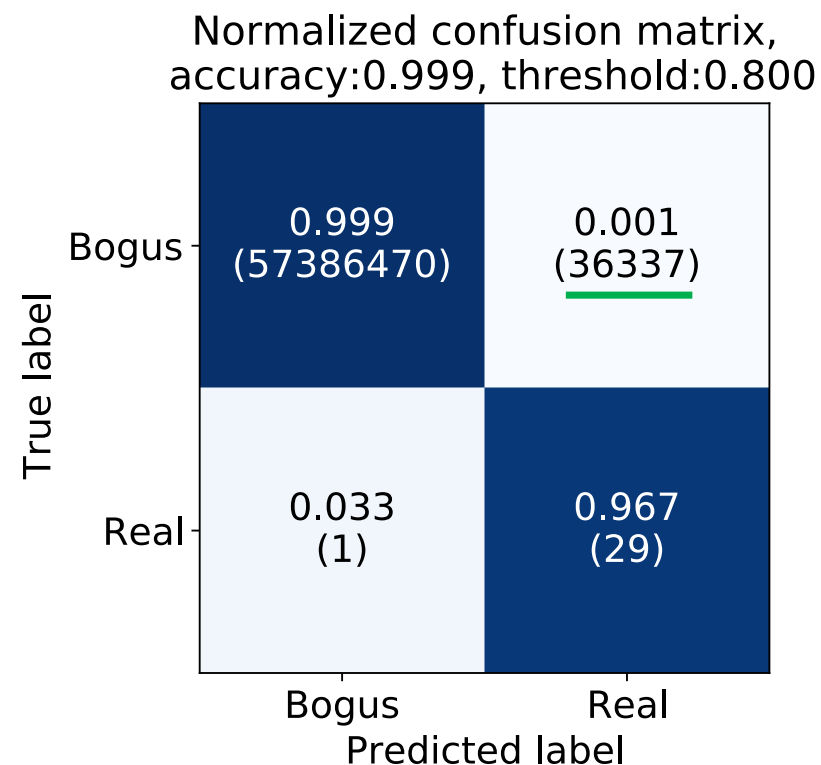
Implementation of new model

- 5月末にパイプラインに新モデルを実装
- 実装後の分類成績
 - 1週間の中に検出された候補（約6000万サンプル）

Simple CNN



VAT



Misclassified samples

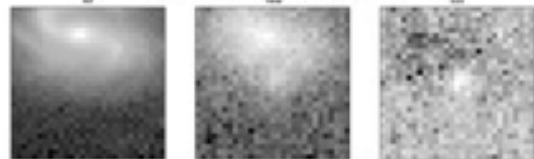
- 誤分類しているものをチェック
 - FNの1つとFPの~1%

Normalized confusion matrix,
accuracy:0.999, threshold:0.800

True label	Bogus	0.999 (57386470)	0.001 (36337)
	Real	0.033 (1)	0.967 (29)
		Bogus	Real
		Predicted label	

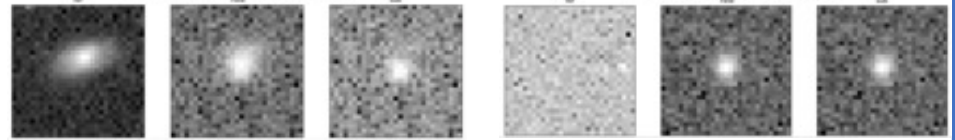
False Negative (1)

Low S/N, Large galaxy (1)

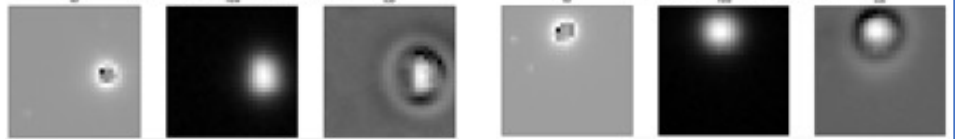


False Positive (341/36337)

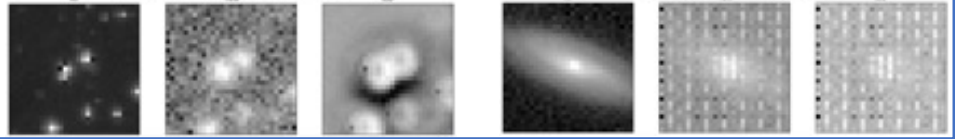
Real? (106)



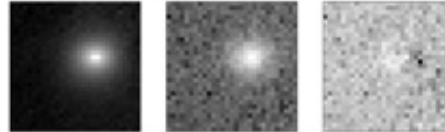
Mask in ref (65)



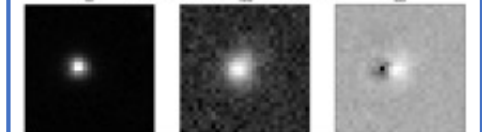
Bogus (64)



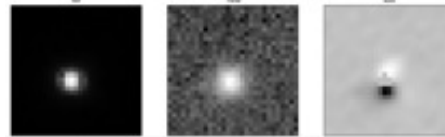
Low S/N (43)



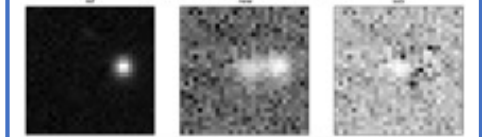
Moving? (29)



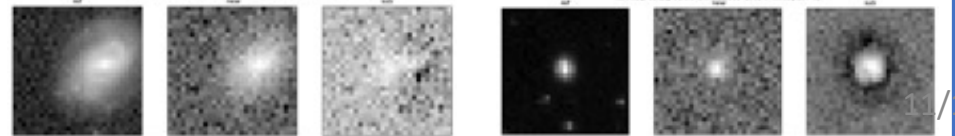
Mask in sub (6)



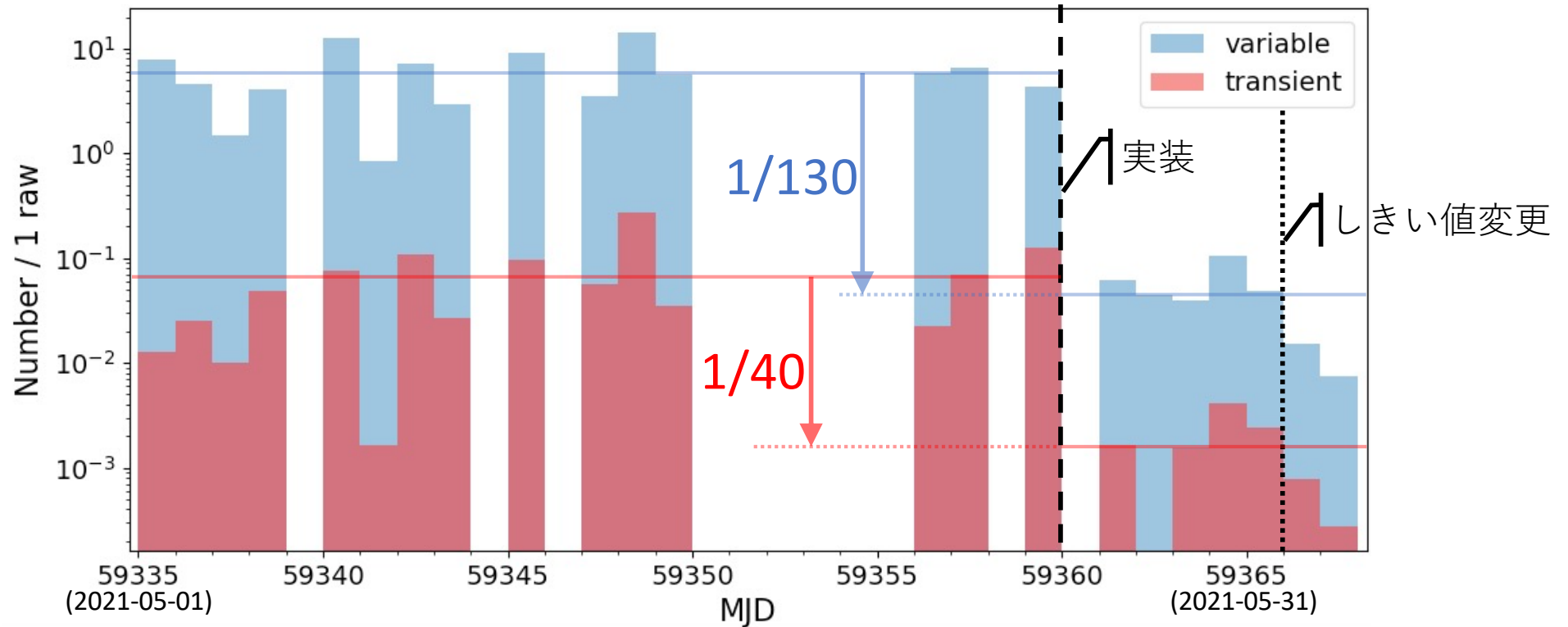
Tracking error (1)



Unclear (27)



Number of transient candidates registered in DB



		実装前	実装後
本物 (TNS天体) の 確率	variable	1 / 640494	1 / 1004
	transient	1 / 8738	1 / 90

transient候補が6000天体/dayから150天体/day (人がチェックできるレベル) へ

まとめと今後の予定

- 1/100のラベル間違いを機械が修正することで
Tomo-e のReal/Bogus分類性能が100倍になった
 - 前回シンポジウムの課題はクリア
- reference更新後にモデルの更新
 - 学習データの作り直しから
 - FPRがさらに減ることが期待される
 - 既に試作referenceで更新手順を確認済み

