# 인공지능을 이용한 암 패턴 분석

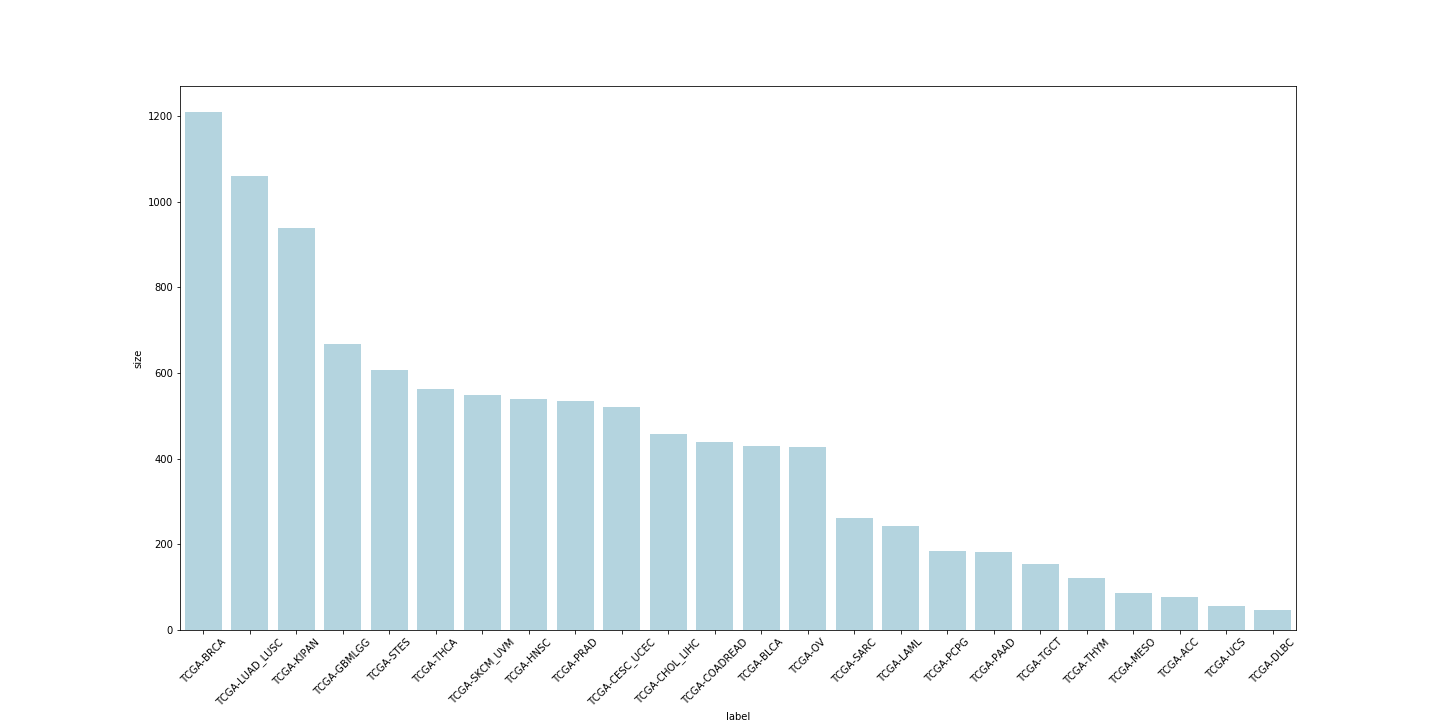
## 배경설명

유전자는 단백질을 암호화하고 단백질은 세포의 기능을 결정한다. 따라서 특정 세포에서 발현된 수많은 유전자는 그 세포가 무엇을 할지를 결정한다. 이렇듯 세포의 역할을 결정하는 유전자의 발현 패턴은 정상과 암을 구분하고, 어떤 암인지를 판별할 수 있는 데이터가 된다. 2006년에 시작한 The Cancer Genome Atlas (TCGA) 프로젝트는 대규모 유전체 시퀀싱 및 게놈분석 기술을 통해 암의 분자생물학적 기초에 대한 이해를 촉진하기 위해 표준화된 방법으로 33개 암종의 다양한 데이터를 연구자들에게 제공한다. 이 중 유전자 발현에 관련된 gene expression, DNA methylation, microRNA expression 데이터를 이용하여 인공지능 기술을 통해 암 패턴 분석을 진행해 보았다.

## 데이터

TCGA는 33개의 암종에 대해서 10,361( ? 10,255)명의 암 유전체 데이터를 방대하게 모아 놓았다. 우리는 샘플 수를 늘리기 위해 조직 위치를 기반으로 24개의 암종으로 합병을 시도하였다.

|  |  |  |  |
| --- | --- | --- | --- |
| 번호 | 암종 코드 | 암종 | 샘플 수 |
| 1 | TCGA-BRCA | 유방암 | 1208 |
| 2 | TCGA-LUAD\_LUSC | 폐선암, 폐편평상피세포암 | 1117 |
| 3 | TCGA-KIPAN | 신장암 | 1008 |
| 4 | TCGA-GBMLGG | 신경교종 | 690 |
| 5 | TCGA-STES | 위-식도암 | 574 |
| 6 | TCGA-THCA | 갑상선암 | 567 |
| 7 | TCGA-PRAD | 전립선암 | 546 |
| 8 | TCGA-HNSC | 두경부암 | 544 |
| 9 | TCGA-CESC\_UCEC | 자궁경부암, 자궁내막암 | 511 |
| 10 | TCGA-SKCM\_UVM | 악성흑색종, 맥락막흑색종 | 469 |
| 11 | TCGA-CHOL\_LIHC | 담도암, 간암 | 466 |
| 12 | TCGA-COADREAD | 대장암 | 434 |
| 13 | TCGA-BLCA | 방광암 | 426 |
| 14 | TCGA-OV | 장액성 난소상피암 | 378 |
| 15 | TCGA-SARC | 육종 | 264 |
| 16 | TCGA-PCPG | 부신암 | 184 |
| 17 | TCGA-PAAD | 췌장암 | 182 |
| 18 | TCGA-TGCT | 고환암 | 154 |
| 19 | TCGA-LAML | 급성골수성백혈병 | 147 |
| 20 | TCGA-THYM | 흉선종 | 120 |
| 21 | TCGA-MESO | 악성중피종 | 86 |
| 22 | TCGA-ACC | 부신피질암 | 77 |
| 23 | TCGA-UCS | 자궁육종 | 56 |
| 24 | TCGA-DLBC | 미만성거대B세포림프종 | 47 |



눈으로 살펴봐도 클래스 불균형(class imbalance)이 보이기에 ID(Imbalance Degree)를 측정해 보았습니다[1]. 클래스 불균형 문제는 1) balanced problem, 2) unbalnaced problem showing multi-majority, 3) a multi-minority unbalanced problem 세 가지로 분류되는데[2,3], ID 값으로 어떤 문제에 속하는지 알 수 있습니다. 값은 11.56으로 unbalanced problem showing multi-majority 와 a multi-minority unbalanced problem 중간 정도로 imbalance 한 것을 확인했습니다.

우리는 이러한 클래스 불균형에 의한 작은 샘플 수를 가진 클래스 예측률 저하 문제를 해결하기 위해서 data-level methods의 접근 방법을 선택했고, sampling이 아닌 여러 타입의 데이터로 feature selection으로 모델의 예측 성능을 개선하는 전략을 취했습니다. 물론 sampling이나 algorithm-level methods 또는 hybrid methods 기반의 전략도 좋지만, 이번에는 논외로 합니다.

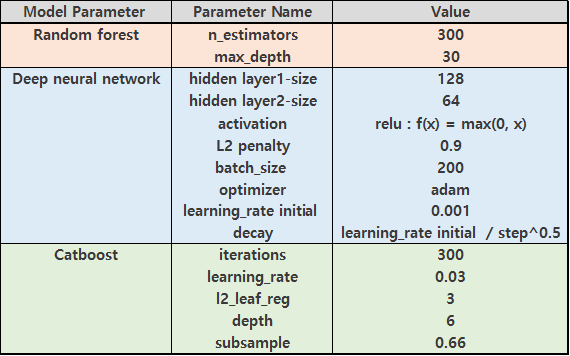
## Feature 선택

RNA는 약 2만개(gene expression), methylation은 약 48만개(DNA methylation) 정도의 feature가 있습니다. 우리는 이 중 1,000개에서 5,000개까지 선택을 한 후에, 같은 숫자의 feature 개수 시 어떤 타입의 데이터가 예측하는 데 유리한지, 그리고 얼만큼의 feature 수가 최적인지를 알아보려고 했습니다.

\*\* feature 선택 방법 선택 기준은 데이터값 자체의 variance, xgboost 및 lightgbm의 그래디언트 부스팅 모델의 트리 생성 시 사용하는 엔트로피값 기준, 그리고 Support Vector Machine에 L1 페널티를 적용하여 feature를 선정하고, 이것을 weighted voting 방식으로 선택했습니다.

## 모델

불균형 데이터에는 Bagging 및 Boosting 기반의 알고리즘이 잘 동작한다고 알려져 있고 많이 사용하는 방법입니다[5,6,7]. 따라서 Bagging 방식의 Random Forest, Boosting 방식의 Catboost, Neural Network를 가지고 실험해 보았습니다. 각각의 모델 파라미터는 다음과 같습니다.



## 검증방법

클래스 불균형인 데이터라서 일반적인 accuracy로는 적은 샘플의 클래스에 대한 성능 측정이 어렵기 때문에, 주 성능 지표로는 aucPR(Area Under the Precision-Recall Curve)를 사용했습니다. (Average 방식은 Micro 사용) 그리고 데이터의 샘플 수가 많지 않기 때문에 5-fold cross validation으로 성능을 측정했습니다.

## 실험 결과

다음의 테이블은 각 데이터 세트별, 3가지 모델의 aucPR 평균값입니다. **com = RNA + methylation , rna = RNA , mtl = methylation** 을 나타내고, 뒤의 숫자는 feature 수를 나타냅니다.

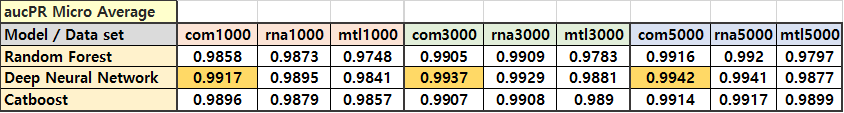


Figure 1: Compare aucPR micro average

현 실험에서는 딥러닝 모델(Deep Neural Network)이 각 feature 수의 데이터세트에 대해서 다른 모델들 보다 높은 성능을 보이는 것을 확인했습니다.

다음은 딥러닝 모델의 각 암종별 성능을 나타낸 테이블입니다. 파란색 음영은 각 값을 시각화 한 것이고, 빨간색 숫자는 각 암종에서 aucPR 값이 가장 높은 데이터 세트를 나타냅니다. 마지막 열 **Count**는 샘플 수를 나타냅니다.

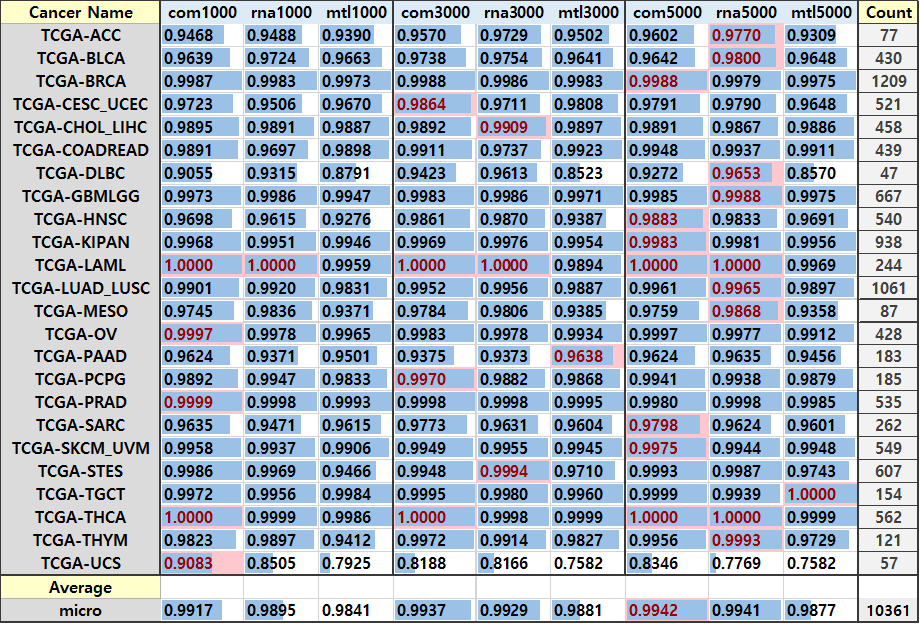


Figure 2: Compare each class aucPR across all dataset

feature 수가 1000개 일 때는 RNA + methylation 조합이, 2000, 3000개는 RNA + methylation 데이터와 RNA 데이터의 성능 이 비슷한 것을 볼 수 있었습니다. 모든 feature 수에 대해서 methylation 데이터는 다른 데이터들보다 낮은 성능이 나왔습니다.

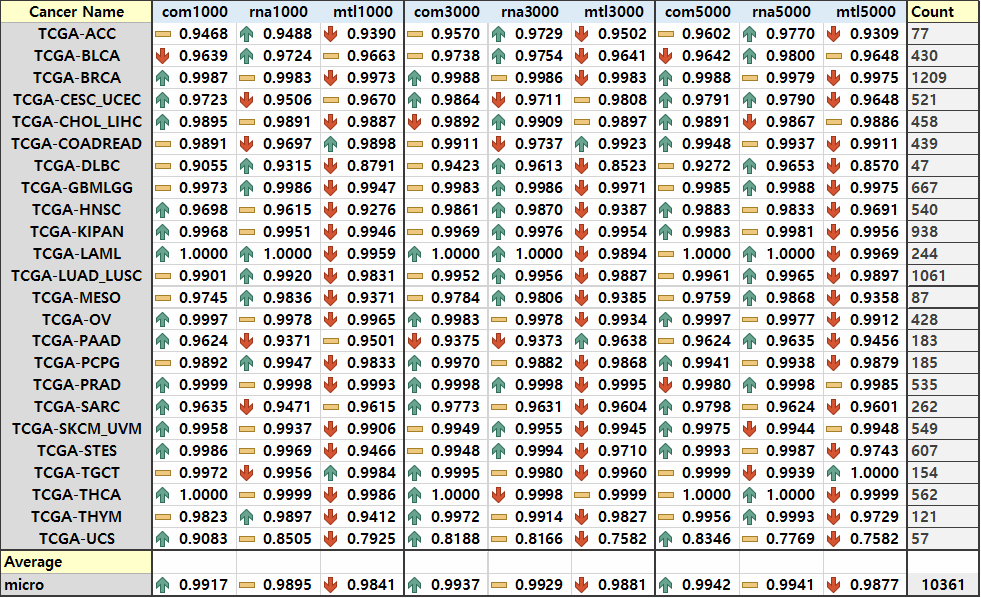


Figure 3: Compare each class aucPR in same feature count

평균적인 성능은 RNA + methylation 조합의 데이터 세트가 모든 feature 수 구간에서 가장 높은 aucPR이 나왔다. 특히 feature 수가 가장 적은 1,000개일 때 RNA와 methylation을 같이 구성 시 각 암종별 패턴이 더욱더 잘 표현되는 것을 알 수 있었습니다.

다음 결과는 Micro Precision-Recall Curve, 샘플 수가 가장 많은 BRCA, LUAD\_LUAC와 샘플 수가 가장 적은 UCS, DLBC를 시각화한 결과입니다.

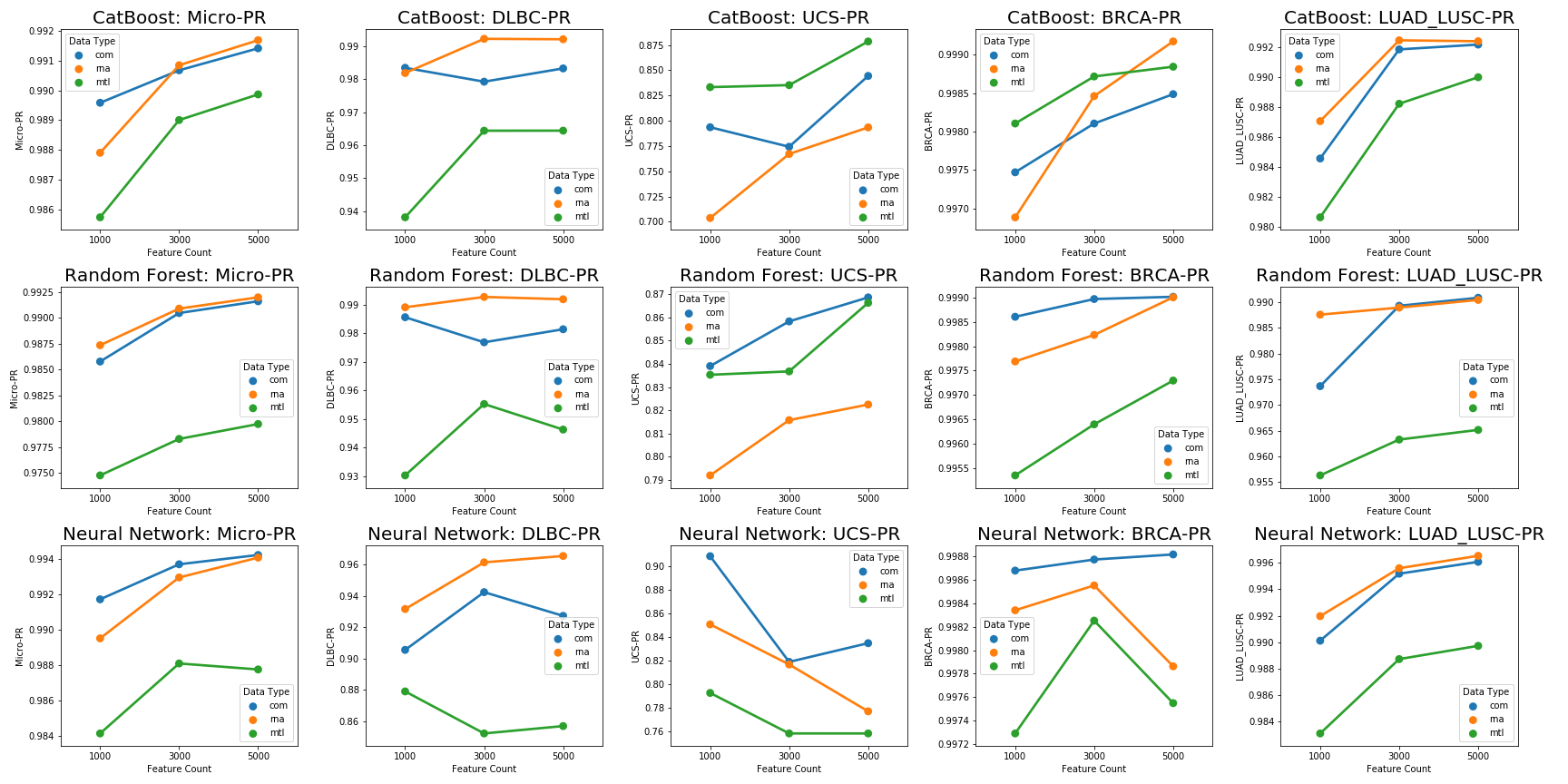


Figure 4: Compare aucPR Value on Average

성능의 평균치를 보았을 때는 feature 수가 많을수록 성능이 올라가는 것을 확인할 수 있었다.

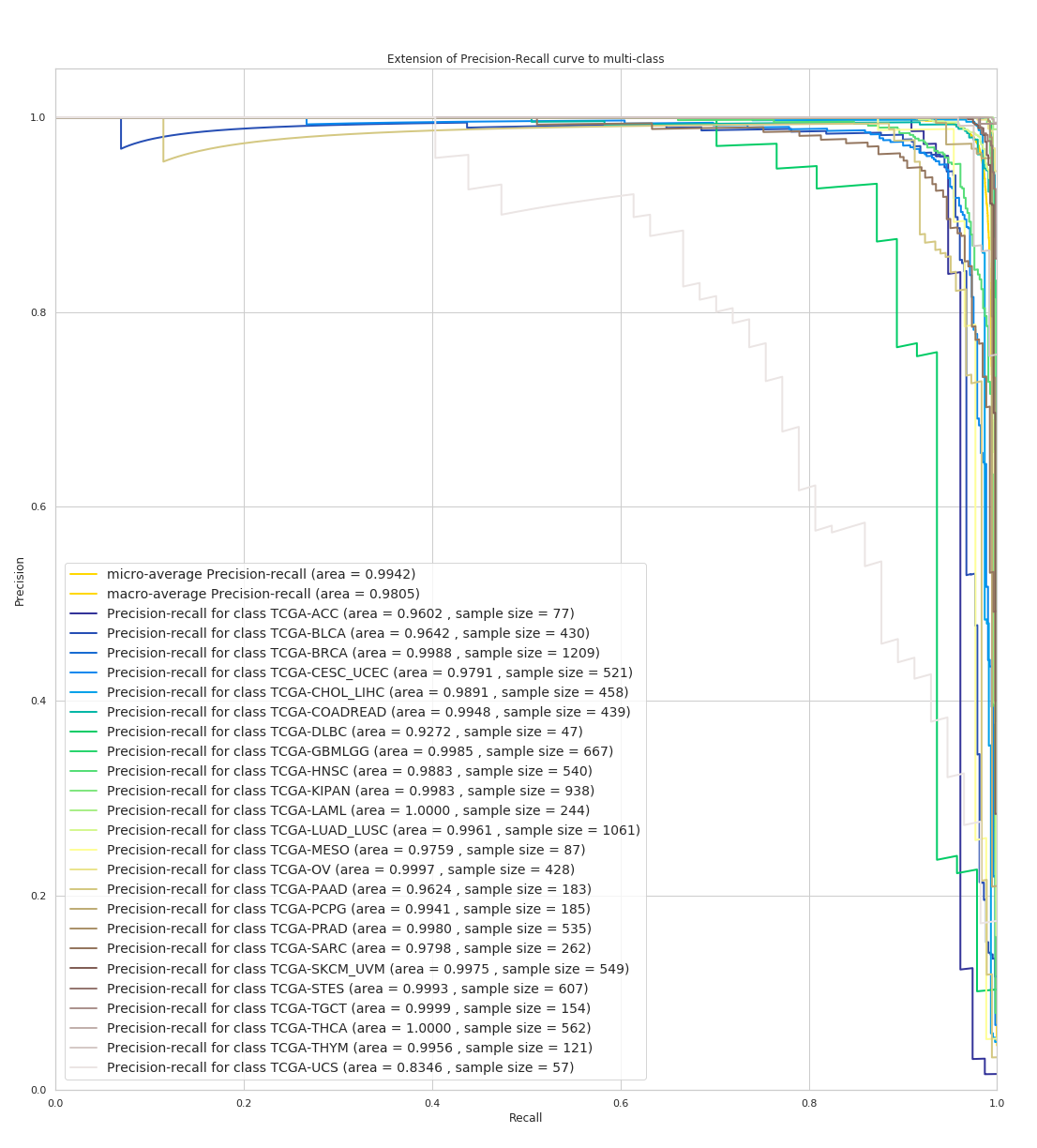


Figure 5: Deep Neural Network PR-Curve on RNA + Methylation , 5000 feature count

## 결론

TCGA의 24종 암종 데이터의 Class imbalance problem에서 작은 샘플 사이즈를 가진 클래스에 대한 예측률을 올리기 위해 RNA와 methylation 데이터를 같이 쓰면 어떠한 효과가 있는지 확인해 보았습니다. 이번 실험을 통해 다음과 같은 결론은 낼 수 있었습니다.

1. 현 실험에서는 같은 종류의 데이터를 사용 시 전반적으로 딥러닝(Deep Neural Network) 모델이 성능이 가장 좋았다.
2. 데이터의 종류로 비교를 해보면 RNA + methylation 구성의 데이터가 조금 더 암종 구분을 위한 패턴이 들어있다고 볼 수 있고, 암종에 따라서 RNA 또는 methylation 한종류로 구성한 데이터에서 더 높은 예측률을 보일 때도 있었습니다. 따라서 같은 feature 수인 경우 모델의 전반적인 예측률을 높이는데 가장 좋은 방법은 RNA와 methylation 둘 다 사용하는 것이라고 볼 수 있겠습니다. 현 실험에서는 RNA와 methylation의 비율은 5대5로 했지만, 최적화된 비율을 찾아낸다면 성능은 더 개선될 것입니다.
3. feature 수는 많을수록 평균적인 성능이 향상했습니다.
4. 여러가지 feature selection method를 비교 및 최적화하는 작업도 도움이 될 것입니다. 특히 이번에는 머신러닝 모델 기반의 선택 방법을 사용했지만, 도메인 지식 기반의 선택 방법도 같이 사용한다면, 좋은 결과를 끌어낼 수 있을 것입니다.

## Appendix A : Feature Importance Result

모든 값은 0~1 사이로 스케일링 했다.

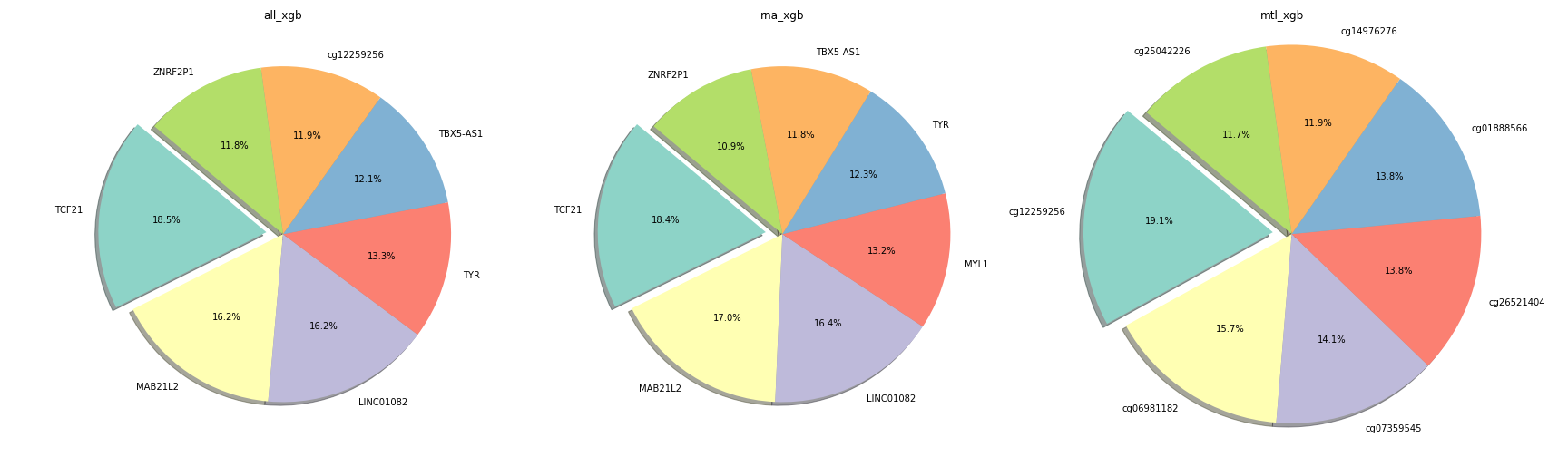


Figure A.1: Top7 Feature from lightgbm

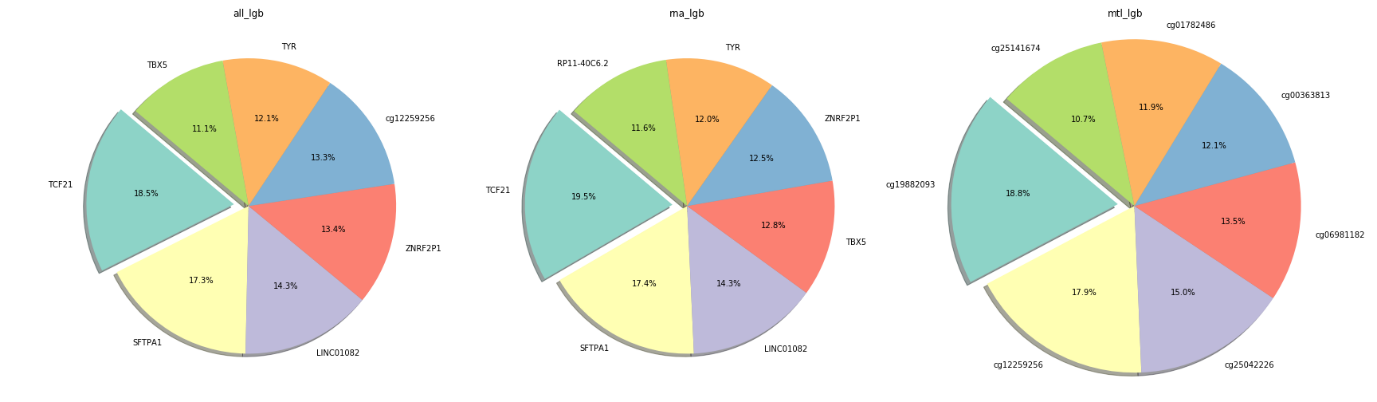


Figure A.2: Top7 Feature from xgboost

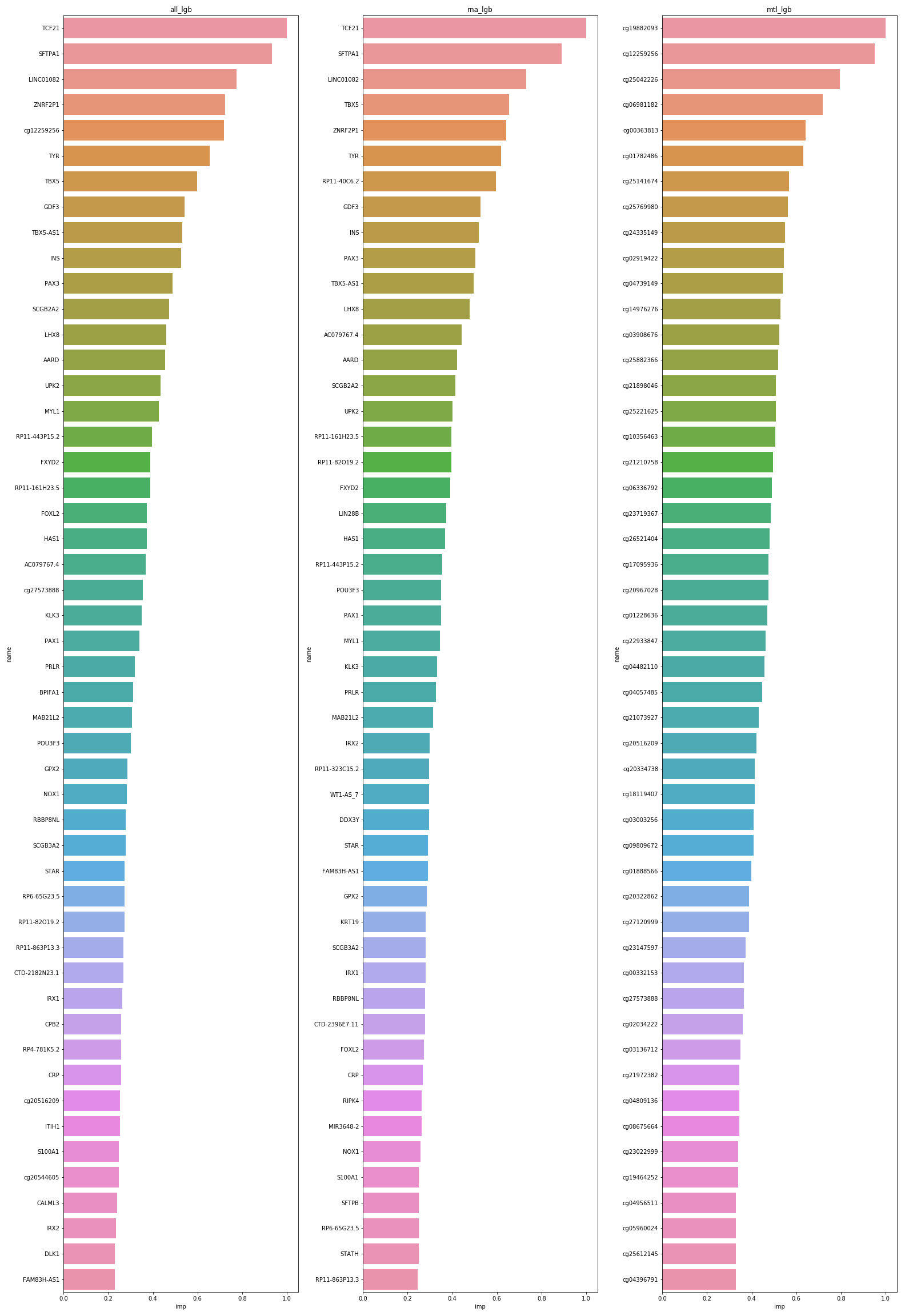
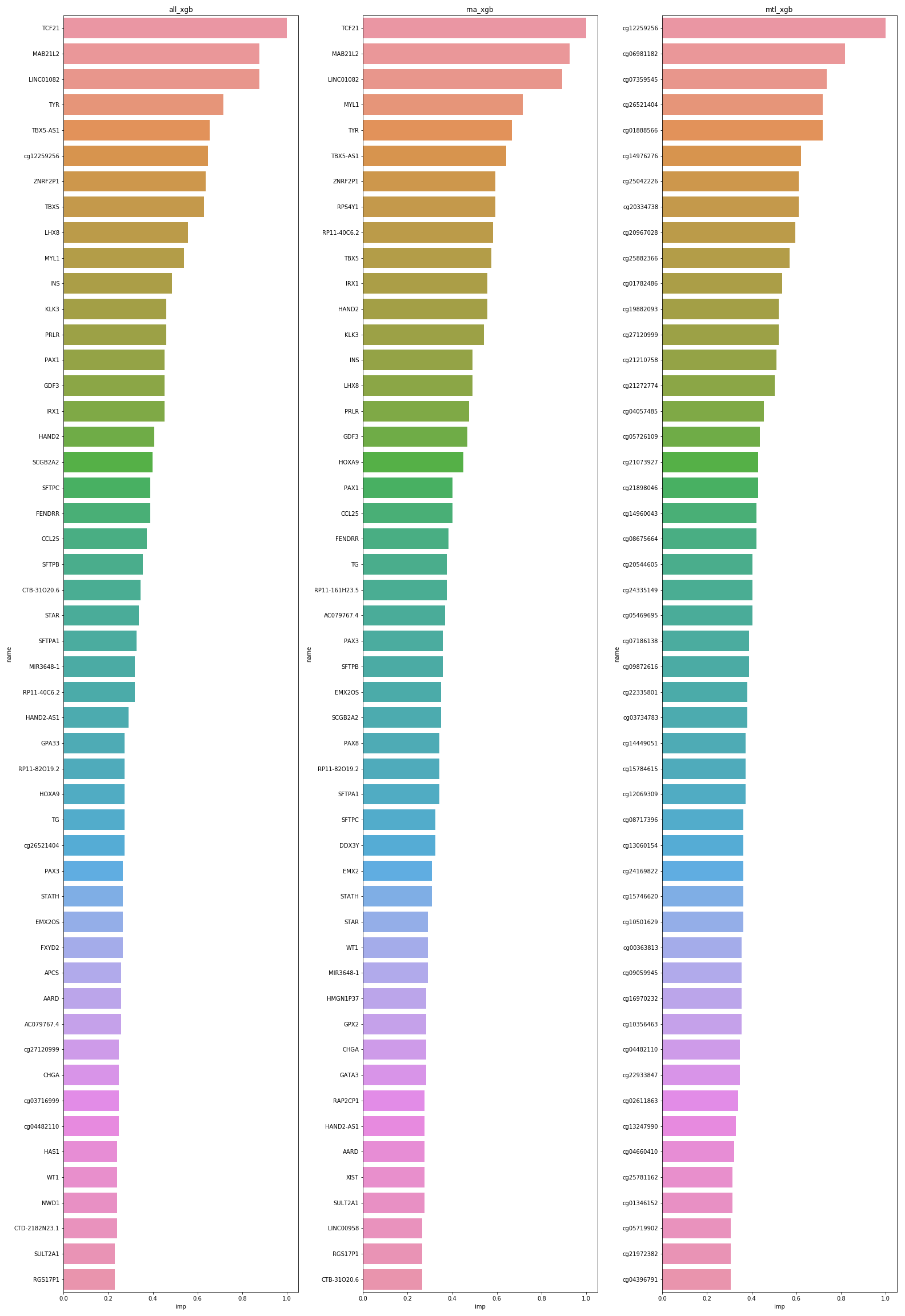
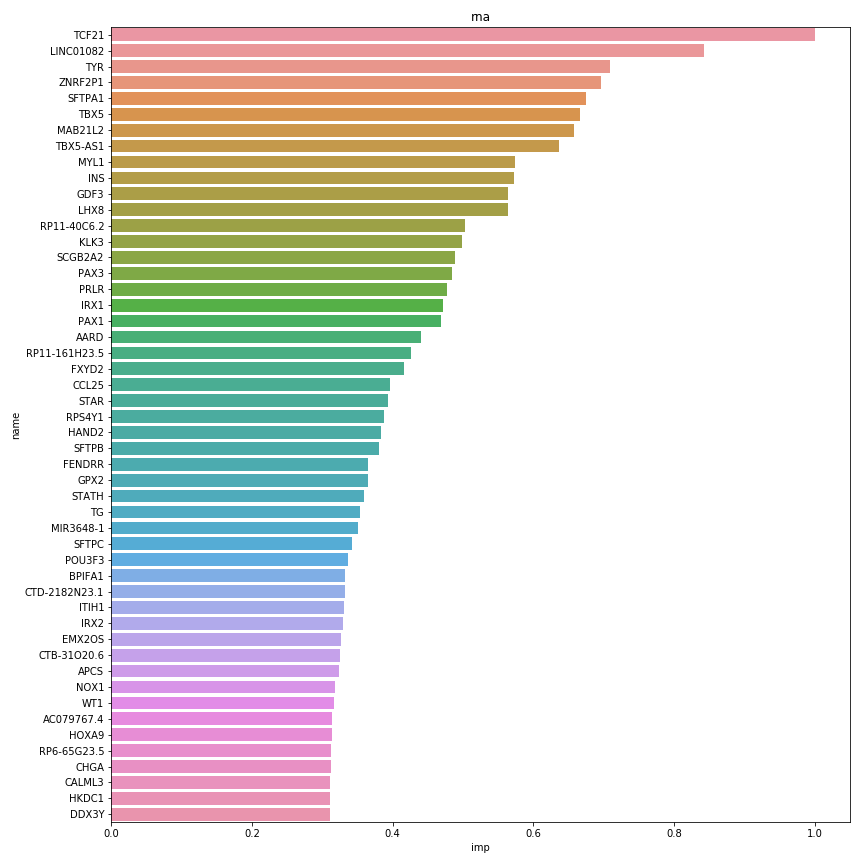
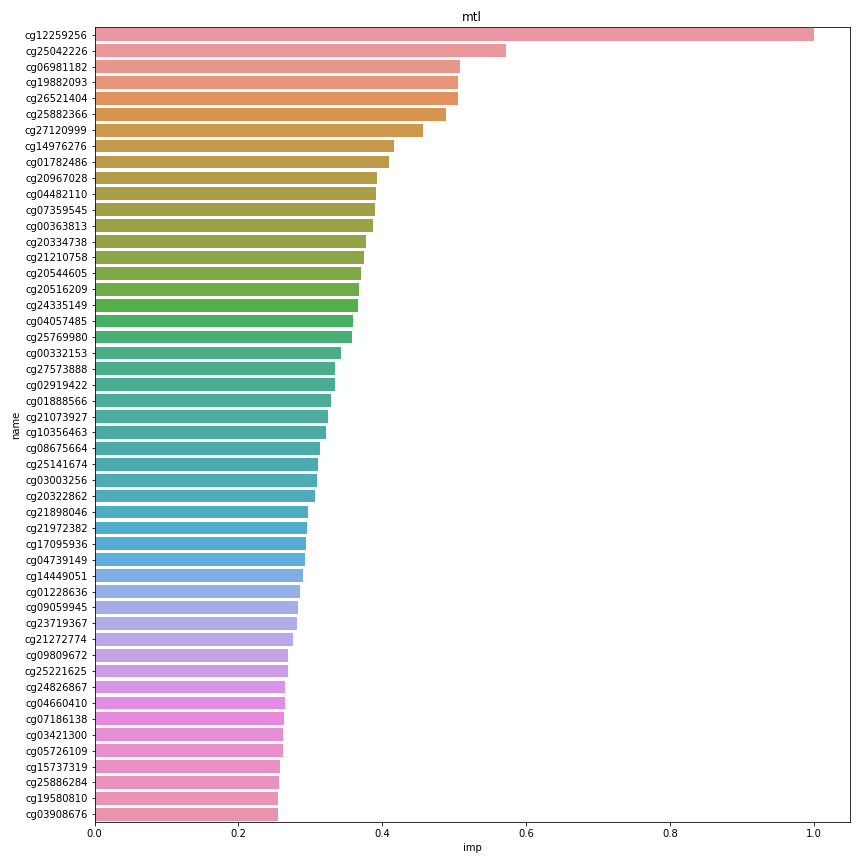


Figure A.3: Top50 Feature of RNA+Methylation , RNA , Methylation from lightgbm

Figure A.4: Top50 Feature of RNA+Methylation , RNA , Methylation from xgboost



Top50 Feature of mRNA from weighted voting



Top50 Feature of Methylation from weighted voting

## 앞으로 ???

추후 각 암종을 구분하는데 있어서 중요하게 작용한 feature들을 분석을 해보려고 합니다. 이러한 예측에 중요하게 기여하는 feature를 분석할떄는 feature들간의 관계를 분석이 불가능 하다는 한계가 있습니다.

#### 작업 미완 부분

* XGboost 결과 넣기 : 약 >4.5 시간 예상됨
* 중요도 분석 관련 의미 있으면 같이 써주기?
* 결론 방향
* TCF21 : 암에 관련 ( In humans, TCF21 has been identified as a candidate tumor suppressor gene and is frequently epigenetically silenced in various human cancers.[7])

[1]Ortigosa-Hernandez, J., Inza, I., Lozano, J.A., 2017. Measuring the class- ´ imbalance extent of multi-class problems. Pattern Recognition Letters 98, 32–38. [2] He, H., and Garcia, E. A. Learning from imbalanced data. IEEE Trans. on Knowledge and Data Engineering 21, 9 (Sep. 2009), 1263–1284. [3] Wang, S., and Yao, X. Multi-class imbalance problems: Analysis and potential solutions. IEEE Trans. on Systems, Man and Cybernetics - Part B 42, 4 (Aug. 2012), 1119–1130. [4]Blaszczynski, J., Stefanowski, J.: Neighbourhood sampling in bagging for imbalanced data. Neurocomputing 150, 529–542 (2015) [5]Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C 42(4), 463–484 (2012) [6]Krawczyk, B., Wo´ zniak, M., Schaefer, G.: Cost-sensitive decision tree ensembles for effective imbalanced classification. Appl. Soft Comput. 14, 554–562 (2014) [7] https://en.wikipedia.org/wiki/TCF21\_(gene)#Role\_in\_development