

機械学習によるTransient検出の 現状と展望

甲南大学

M2 浜崎 凌

• 目次

1. 機械学習によるTransient天体の検出とは？

1. Transient天体の数
2. Transient天体の検出手順
3. Modelの学習
4. Modelの評価について

2. 学習結果

1. 学習データ
2. ROC curve
3. False Positive の S/N について

3. 結論と展望

1. 今後の展望
2. まとめ

• 目次

1. 機械学習によるTransient天体の検出とは？

1. Transient天体の数
2. Transient天体の検出手順
3. Modelの学習
4. Modelの評価について

2. 学習結果

1. 学習データ
2. ROC curve
3. False Positive の S/N について

3. 結論と展望

1. 今後の展望
2. まとめ

1. 機械学習によるTransient天体の検出とは？

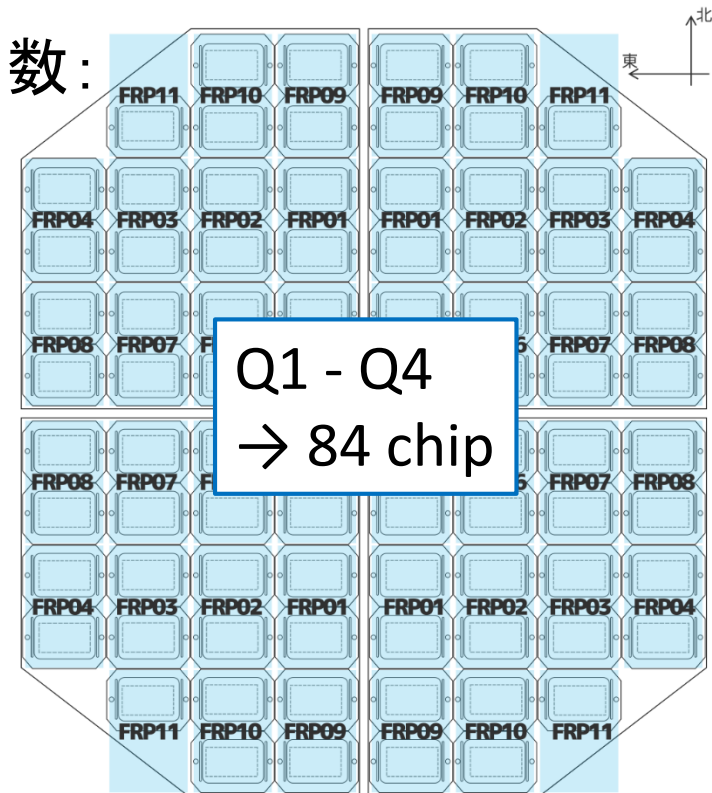
1. Transient天体の数

Q1 - Q4 で撮影、Supernova surveyで

検出されるTransient候補の数: 20000 天体 (in 1 night)

Supernova survey で見つかる超新星の数:

数個 (in 1 night)



1. 機械学習によるTransient天体の検出とは？

1. Transient天体の数

Q1 - Q4 で撮影、Supernova surveyで

検出されるTransient天体のほとんどが本 (in 1 night)

Supernova survey

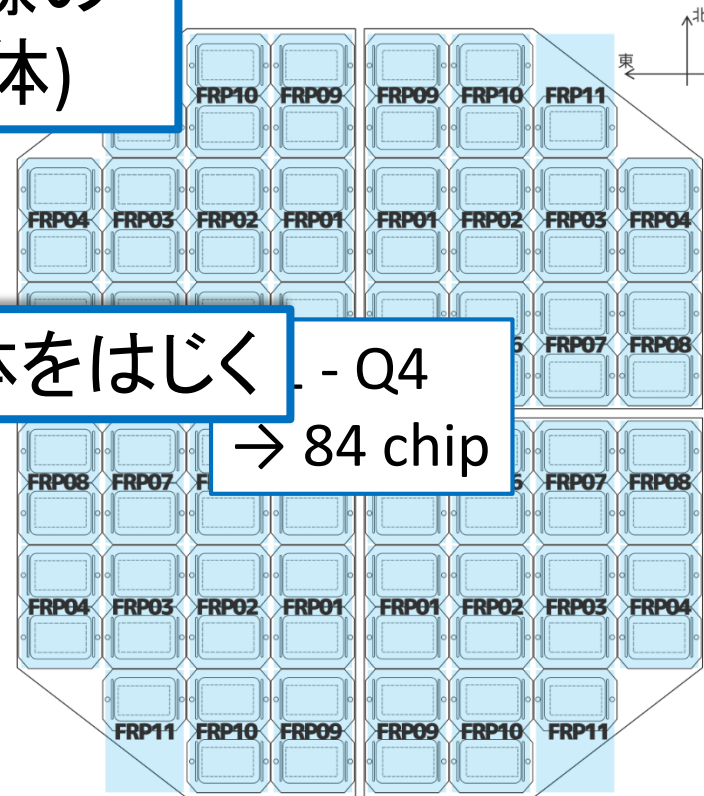
20000天体のほとんどが
エラー天体や宇宙線の
イベント (Artifact天体)

数個 (in 1 night)



画像認識でArtifact天体をはじく - Q4

→ 84 chip



1. 機械学習によるTransient天体の検出とは？

2. Transient天体の検出手順

同じ領域を別時刻で撮った天体画像について

それぞれの画像の位置合わせ



画像同士で引き算

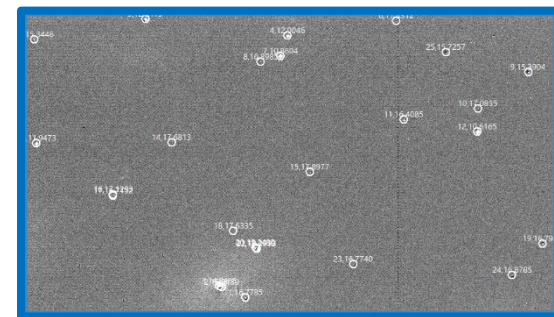
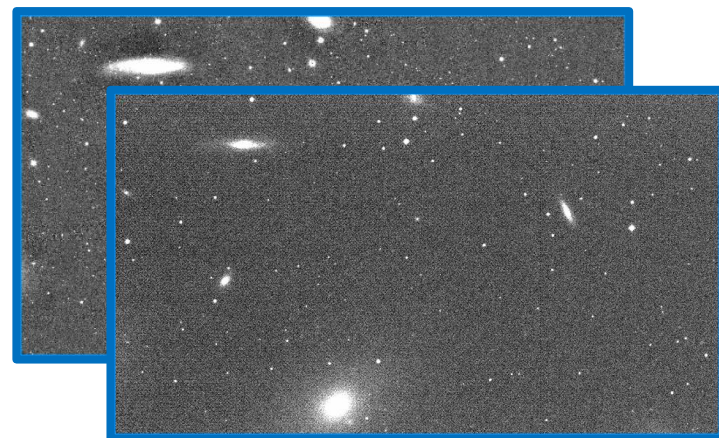
その時刻間で変化した天体のみが写った差分画像



Transient候補天体を検出



機械学習を用いて候補からTransient天体を取り出す



1. 機械学習によるTransient天体の検出とは？

2. Transient天体の検出手順

同じ領域を別時刻で撮った天体画像について

それぞれの画像の位置合わせ



画像同士で引き算

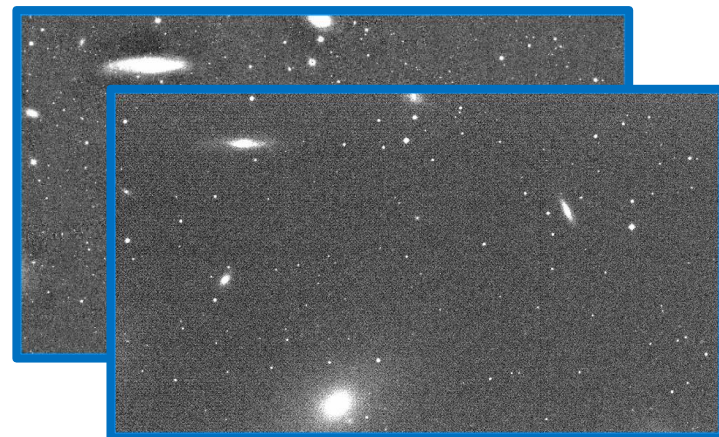
その時刻間で



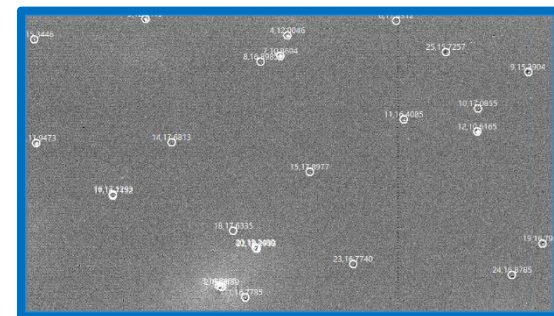
Transient候補天体を検出



機械学習を用いて候補からTransient天体を取り出す



学習器 (Model) の学習

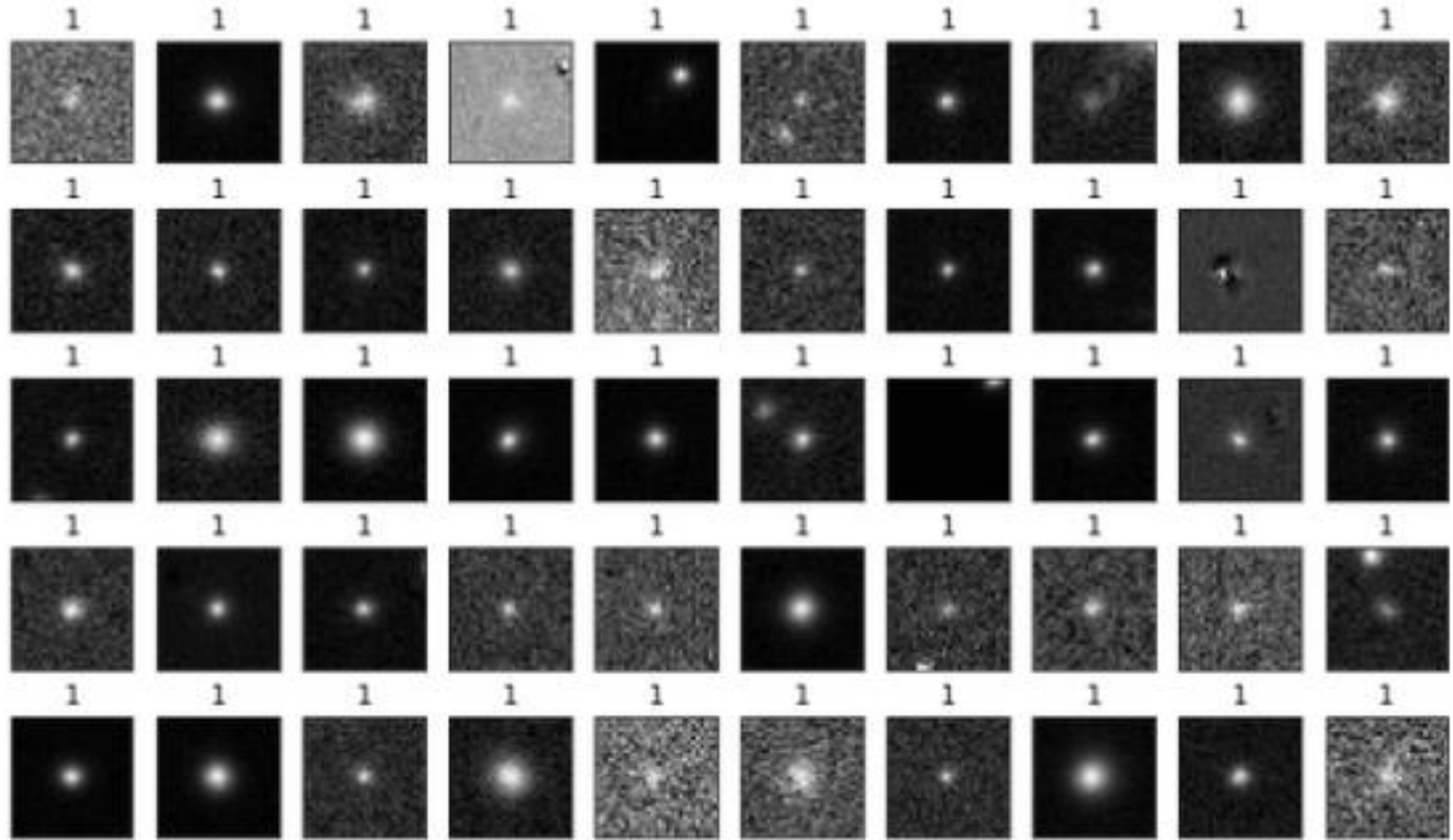


1. 機械学習によるTransient天体の検出とは？

3. Modelの学習

Modelの教師データ

1 : Transient (Artificial Real) 天体の教師データ

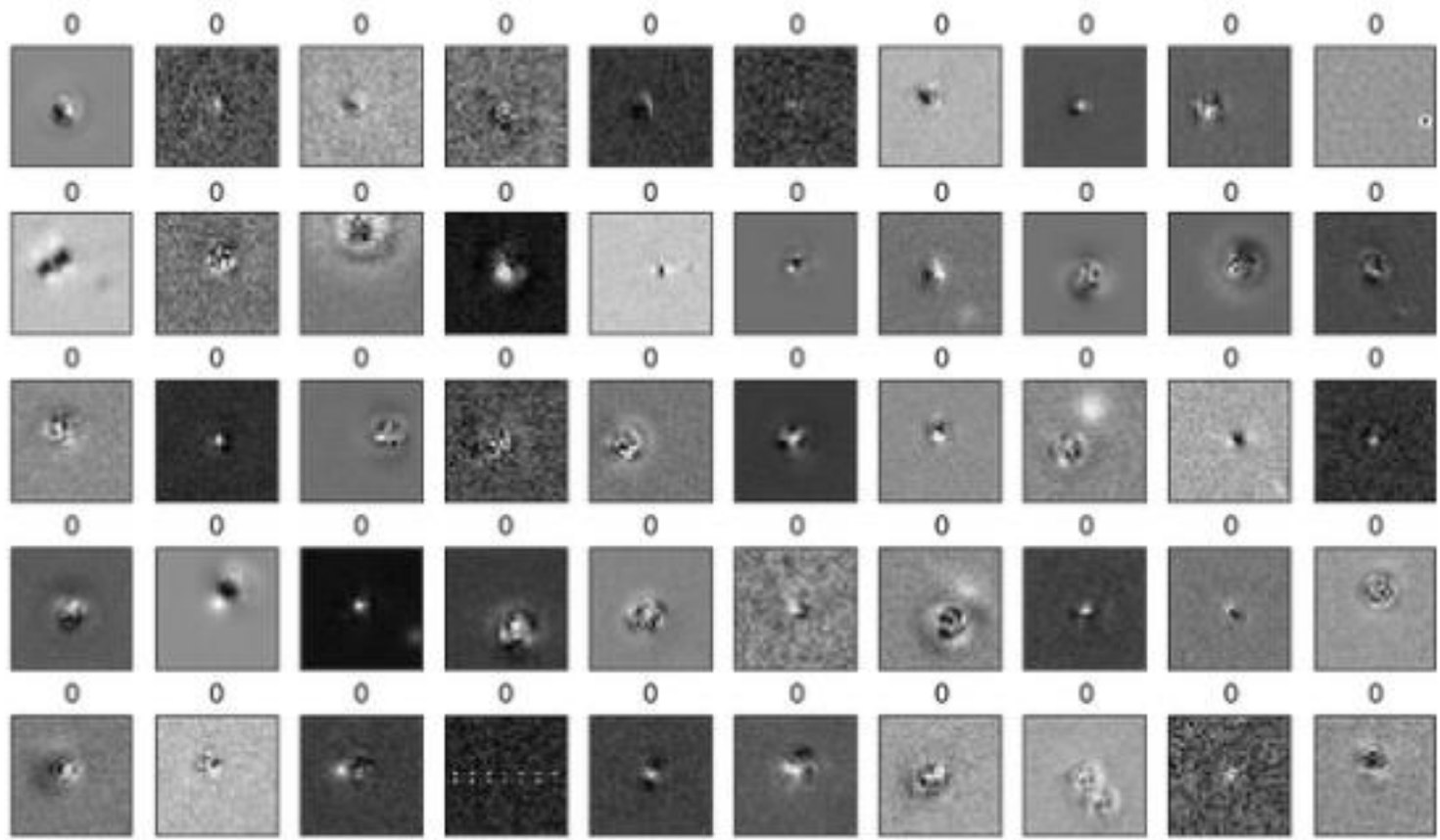


1. 機械学習によるTransient天体の検出とは？

3. Modelの学習

Modelの教師データ

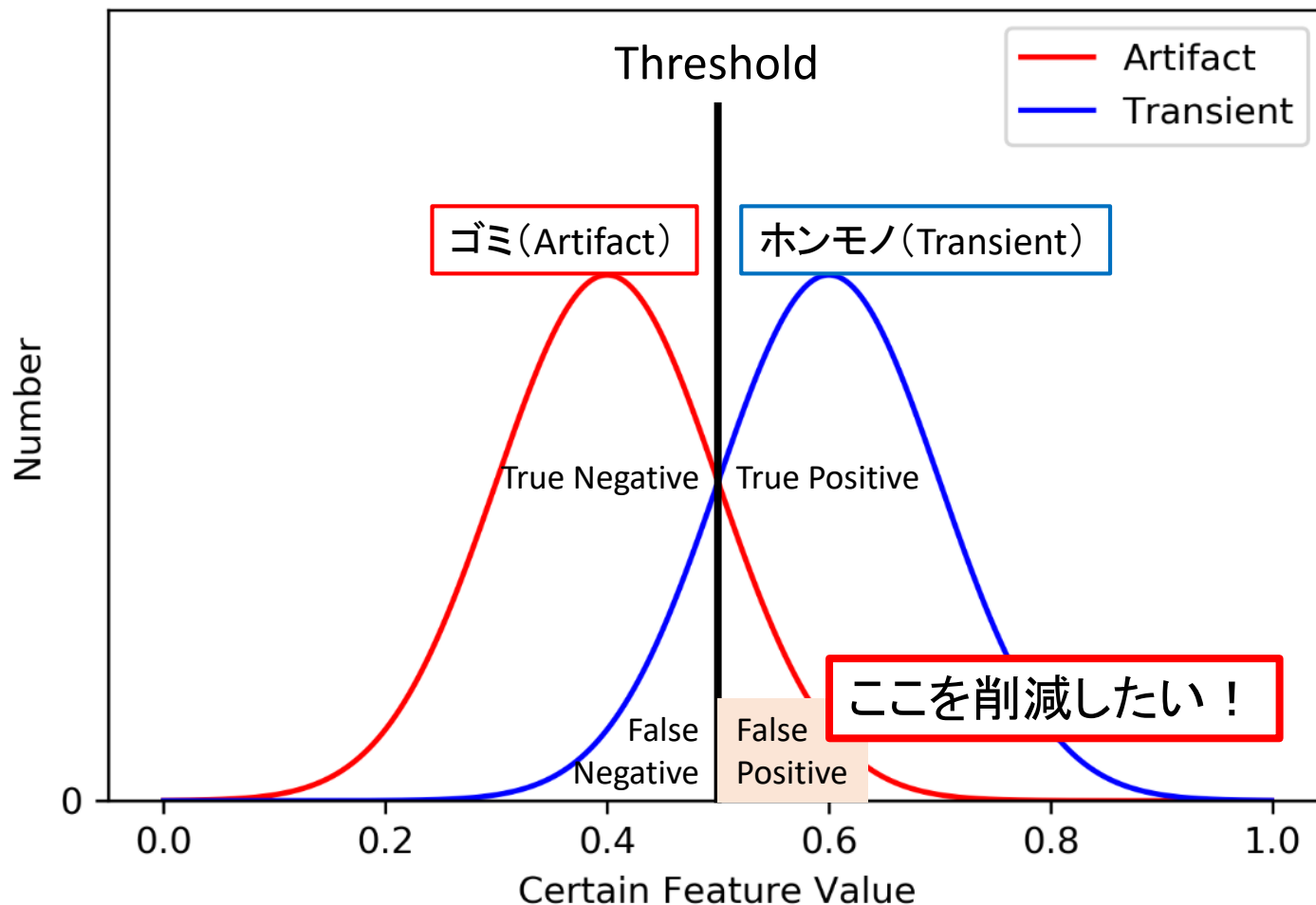
0 : Artifact天体の教師データ



1. 機械学習によるTransient天体の検出とは？

4. Modelの評価について

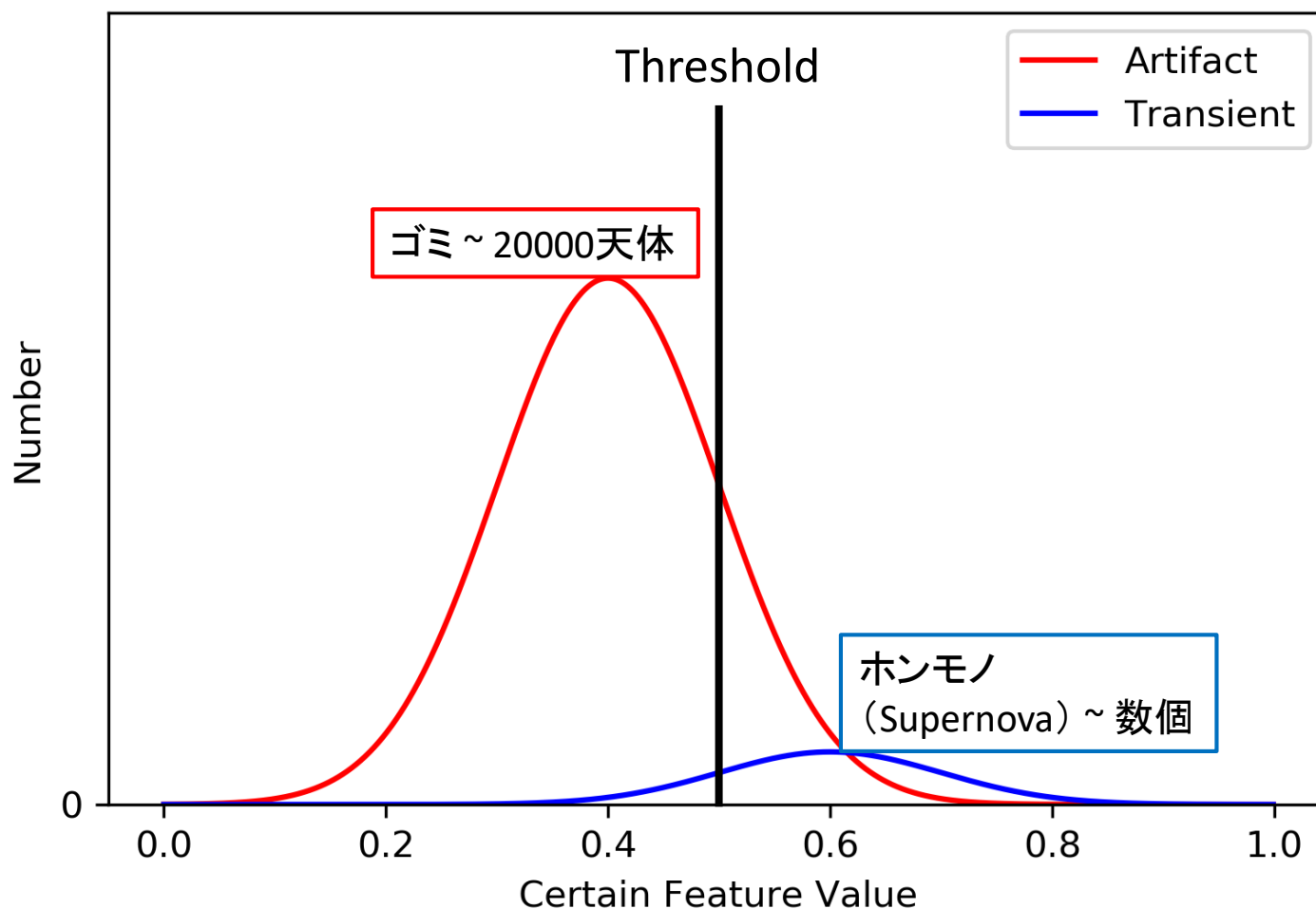
Modelの評価指標



1. 機械学習によるTransient天体の検出とは？

4. Modelの評価について

Modelを使う環境



• 目次

1. 機械学習によるTransient天体の検出とは？

1. Transient天体の数
2. Transient天体の検出手順
3. Modelの学習
4. Modelの評価について

2. 学習結果

1. 学習データ
2. ROC curve
3. False Positive の S/N について

3. 結論と展望

1. 今後の展望
2. まとめ

2. 学習結果 DET Q1 1-1

1. 学習データ

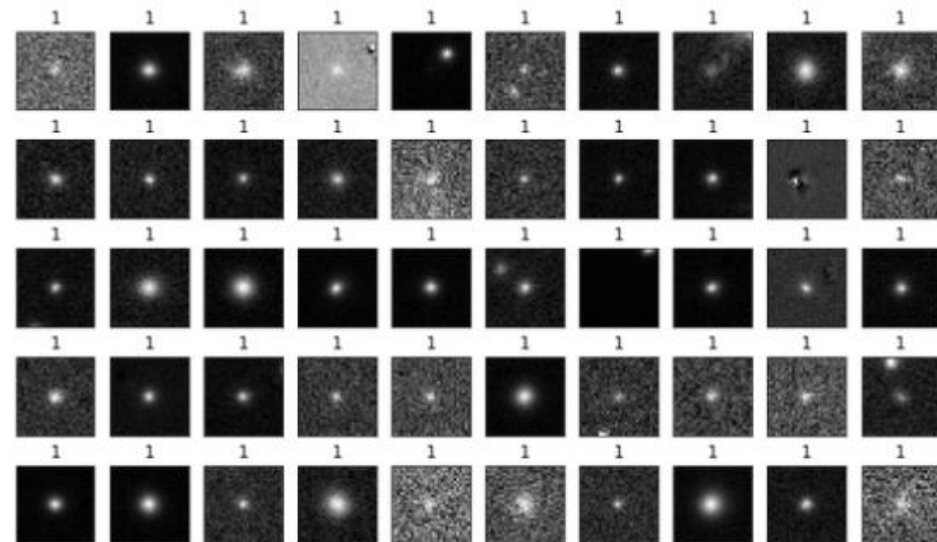
- Label: **Artificial Real**(ホンモノのTransientの代わりとして), **Artifact**
- データセット

Train-Data : 208,990 (104,495 : 104,495)

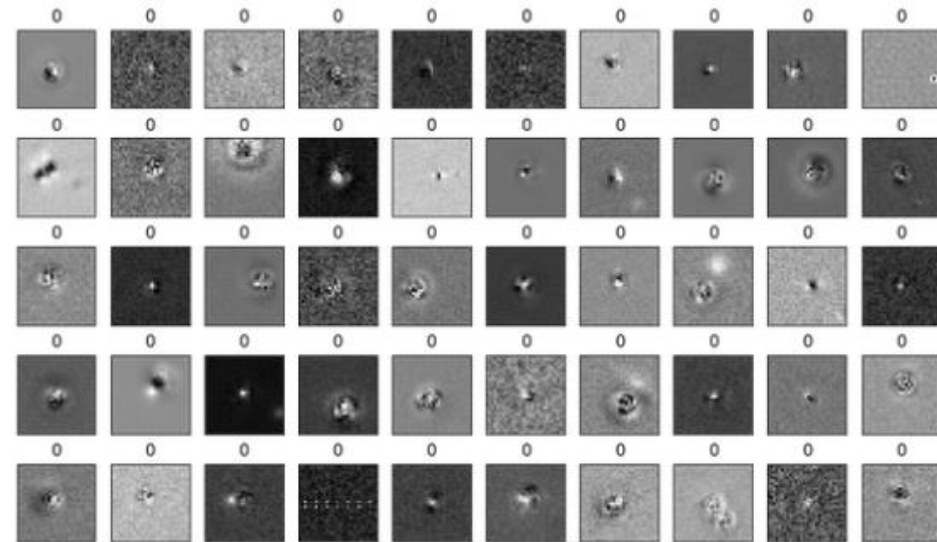
Validation-Data : 104,496 (52,248 : 52,248)

Test-Data : 104,496 (52,248 : 52,248)

Example of input images (1: FAKE 0: NOT FAKE) (1)



Example of input images (1: FAKE 0: NOT FAKE) (1)



2. 学習結果 DET Q1 1-1

2. ROC curve

Threshold

FPR = 2.1 % (at) TPR = 98.0 %

Artifact: 2万個 → ~400個

$TPR = TP / (TP + FN)$

全Artificial Realのうち、

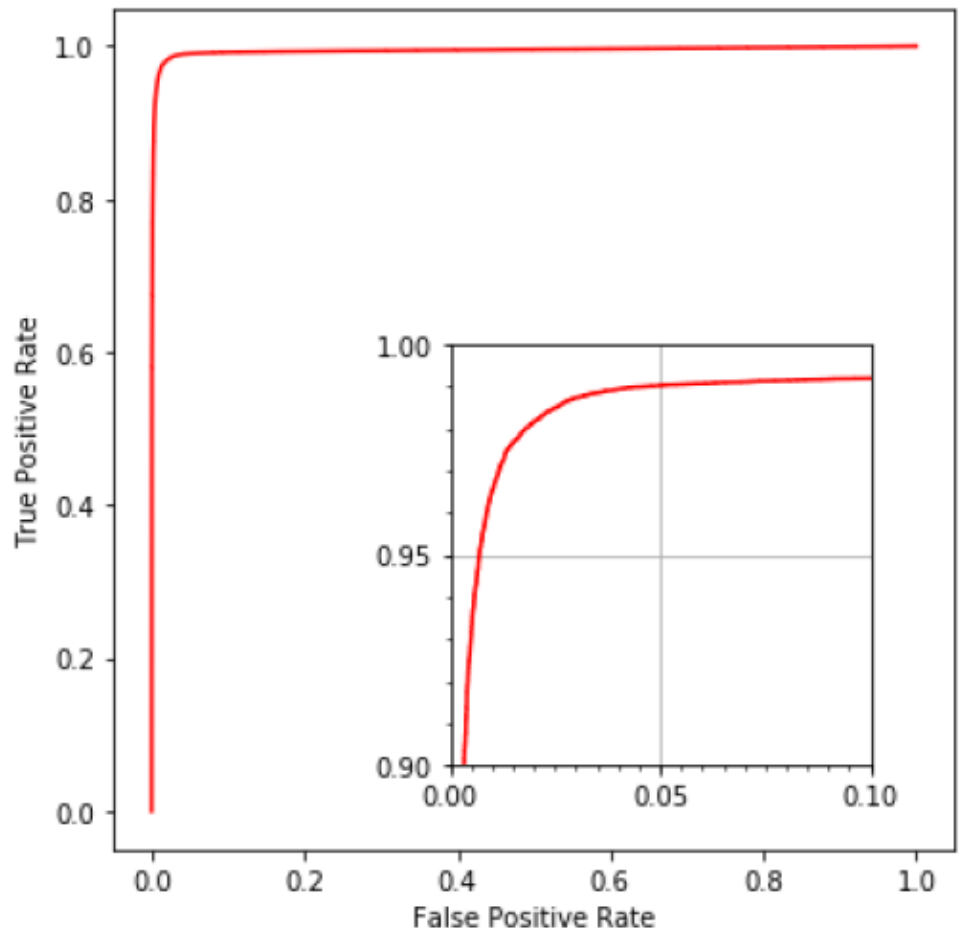
Modelが拾えた割合

$FPR = FP / (FP + TN)$

全Artifactのうち、

Modelが拾ってしまった割合

ROC curve for Test Set (AUC=0.994)

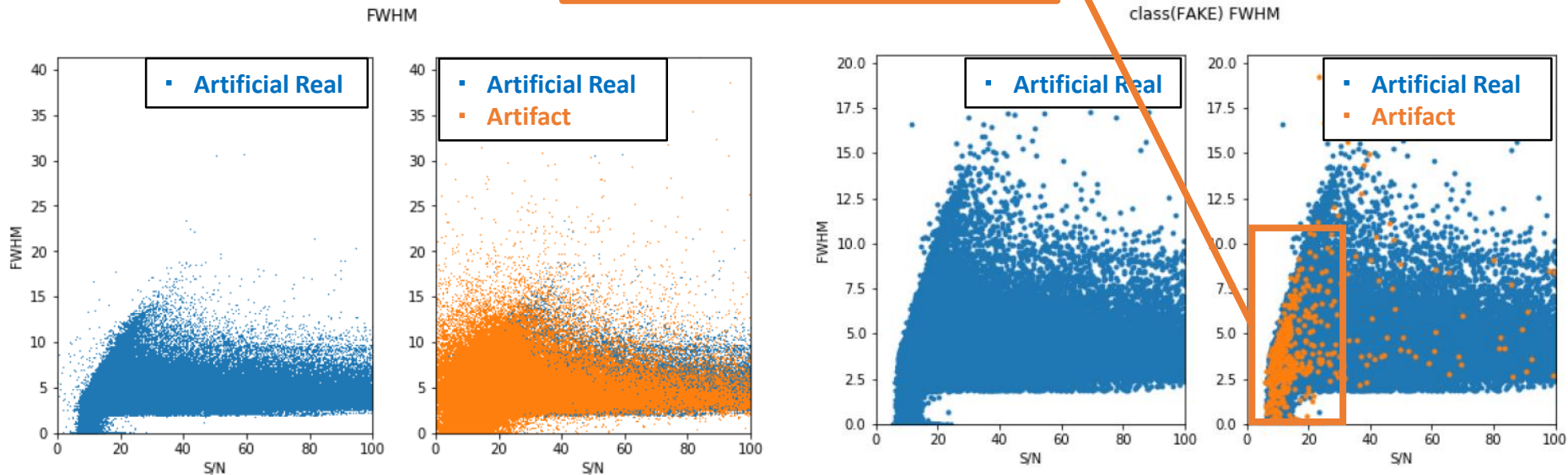


2. 学習結果 DET Q1 1-1

3. False Positive の S/N について

(左図) 全データ (右図) 学習器がホンモノと判定したTestデータ

False Positive は
S/N が低いゴミが多い？



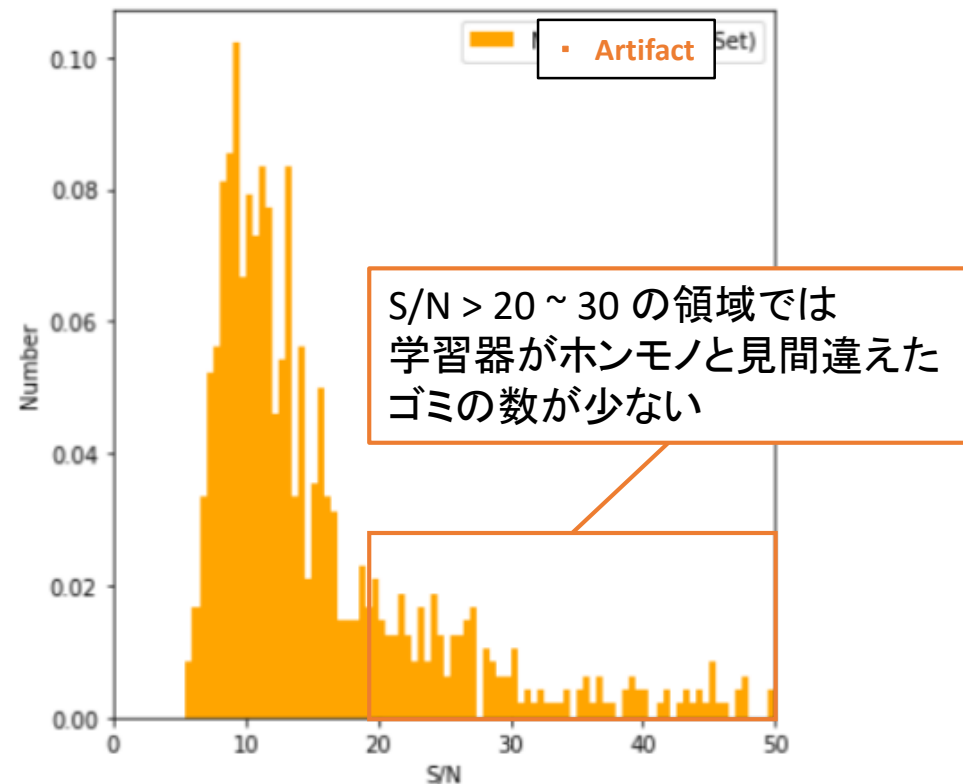
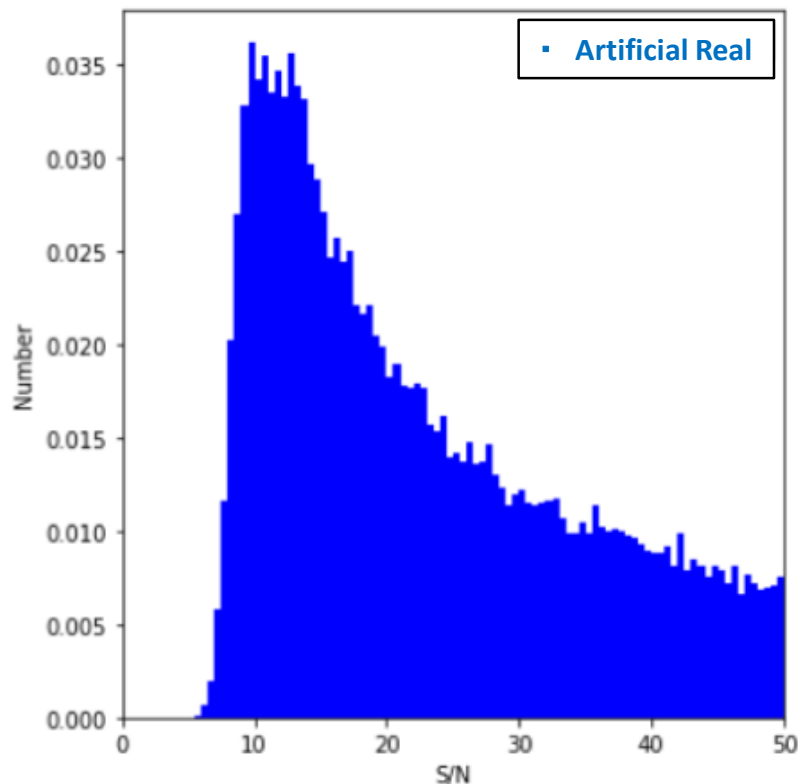
2. 学習結果 DET Q1 1-1

3. False Positive の S/N について

(左図) 学習器がホンモノと判定したTestデータのうち、Label : ホンモノ

(右図) 学習器がホンモノと判定したTestデータのうち、Label : ゴミ

Histogram of S/N using input Class(FAKE)=True label



2. 学習結果 DET Q1 1-1

3. False Positive の S/N について

- S/N に注目して性能評価

分類結果を $S/N = 20$ を基準に
ROC curve を比較。

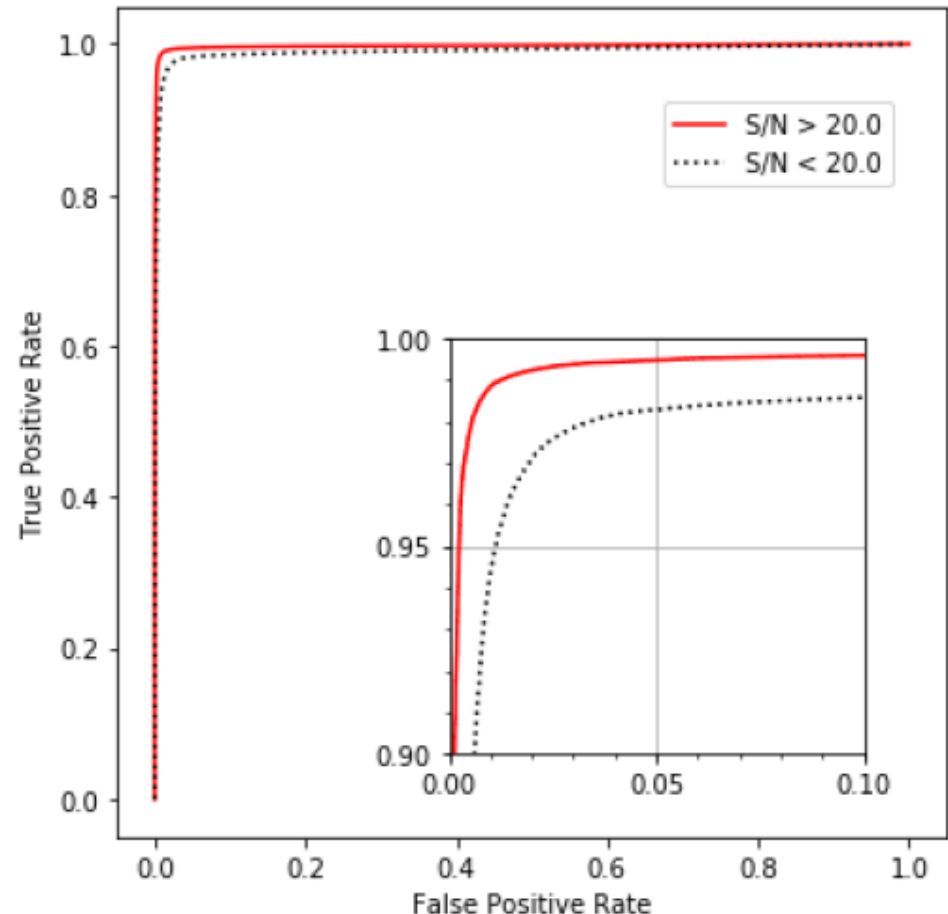
$S/N > 20$: **AUC = 0.997**

$S/N < 20$: AUC = 0.991

- $S/N > 20$

FPR = 0.4 % (at) TPR = 98.0 %

ROC curve (AUC=0.997/0.991)



2. 学習結果 DET Q1 1-1

3. False Positive の S/N について

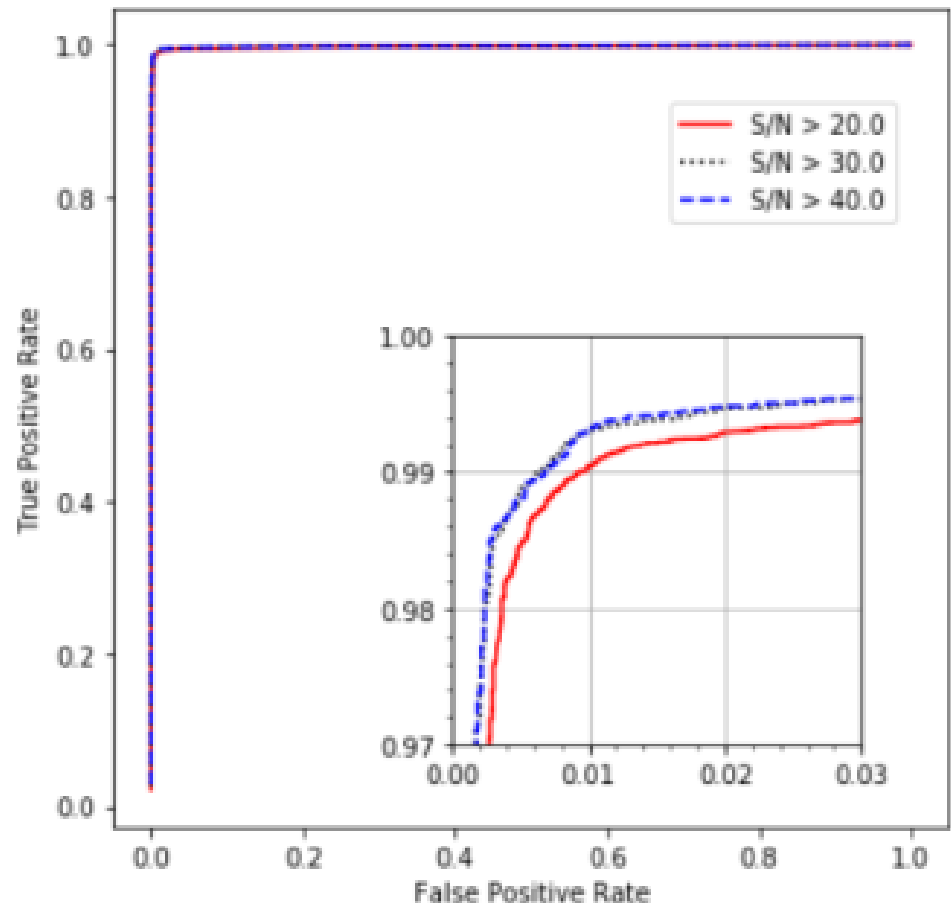
- S/N に注目して性能評価

S/N > 20 : AUC = 0.9977

S/N > 30 : AUC = 0.9982

S/N > 40 : AUC = 0.9983

ROC curve (AUC=0.99770/0.99824/0.99828)



• 目次

1. 機械学習によるTransient天体の検出とは？

1. Transient天体の数
2. Transient天体の検出手順
3. Modelの学習
4. Modelの評価について

2. 学習結果

1. 学習データ
2. ROC curve
3. False Positive の S/N について

3. 結論と展望

1. 今後の展望
2. まとめ

3. 展望 とまとめ

1. 今後の展望

- これらはシミュレーション環境における結果
→ 人工的に作ったTransient (Artificial Real) を使った学習結果



このModelが実際のTransientを拾えるか、
どれだけゴミを拾ってしまうのかを検証する。

また、S/Nの高いTransientが見つかった場合、

Slackワークスペース [Tomo-e Gozen](#) の

[# tomoe-sn](#) チャンネルにて、SN bot に呟かせる予定

高S/Nに関しては、自動アラートを実施できる？

3. 展望 とまとめ

2. まとめ

- シミュレーションでの結果

全体 : AUC = 0.994

S/N > 20 : AUC = 0.9977

S/N > 30 : AUC = 0.9982

S/N > 40 : AUC = 0.9983

→ S/Nが高いほど、結果は良い。

このModelが実際のTransientを拾えるか、
どれだけゴミを拾ってしまうのかを検証していく。

付録

- 付録

1. 全ChipのModel

-24

2. 学習データのサイズ

-26

3. False学習データ汚染、CNNの問題

-30

4. False Negativeについて

-48

5. 最新資料

-51

付録1：全ChipのModel

4. 付録

- 各検出器のModelのAUCと、平均AUCについて

各検出器のModelのROC - curveについては、Shinohara:

/home/tomoesn/script/CNN/cnn_model_v2/Qxxx/

のディレクトリ内に各ROC画像があります。

- 19/06/07 現在の
CNN model の実装状況

以下のChip以外は実装済

Q3

2-1, 2-2, 2-3, 2-4, 2-5, 2-6, 4-1, 4-2, 4-3, 4-4

AUC All mean AUC 0.993959

Q1 mean AUC 0.993143

11	12	13	14	15	16
0.994	0.994	0.994	0.994	0.991	0.991
21	22	23	24	25	26
0.989	0.992	0.993	0.993	0.994	0.995
31	32	33	34	35	
0.994	0.994	0.993	0.994	0.995	
41	42	43	44		
0.993	0.991	0.993	0.995		

Q2 mean AUC 0.994048

11	12	13	14	15	16
0.995	0.995	0.992	0.995	0.994	0.995
21	22	23	24	25	26
0.994	0.994	0.994	0.994	0.992	0.994
31	32	33	34	35	
0.993	0.995	0.993	0.995	0.994	
41	42	43	44		
0.994	0.994	0.994	0.995		

Q3 mean AUC 0.993455

11	12	13	14	15	16
0.992	0.991	0.994	0.995	0.994	0.994
21	22	23	24	25	26
31	32	33	34	35	
0.993	0.994	0.993	0.994	0.994	
41	42	43	44		

Q4 mean AUC 0.994952

11	12	13	14	15	16
0.996	0.995	0.994	0.996	0.995	0.996
21	22	23	24	25	26
0.995	0.994	0.994	0.994	0.994	0.995
31	32	33	34	35	
0.996	0.996	0.996	0.996	0.996	
41	42	43	44		
0.994	0.994	0.994	0.994		

付録2: 学習データのサイズ

DET Q1 1-1 学習結果

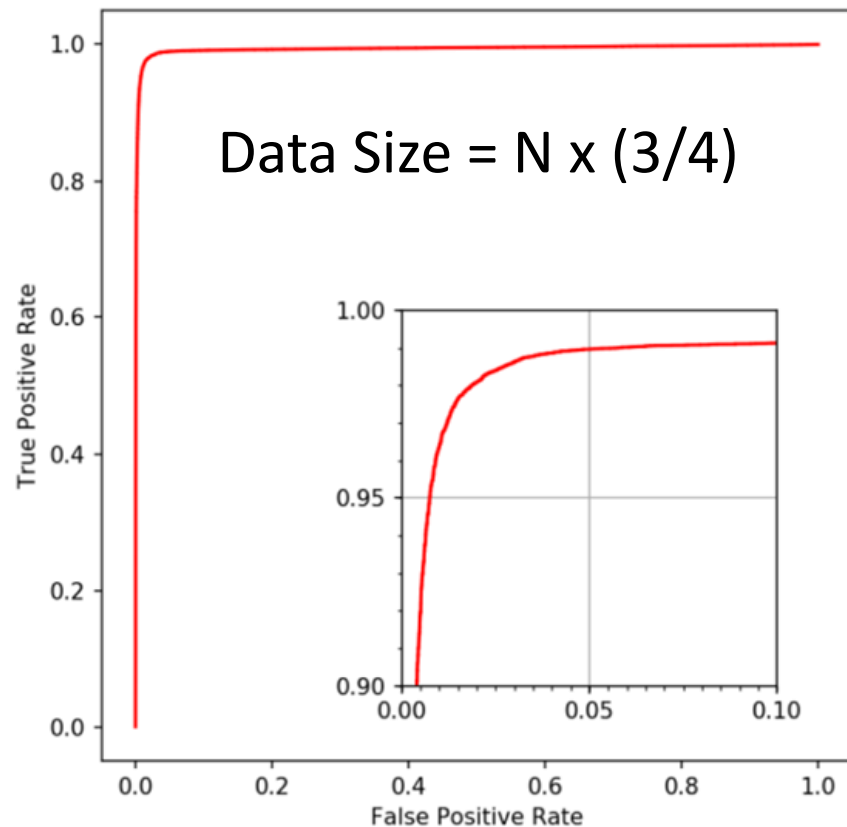
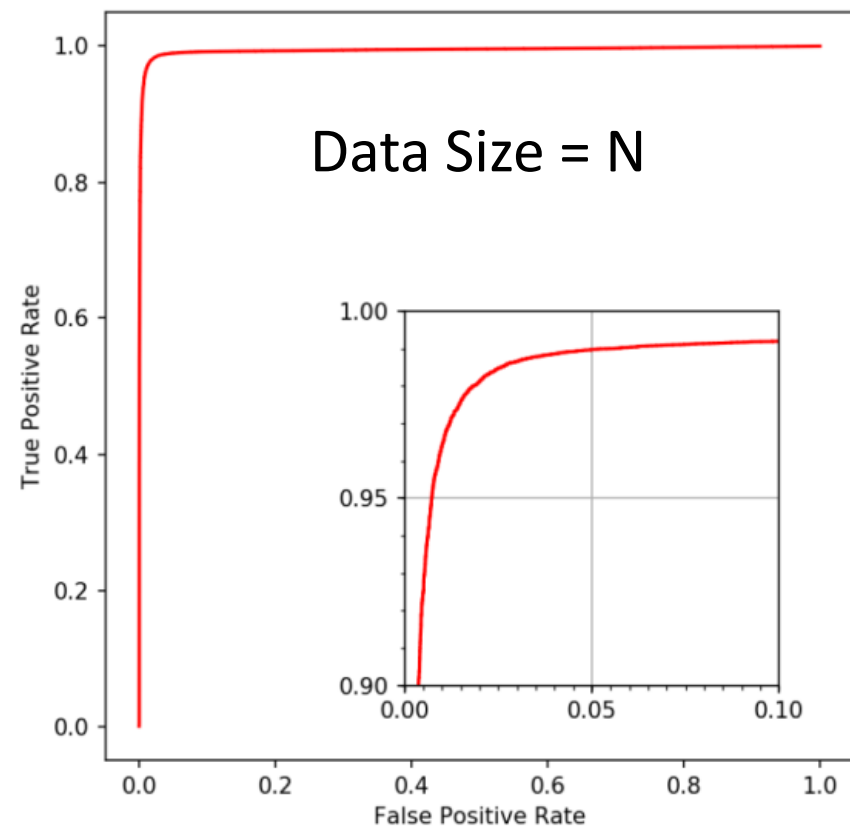
• 学習データのサイズ

N: Train Size = 20万

ROC-curveを比較。

ROC curve for Test Set (AUC=0.994)

ROC curve for Test Set (AUC=0.994)



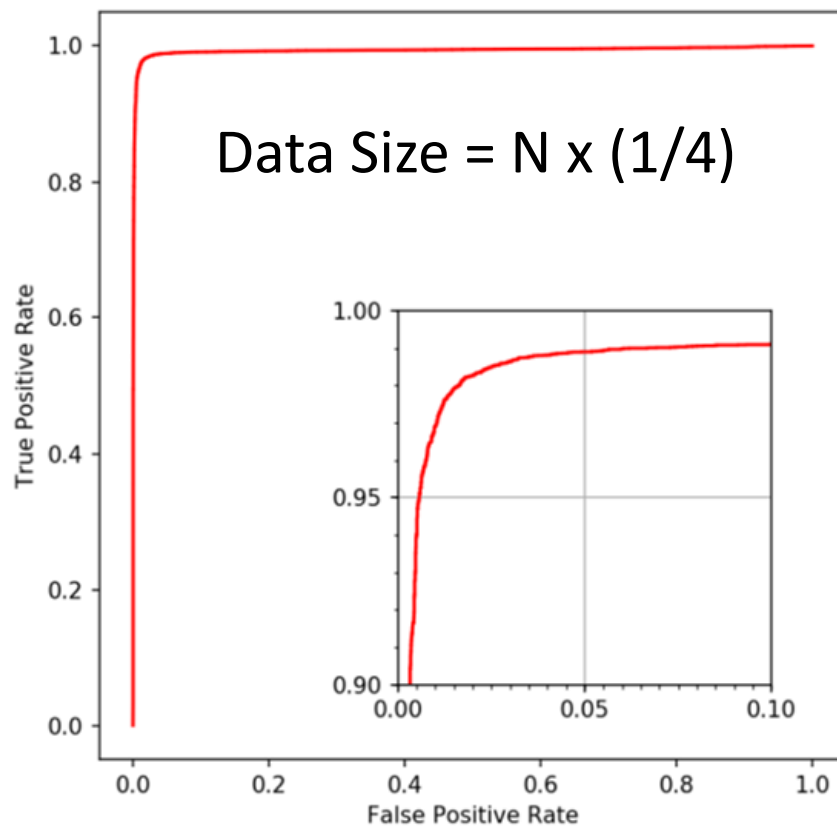
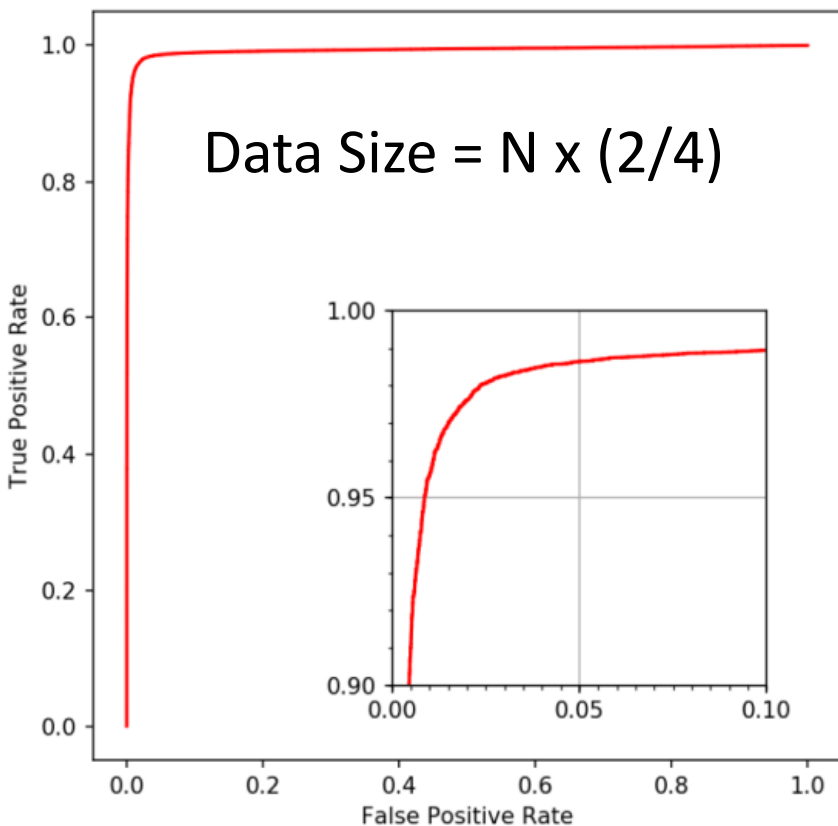
DET Q1 1-1 学習結果

- 学習データのサイズ

ROC-curveを比較。

ROC curve for Test Set (AUC=0.993)

ROC curve for Test Set (AUC=0.993)

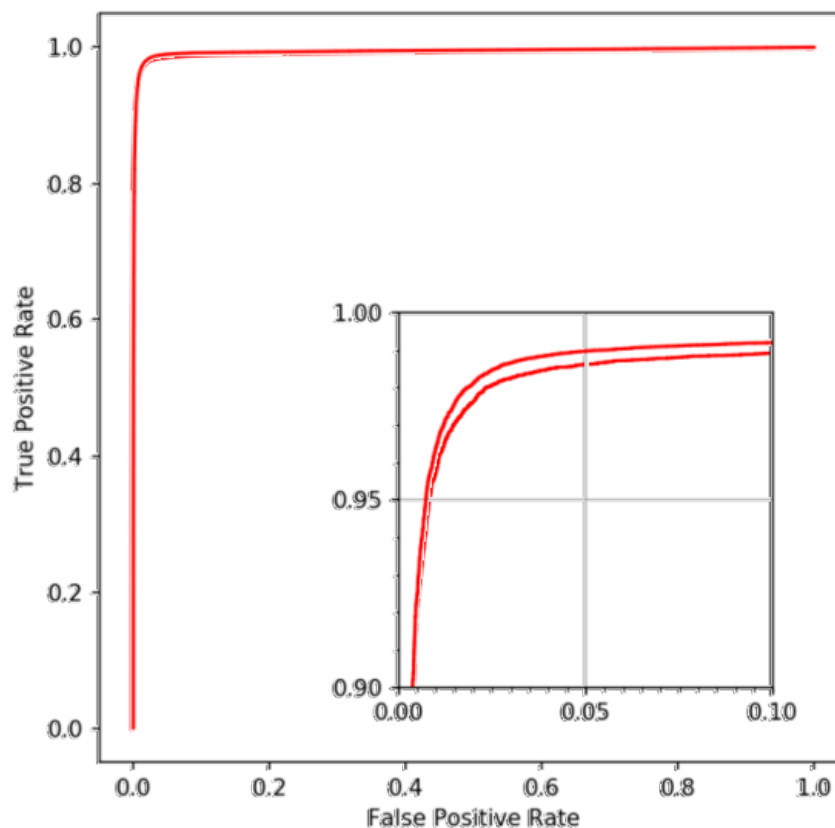


DET Q1 1-1 学習結果

- 学習データのサイズ

ROC-curveを比較。

Data Size = $N \times (2/4)$ vs Data Size = N



付録3: False学習データ汚染、CNNの問題

- 学習データが抱える問題の背景

Fake label 学習データについて、既知の汚染の要因

- まったく差分できていない天体
- 明るい天体付近の差分失敗天体
- 差分画像に写った本物のTransient天体
- 小惑星

• Abstract

1. 新旧Model の ROC - curve の比較

(DET Q1 1-1) 新: AUC = 0.994 旧: AUC = 0.984

2. 旧Modelの問題点とその解決

- まったく差分できていない天体による False label 学習データの汚染を排除し、学習データを改善

3. 未だ残る、新Modelの問題点と展望

- 明るい天体付近の差分失敗天体、差分画像に写った本物のTransient天体、小惑星による、False label 学習データの汚染
- S/Nの低いFalse labelをTrueと判別している。
- 差分失敗天体付近のホンモノがFalse Negativeに多い。

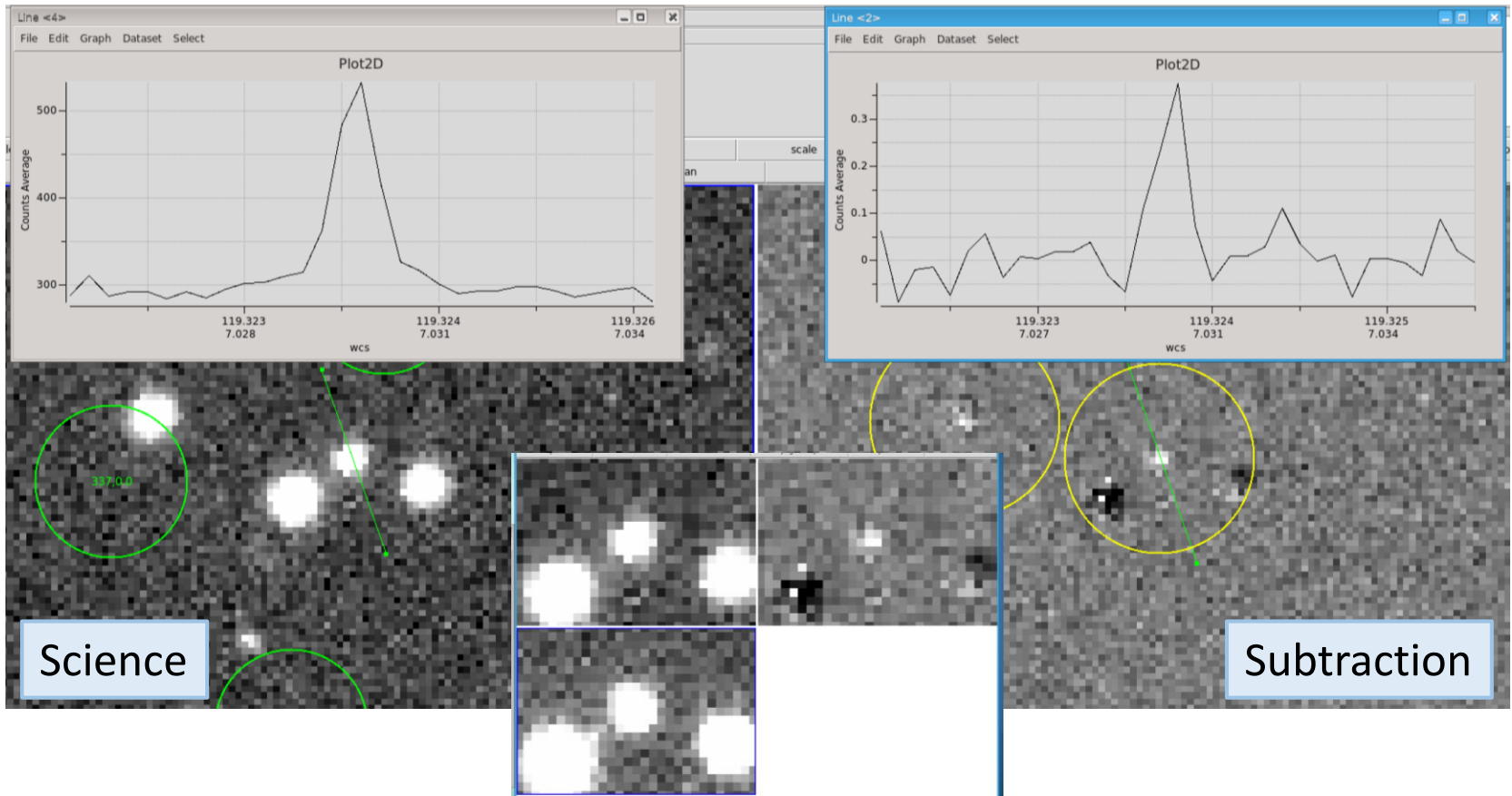
4. 付録

各検出器のModelの ROC - curve について

3. False Positive の image について

- 新Model (DET Q1 1-1)

明るい天体付近の差分失敗天体がFalse Positiveに紛れていることを確認。



False Positive データ汚染の調査

• ファイルごとのFalse Positive数 – TOP20file

含まれるFP天体の数...FP全体の約26%

大幅な位置ズレ : 6/20 改善済み

PSFの形が異常 : 6/20 (ホンモノなはずだが、labelがニセモノ)

天体数が多いimageであり、差分が失敗している : 5/20

CNNが拾ってしまったゴミ : 3/20

Number of False Positive : 3942

Top 11 - 20

```
file name : sTMQ1201903190009423811, count : 45
file name : sTMQ1201902210008355311, count : 42
file name : sTMQ1201902210008345011, count : 42
file name : sTMQ1201901050007090211, count : 32
file name : sTMQ1201903220009487411, count : 31
file name : sTMQ1201903290009618011, count : 31
file name : sTMQ1201903290009611411, count : 29
file name : sTMQ1201902010007790111, count : 28
file name : sTMQ1201902210008352911, count : 26
file name : sTMQ1201903290009617111, count : 26
```

Top 10

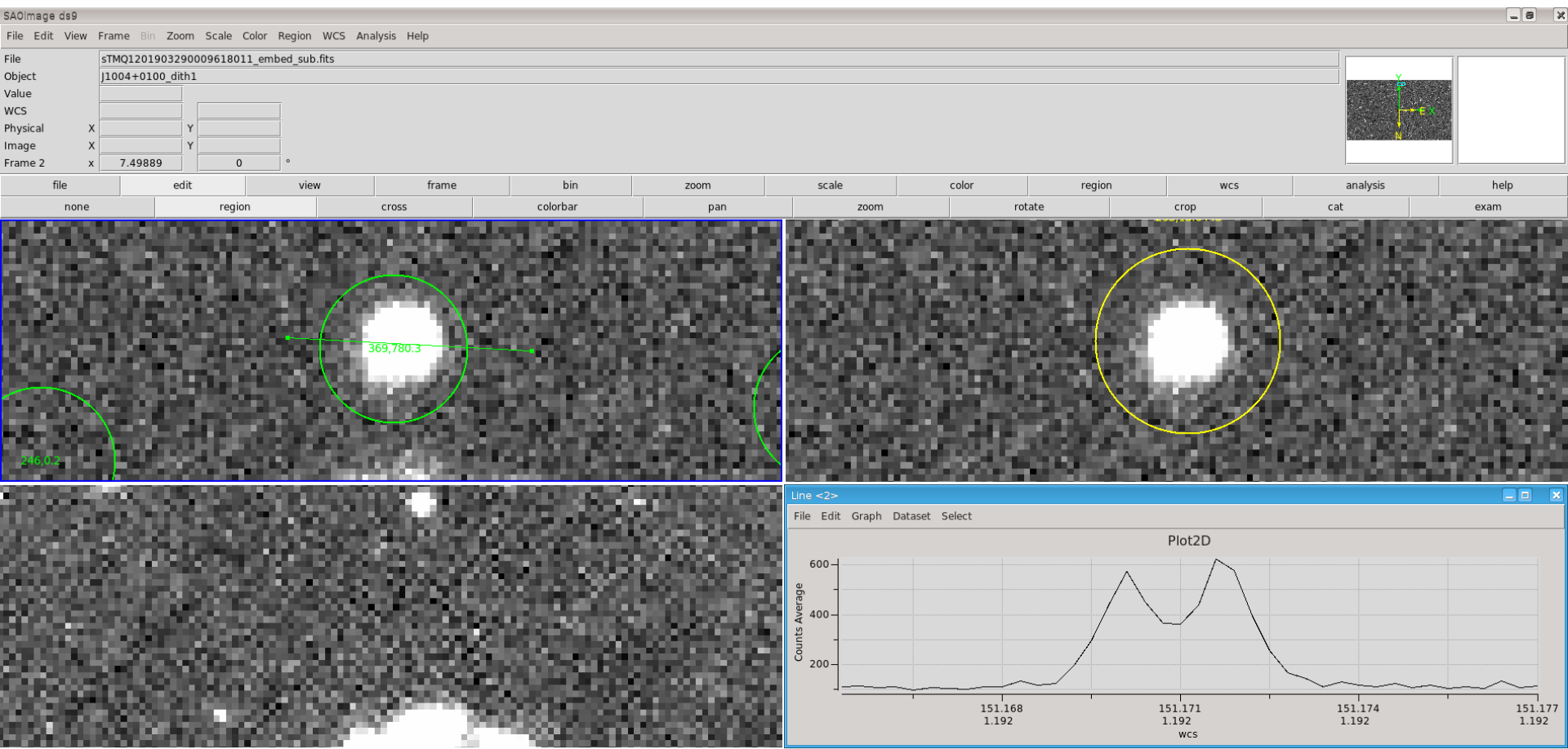
```
file name : sTMQ1201902230008479111, count : 148
file name : sTMQ1201905100011144211, count : 73
file name : sTMQ1201903170009242611, count : 71
file name : sTMQ1201902170008208811, count : 69
file name : sTMQ1201902210008350011, count : 69
file name : sTMQ1201903290009619411, count : 63
file name : sTMQ1201903290009629011, count : 63
file name : sTMQ1201902210008342311, count : 55
file name : sTMQ1201901220007635011, count : 51
file name : sTMQ1201903290009622611, count : 49
```

```
sTMQ1201903190009423811 : Misalignment
sTMQ1201902210008355311 : SUB error (Garactic plane ?)
sTMQ1201902210008345011 : SUB error (Garactic plane ?)
sTMQ1201901050007090211 : CNN error ? ----- Check image later
sTMQ1201903220009487411 : Misalignment ----- Check image later
sTMQ1201903290009618011 : PSF error ----- Check image later
sTMQ1201903290009611411 : PSF error ----- Check image later
sTMQ1201902010007790111 : Misalignment
sTMQ1201902210008352911 : SUB error (Garactic plane ?)
sTMQ1201903290009617111 : PSF error
```

```
sTMQ1201902230008479111 : Misalignment ?
sTMQ1201905100011144211 : CNN error (Bright star !)
sTMQ1201902210008350011 : SUB error (Garactic plane)
sTMQ1201901220007635011 : CNN error (Background is high brightness)
sTMQ1201903170009242611 : Misalignment
sTMQ1201902170008208811 : Misalignment
sTMQ1201903290009619411 : PSF error
sTMQ1201903290009629011 : PSF error
sTMQ1201902210008342311 : Misalignment SUB error
sTMQ1201903290009622611 : PSF error
```

False Positive データ汚染の調査

- 天体のPSFの形が異常

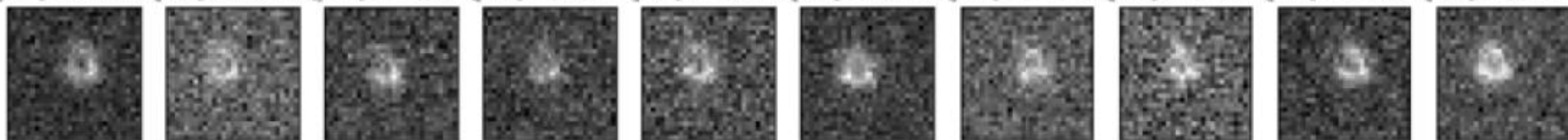


False Positive データ汚染の調査

- 天体のPSFの形が異常

False Positive (sTMQ1201903290009618011) PSF error ! (1)

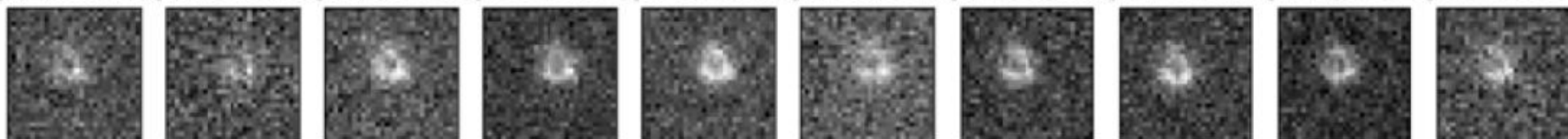
0/1 (212180) 0/1 (223374) 0/1 (223615) 0/1 (230666) 0/1 (236938) 0/1 (250778) 0/1 (253231) 0/1 (275528) 0/1 (281691) 0/1 (283449)



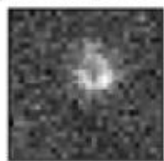
0/1 (283830) 0/1 (290777) 0/1 (293026) 0/1 (294144) 0/1 (297979) 0/1 (305938) 0/1 (314671) 0/1 (318111) 0/1 (330641) 0/1 (331338)



0/1 (336980) 0/1 (342989) 0/1 (367119) 0/1 (373676) 0/1 (377081) 0/1 (389249) 0/1 (390681) 0/1 (404128) 0/1 (407174) 0/1 (412883)

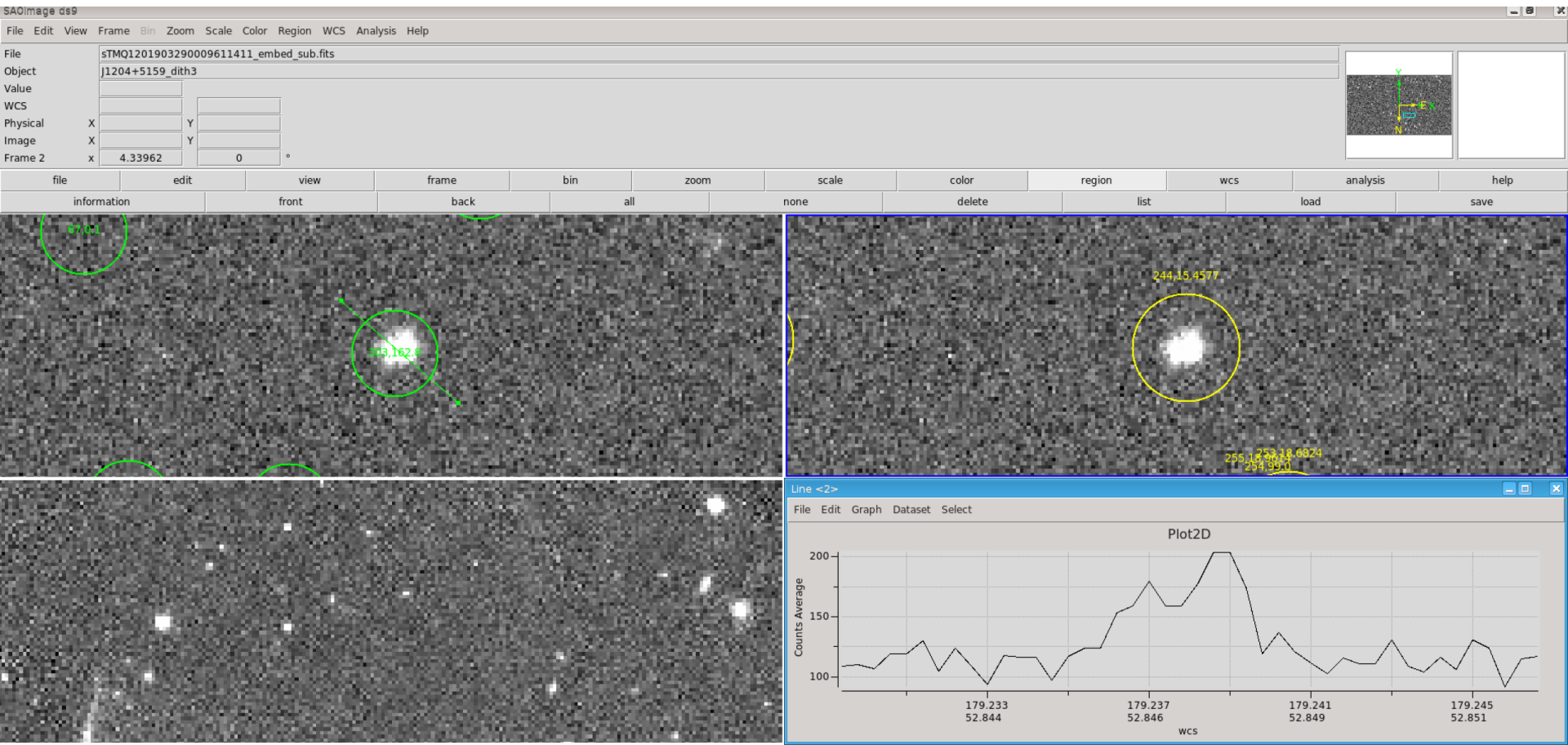


0/1 (416722)



False Positive データ汚染の調査

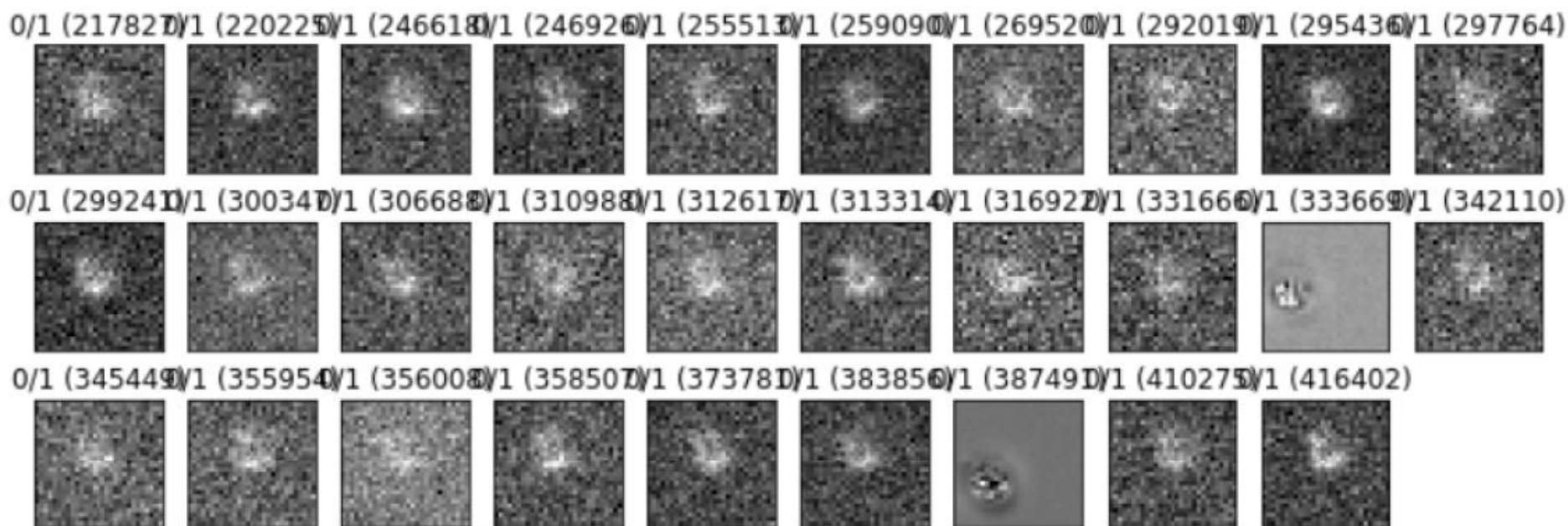
- 天体のPSFの形が異常



False Positive データ汚染の調査

- 天体のPSFの形が異常

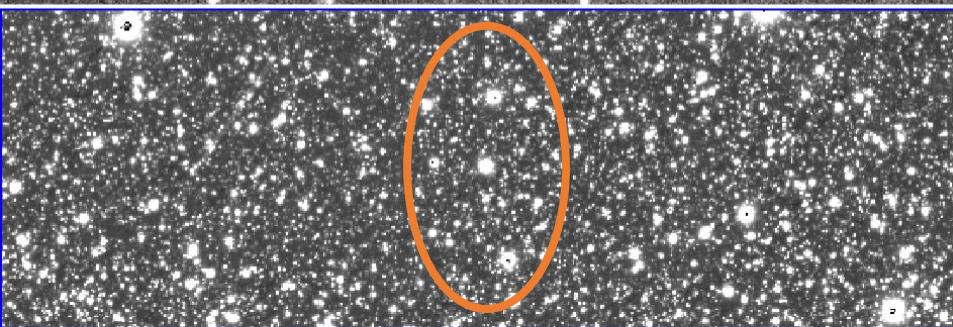
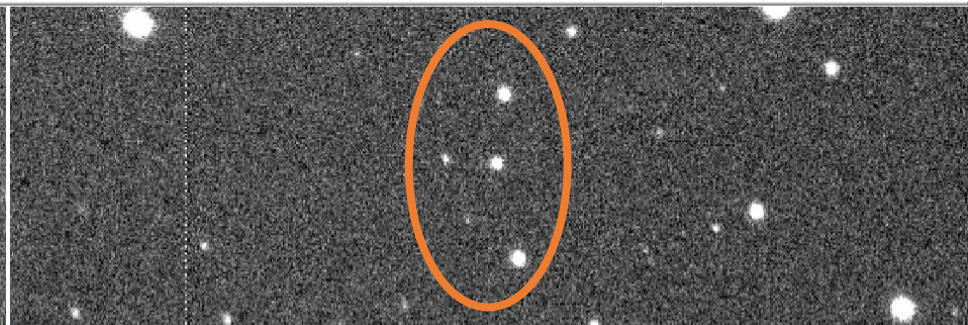
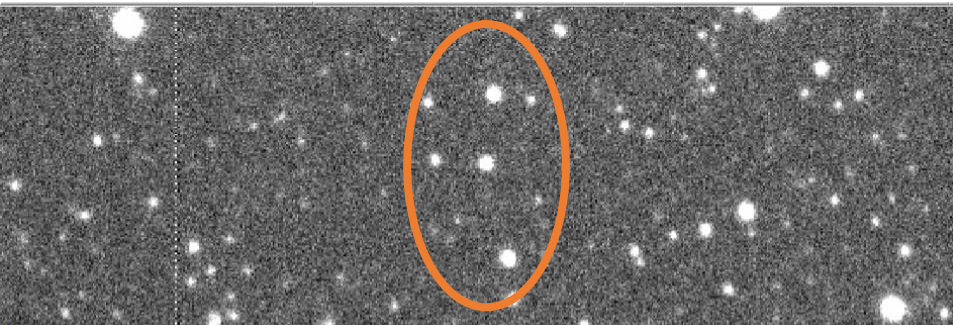
False Positive (sTMQ1201903290009611411) PSF error ! (1)



False Positive データ汚染の調査

- 差分失敗 (天体数が多いimage)

天の川付近 差分ができていない

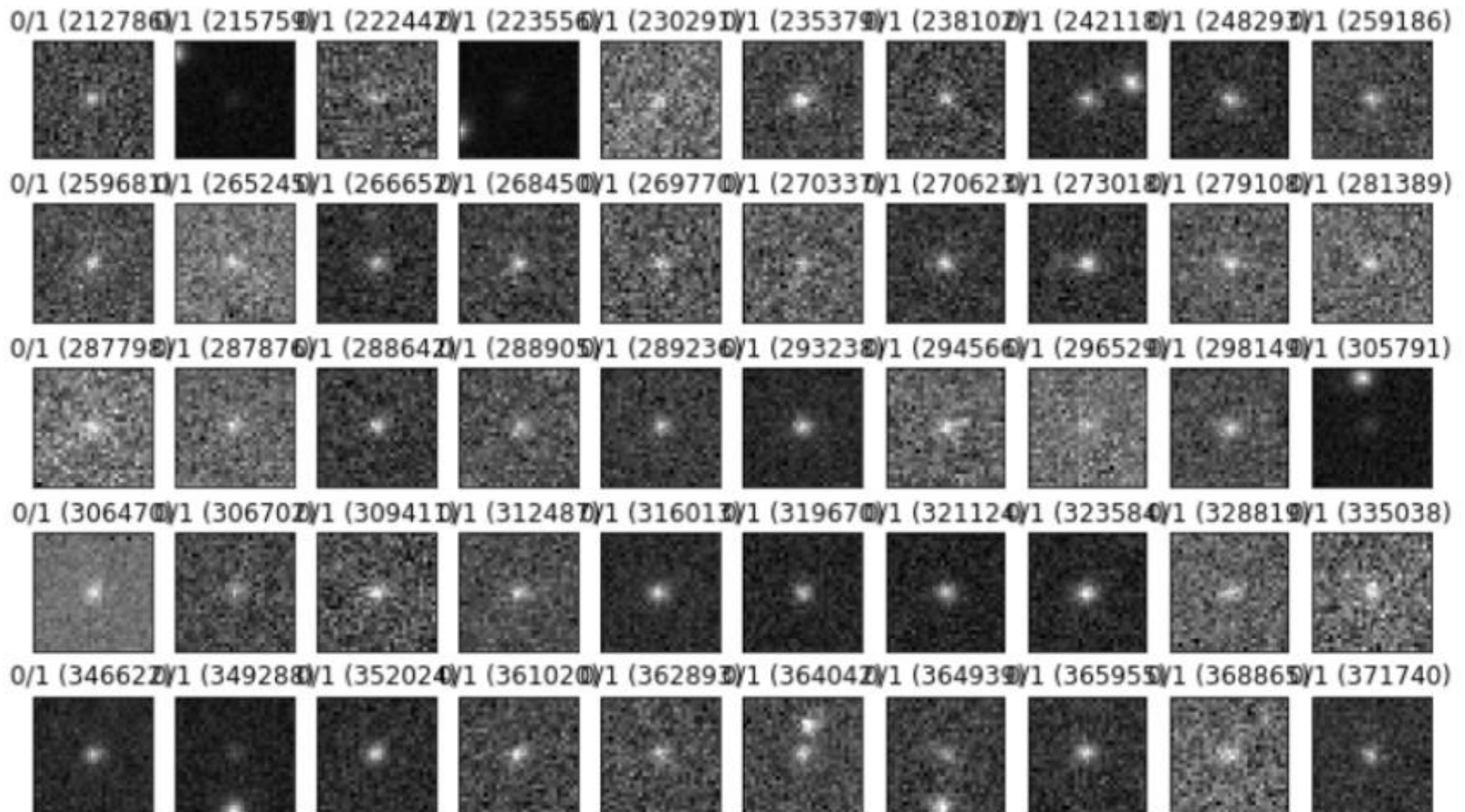


False Positive データ汚染の調査

- 差分失敗 (天体数が多いimage)

天の川付近 差分ができていない

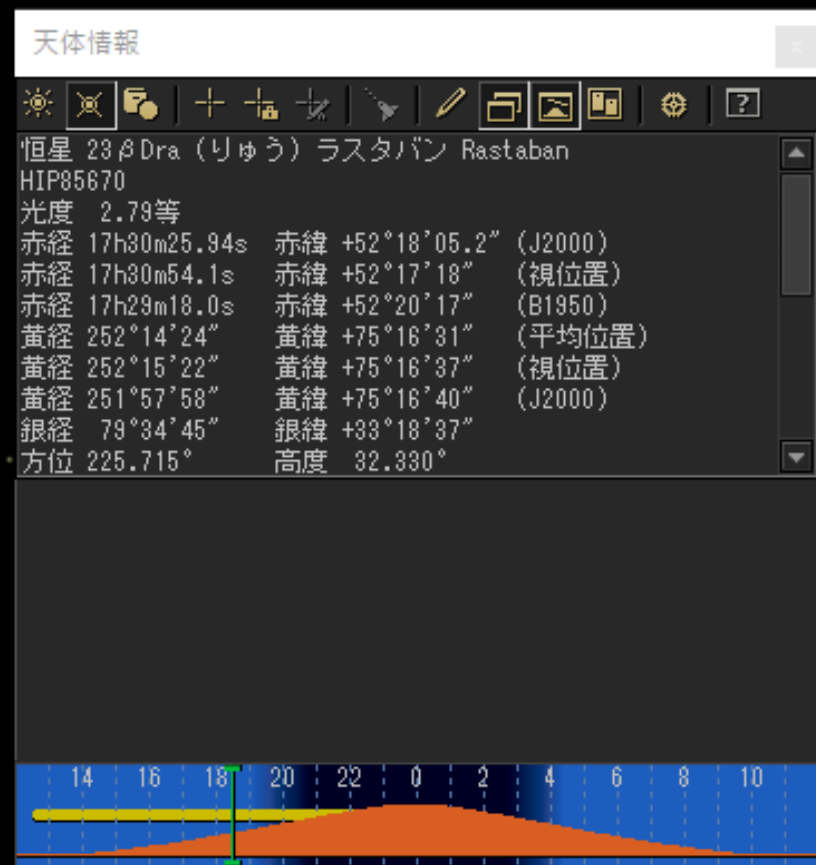
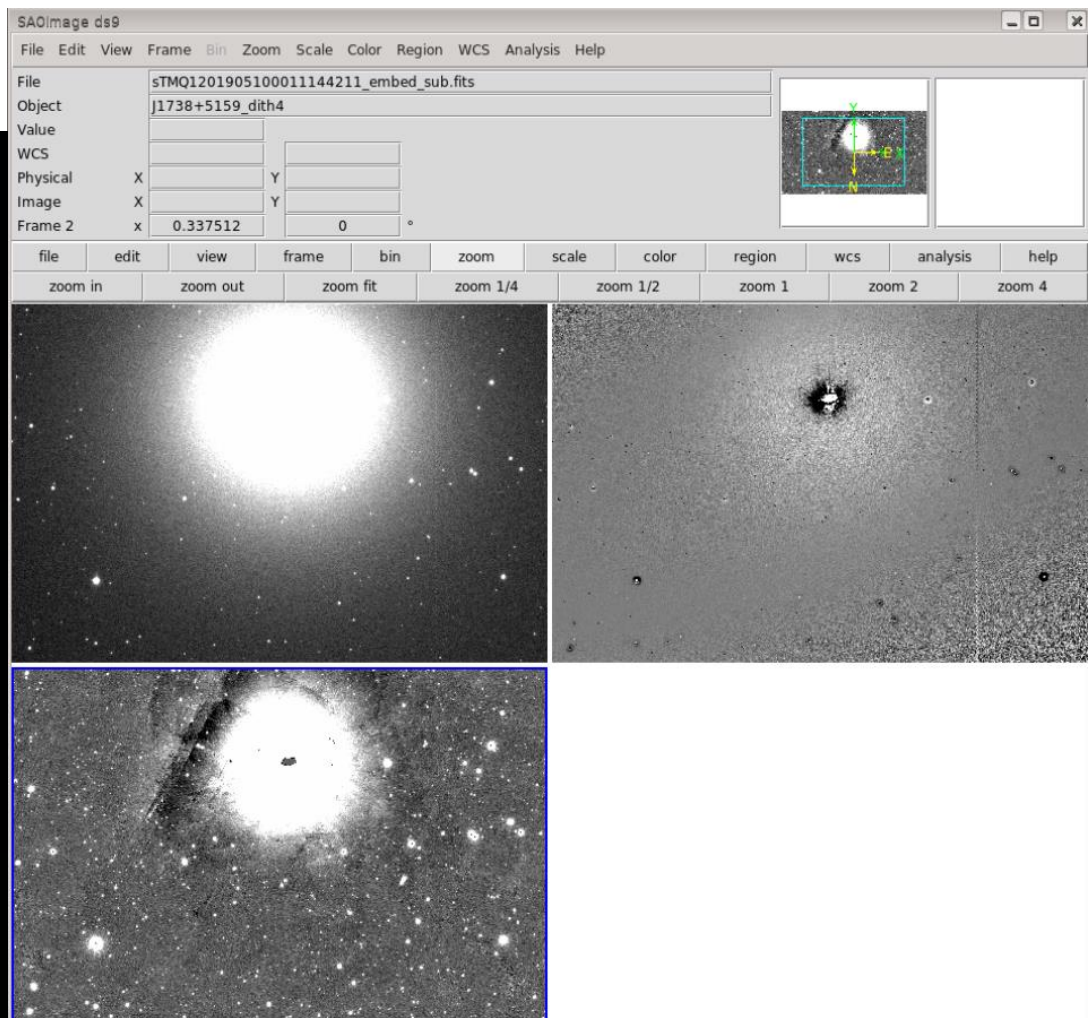
False Positive (sTMQ1201902210008350011) Miss SUB by Galactic plane ? (1)



False Positive データ汚染の調査

- CNNのerror

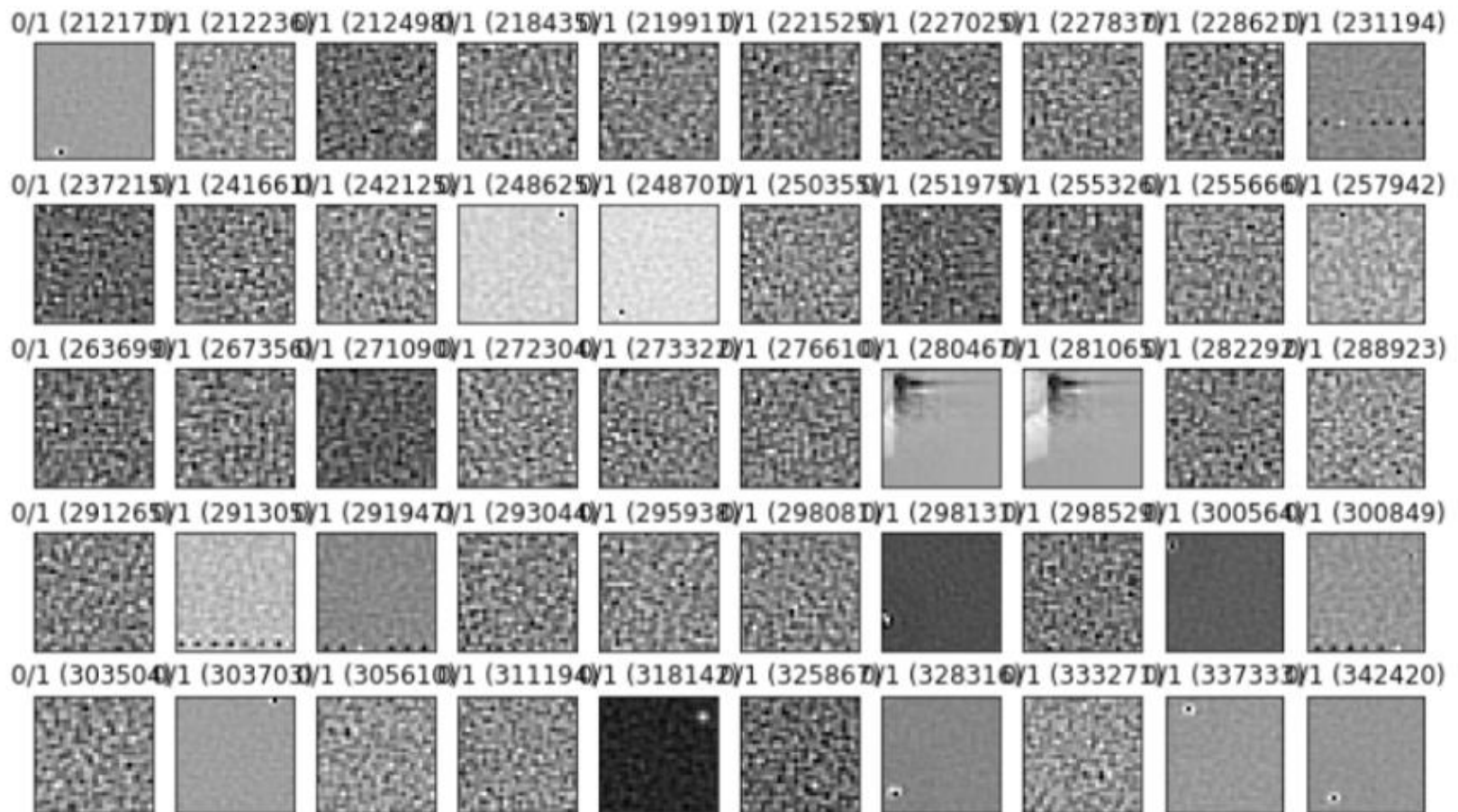
明るい星周り



False Positive データ汚染の調査

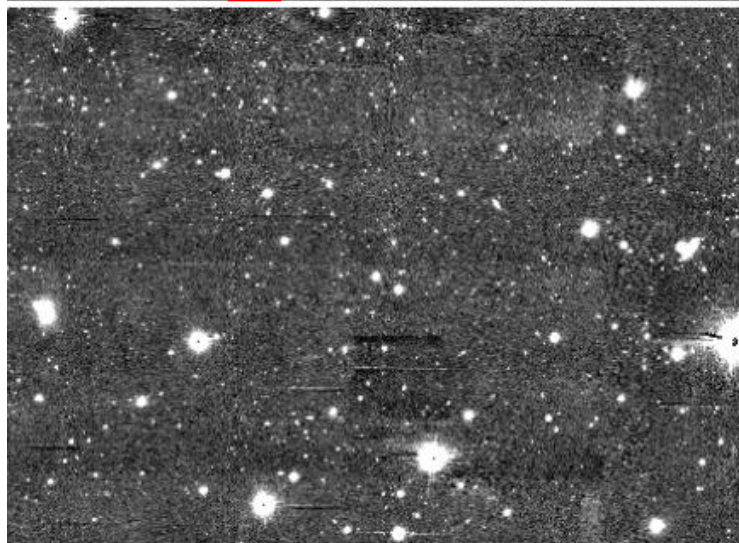
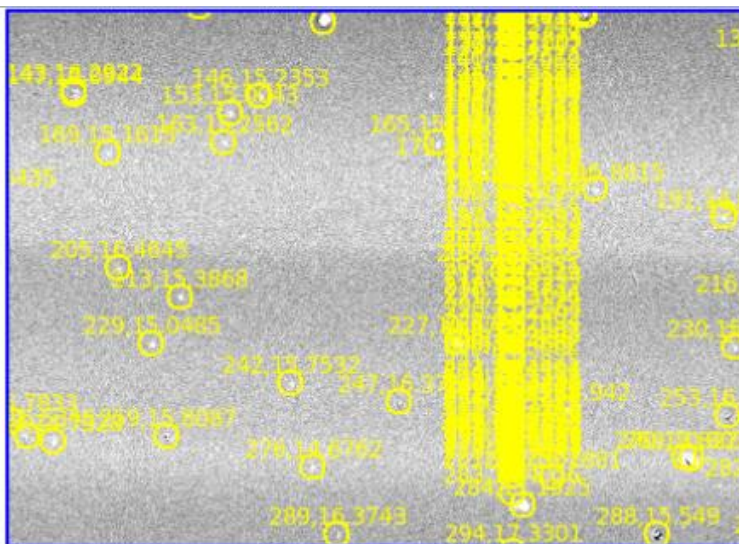
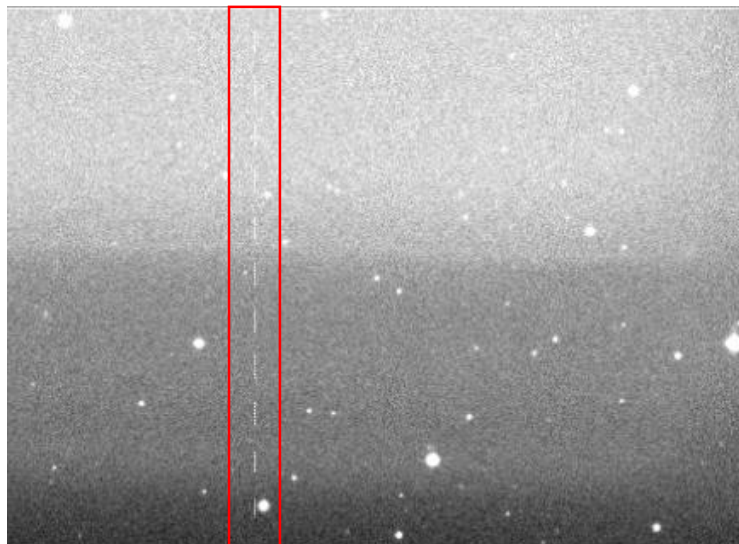
- CNNのerror CNNはどこ見てTrueと判断したのか

False Positive (sTMQ1201905100011144211) Bright star in the image ! (1)



False Positive データ汚染の調査

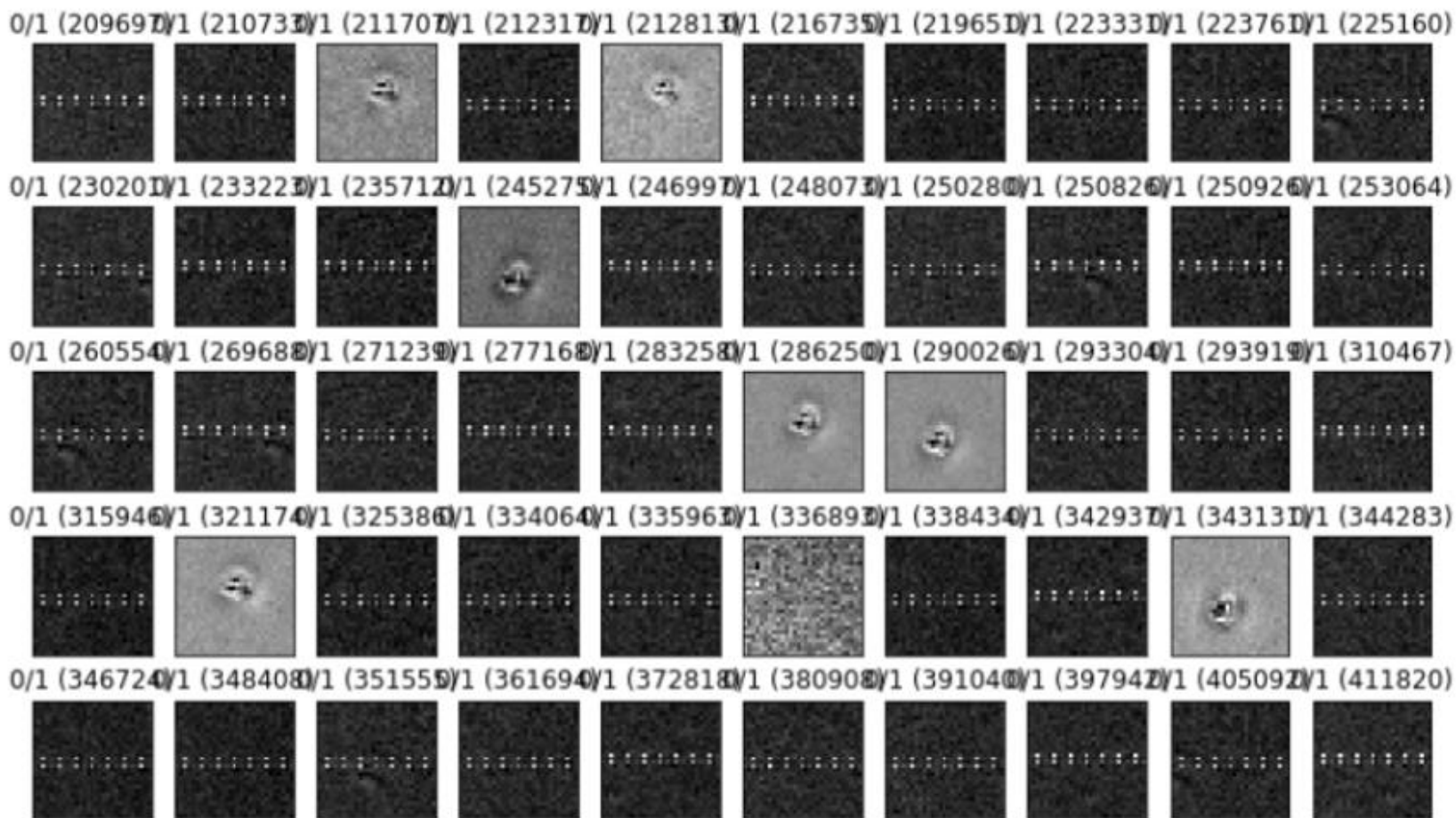
- CNNのerror (imageのエラー)



False Positive データ汚染の調査

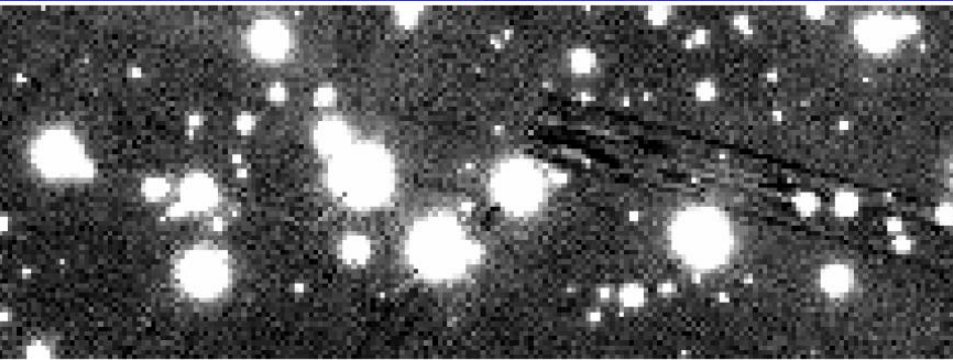
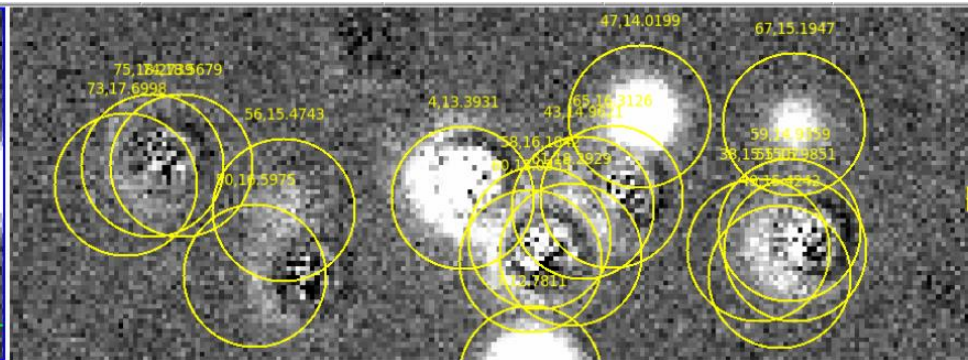
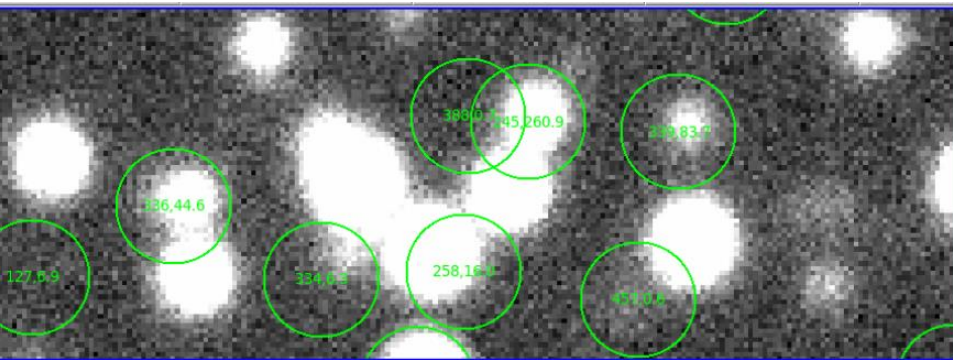
- CNNのerror CNNはどこ見てTrueと判断したのか

False Positive (sTMQ1201901220007635011) (1)



False Positive データ汚染の調査

- CNNのerror

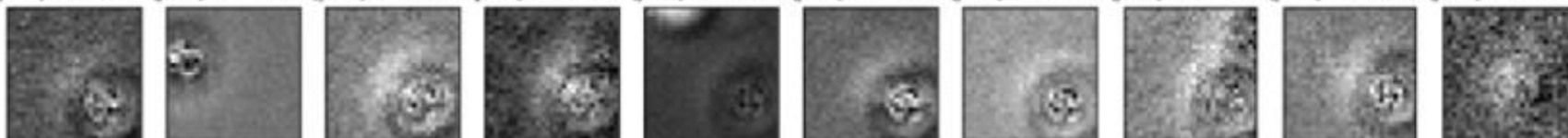


False Positive データ汚染の調査

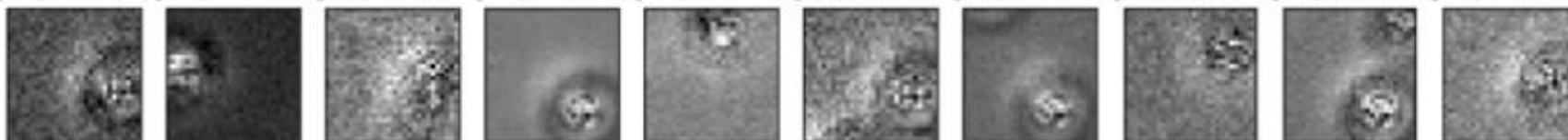
- CNNのerror CNNはどこ見てTrueと判断したのか

False Positive (sTMQ1201901050007090211) ??? (1)

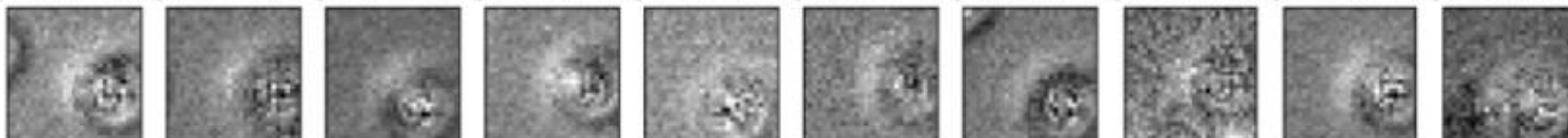
0/1 (210362) 0/1 (210917) 0/1 (231506) 0/1 (237885) 0/1 (246021) 0/1 (254624) 0/1 (255707) 0/1 (257605) 0/1 (270986) 0/1 (275632)



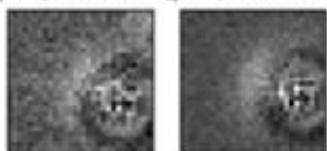
0/1 (286277) 0/1 (289981) 0/1 (292760) 0/1 (293177) 0/1 (293250) 0/1 (297301) 0/1 (319847) 0/1 (323910) 0/1 (330287) 0/1 (330851)



0/1 (331931) 0/1 (341479) 0/1 (342179) 0/1 (346476) 0/1 (349140) 0/1 (354794) 0/1 (362122) 0/1 (371407) 0/1 (382570) 0/1 (400100)



0/1 (403958) 0/1 (407543)



False Positive データ汚染の調査 まとめ

- ファイルごとのFalse Positive数 – TOP20file

含まれるFP天体の数...FP全体の約26%

大幅な位置ズレ : 6/20 改善済み

PSFの形が異常 : 6/20 (ホンモノなはずだが、labelがニセモノ)

天体数が多いimageであり、差分が失敗している : 5/20

CNNが拾ってしまったゴミ : 3/20

Number of False Positive : 3942

Top 10

file name : STMQ1201902230008479111, count : 148
file name : STMQ1201902210008345011, count : 73

Next >>> 差分方法 (hotpants) を変えてみる

file name : STMQ1201903290009619411, count : 63
file name : STMQ1201903290009629011, count : 63
file name : STMQ1201902210008342311, count : 55
file name : STMQ1201901220007635011, count : 51
file name : STMQ1201903290009622611, count : 49

Next >>> CNNがどこ見て判断したのかを見してみる

Top 11 - 20

file name : STMQ1201903190009423811, count : 45
file name : STMQ1201902210008355311, count : 42
file name : STMQ1201902210008345011, count : 42
file name : STMQ1201901050007090211, count : 32
file name : STMQ1201903220009487411, count : 31
file name : STMQ1201903290009618011, count : 31
file name : STMQ1201903290009611411, count : 29
file name : STMQ1201902010007790111, count : 28
file name : STMQ1201902210008352911, count : 26
file name : STMQ1201903290009617111, count : 26

STMQ1201903290009617111 : PSF error
STMQ1201902210008355311 : SUB error (Garactic plane ?)
STMQ1201902210008345011 : SUB error (Garactic plane ?)
STMQ1201901050007090211 : CNN error ? ----- Check image later
STMQ1201903220009487411 : Misalignment ----- Check image later
STMQ1201903290009618011 : PSF error ----- Check image later
STMQ1201903290009611411 : PSF error ----- Check image later
STMQ1201902010007790111 : Misalignment
STMQ1201902210008352911 : SUB error (Garactic plane ?)
STMQ1201903290009617111 : PSF error

STMQ1201902230008479111 : Misalignment ?
STMQ1201905100011144211 : CNN error (Bright star !)
STMQ1201902210008350011 : SUB error (Garactic plane)
STMQ1201901220007635011 : CNN error (Background is high brightness)
STMQ1201903170009242611 : Misalignment
STMQ1201902170008208811 : Misalignment
STMQ1201903290009619411 : PSF error
STMQ1201903290009629011 : PSF error
STMQ1201902210008342311 : Misalignment
STMQ1201903290009622611 : PSF error

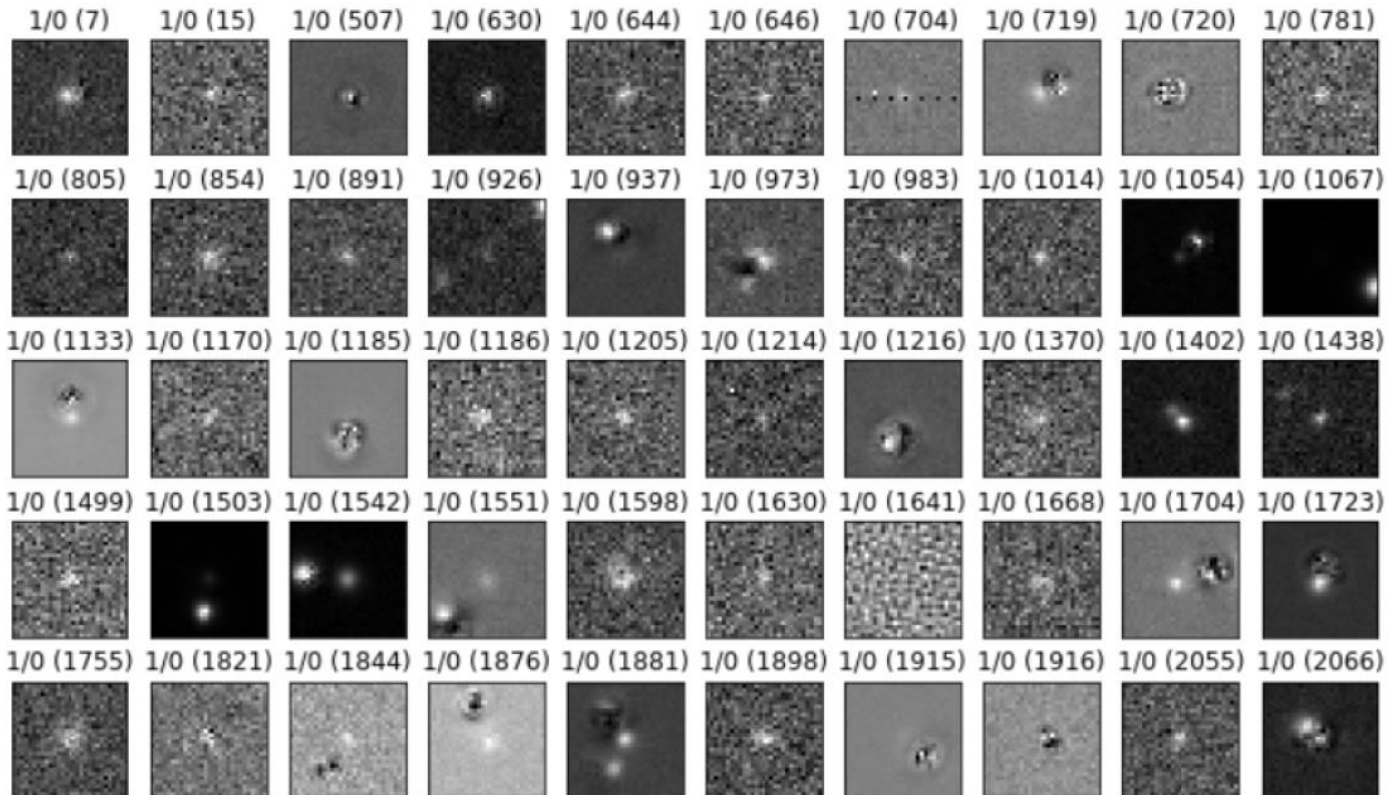
付録4 : False Negativeについて

3. False Negative の image について

- 新Model (DET Q1 1-1)

差分失敗天体と被ってるFAKE(ホンモノ)が多い。
→ 差分失敗した母銀河の超新星が拾えないことを示唆している。

False Negative (test set) (1)



• まとめ

3. 展望・問題点

- 今のところ、 $S/N > 20$ については自動アラートが実施できる可能性がある？ (DET Q1 1-1)
- 明るい天体付近の差分失敗天体、差分画像に写った本物のTransient天体、
小惑星による、False label 学習データの汚染
- S/N の低いFalse labelをTrueと判別している。
- 差分失敗した母銀河の超新星を拾うことが困難か？

>>> Next

差分失敗天体を減らすために、差分に使うツールを検討 (hotpants → ???)

CNNに Reference, Science, Subtraction の画像を用いて、性能が向上するか

差分失敗した母銀河の超新星検出のために、銀河のカタログ情報を用いる

付録5:最新資料

• 目次

1. 機械学習によるTransient天体の検出とは？

1. Transient天体の数
2. Transient天体の検出手順
3. Modelの学習
4. Modelの評価について

2. 学習結果

1. 学習データ
2. ROC curve
3. False Positive の S/N について

3. 結論と展望

1. 今後の展望
2. まとめ

2. 学習結果 DET Q1 1-1

1. 学習データ

- Label: **Artificial Real**(ホンモノのTransientの代わりとして), **Artifact**
- データセット

Train-Data : 109,240 (**54,620** : **54,620**)

Validation-Data : 54,620 (**27,310** : **27,310**)

Test-Data : 54,622 (**27,311** : **27,311**)

- 結果

True Positive : 26,673

False Positive : 552

True Negative : 26,759

False Negative : 638

2. 学習結果 DET Q1 1-1

2. ROC curve

Threshold

FPR = 2.0 % (at) TPR = 97.7 %

Artifact: 2万個 → ~400個

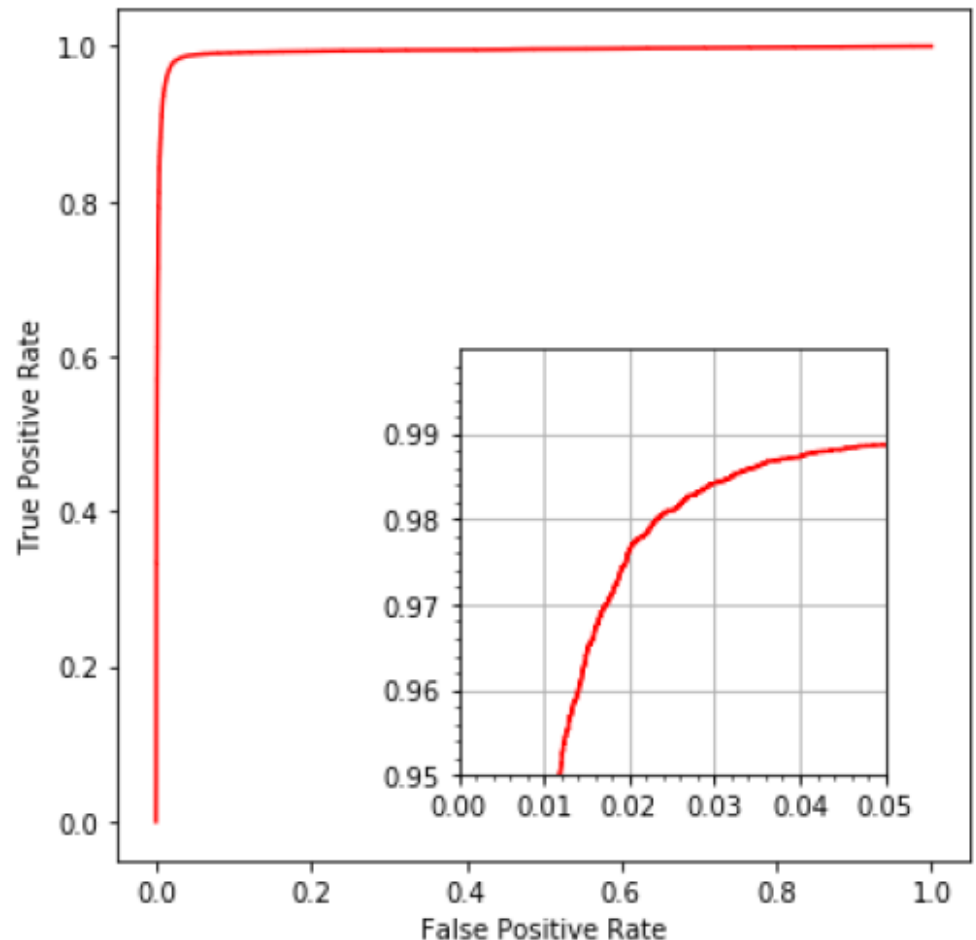
$TPR = TP / (TP + FN)$

全Artificial Realのうち、
Modelが拾えた割合

$FPR = FP / (FP + TN)$

全Artifactのうち、
Modelが拾ってしまった割合

ROC curve for Test Set (AUC=0.993)



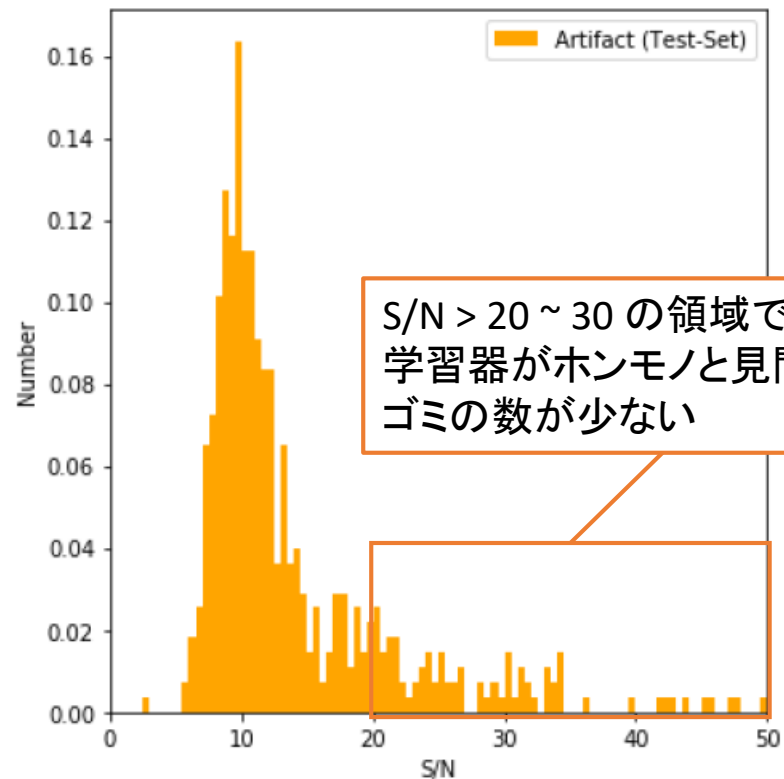
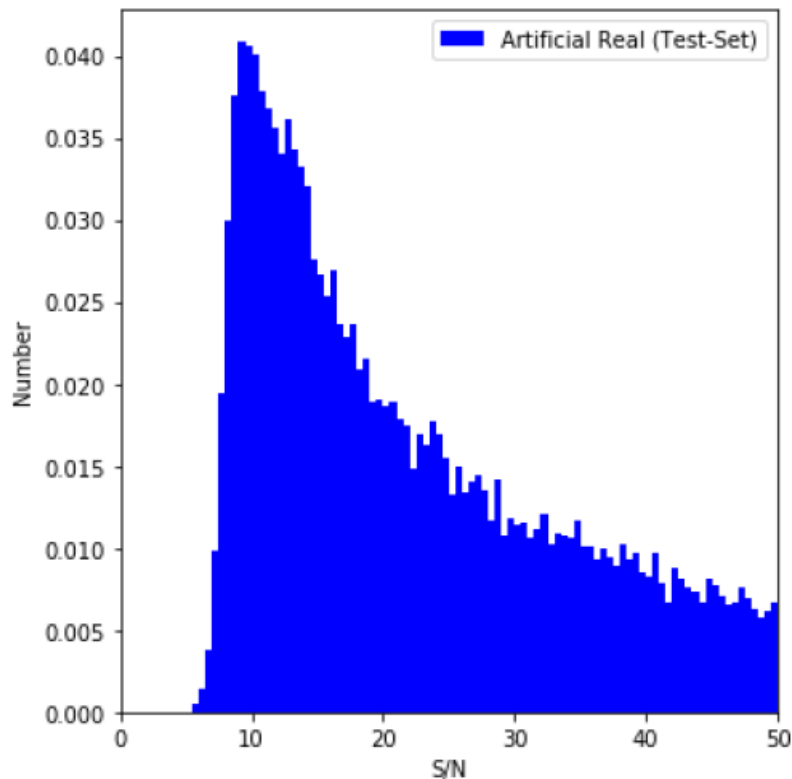
2. 学習結果 DET Q1 1-1

3. False Positive の S/N について

(左図) 学習器がホンモノと判定したTestデータのうち、Label : ホンモノ

(右図) 学習器がホンモノと判定したTestデータのうち、Label : ゴミ

Histogram of S/N using input (class = True)



2. 学習結果 DET Q1 1-1

3. False Positive の S/N について

- S/N に注目して性能評価

分類結果を $S/N = 20$ を基準に
ROC curve を比較。

$S/N > 20$: **AUC = 0.996**

$S/N < 20$: AUC = 0.989

(> 20 : < 20)

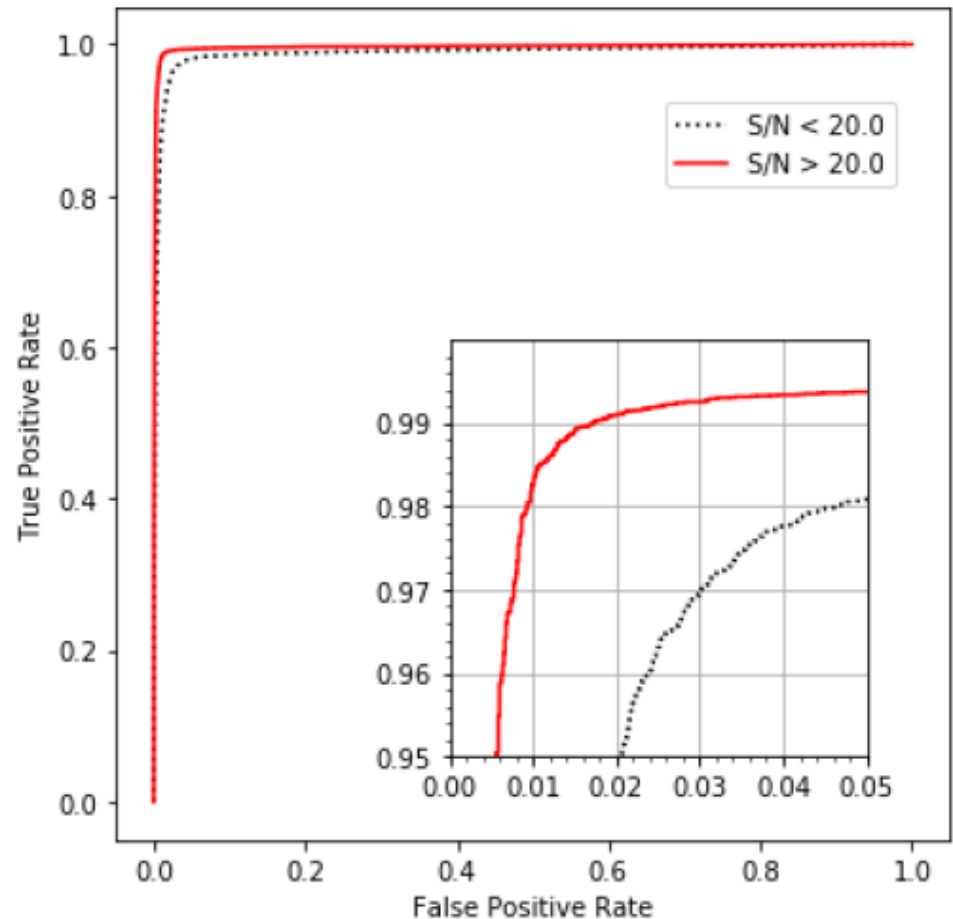
TP = 16598 : 10054

FP = 131 : 417

TN = 7367 : 19396

FN = 171 : 488

ROC curve (AUC = 0.996 / 0.989)



2. 学習結果 DET Q1 1-1

3. False Positive の S/N について

- S/N に注目して性能評価

S/N > 20 : AUC = 0.9961

S/N > 30 : AUC = 0.9967

S/N > 40 : AUC = 0.9970

(> 20 : > 40 : > 60)

TP = 16598 : 12586 : 9851

FP = 131 : 79 : 58

TN = 7367 : 4591 : 3416

FN = 171 : 108 : 82

ROC curve (AUC = 0.9961 / 0.9967 / 0.9970)

