DACON 2022 AI 대학원 챌린지

백신 및 면역치료제 개발을 위한 항원-항체 반응 예측 solution

[Private : 0.76034]

중앙대학교 석사과정
윤건일
rjsdlf45@gmail.com

12th Aug, 2022

CAU AutoML Lab
http://ml.cau.ac.kr/

Contents

Preprocessing(Feature Engineering)

① Sequence Modeling

- Left antigen(64), epitope(128), right antigen(64) protein sequence modeling ESM(Pretrained-model)[1] 이용

② CT-CTD Feature extraction[2]

- Epitope, antigen sequence 데이터로부터 feature extraction
- Conjoint Triad (CT) features : 7x7x7 features for each sequence
- Composition-Transition-Distribution (CTD) : 13x3, 13x3, 13x3x5 features for each sequence

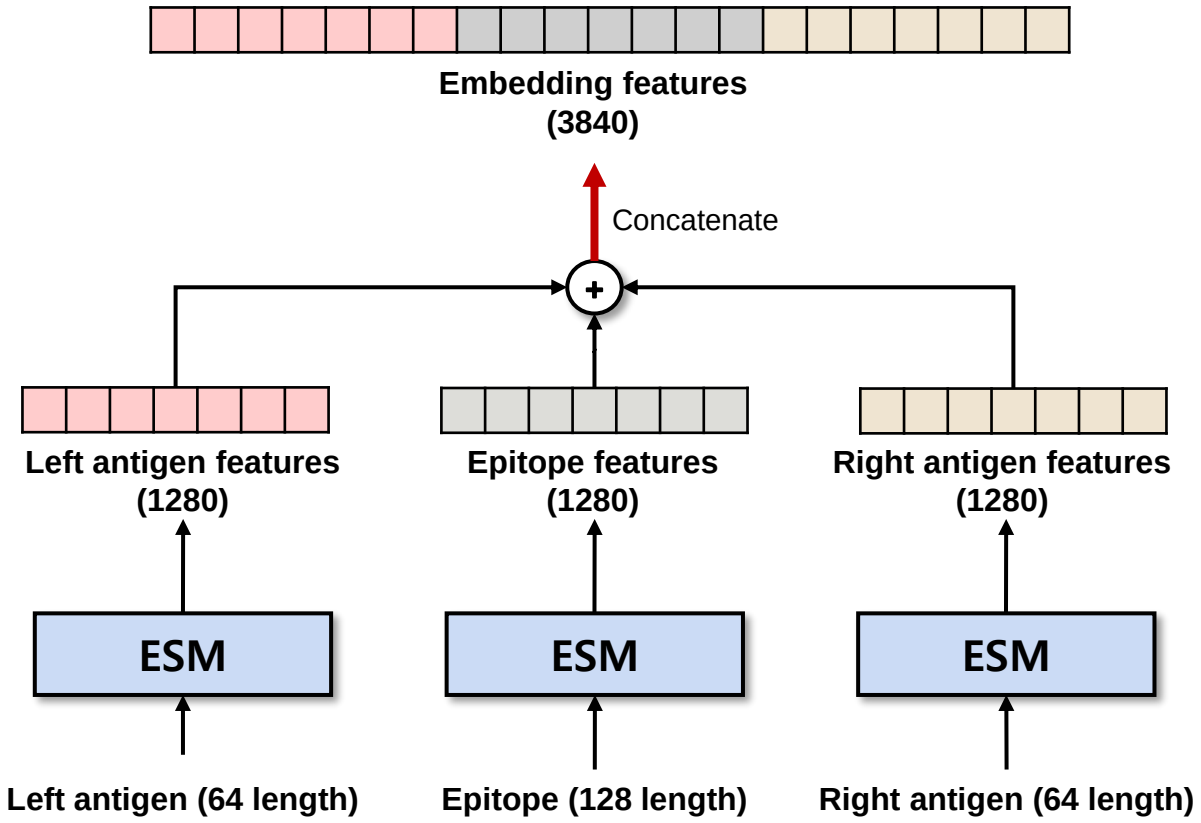
③ Counting Feature extraction[3]

- Epitope, antigen sequence 데이터로부터 feature extraction
- Single, Double features : frequency of amino acids and dipeptides
- 20+400 features for each sequence

Preprocessing(Feature Engineering)

① Sequence Modeling – ESM

- **ESM pretrained model(Encoder)에 left antigen 64, epitope 128, right antigen 64를 각각 입력해서 embedding 추출 (각 1280 features)**
- **각 sequence embedding을 concatenate해서 feature로 사용 (3840 features)**

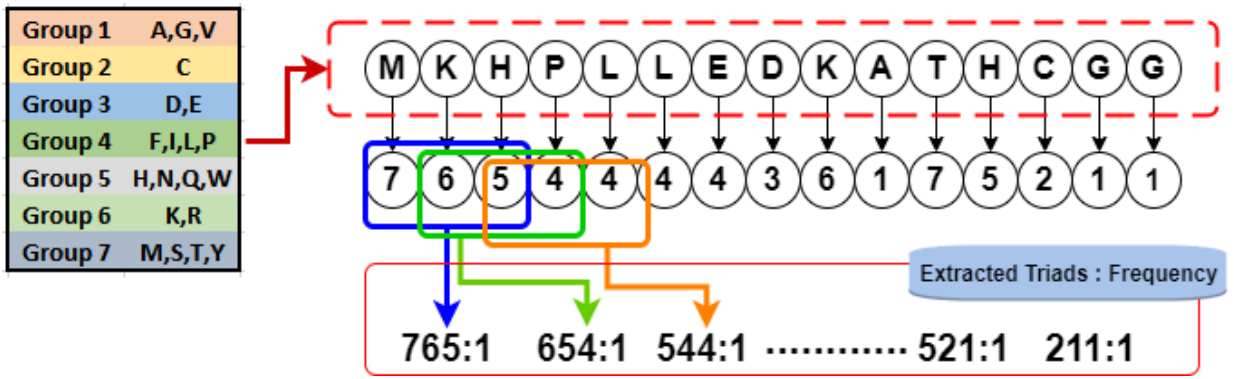


Preprocessing(Feature Engineering)

② CT-CTD features

- CT features : amino acids를 7개 group으로 나눠 연속된 3개(Triad)의 그룹 번호 frequency features
- Window size 3으로 sliding window를 통해 그룹번호 111부터 777까지 총 343개의 feature을 얻음

Group1	Group2	Group3	Group4	Group5	Group6	Group7
A, G, V	C	D, E	F, I, L, P	H, N, Q, W	K, R	M, S, T, Y



$$CT = (f_{111}, f_{112}, \ldots, f_{776}, f_{777})$$ 343 (7x7x7) features

Preprocessing(Feature Engineering)

② CT-CTD features

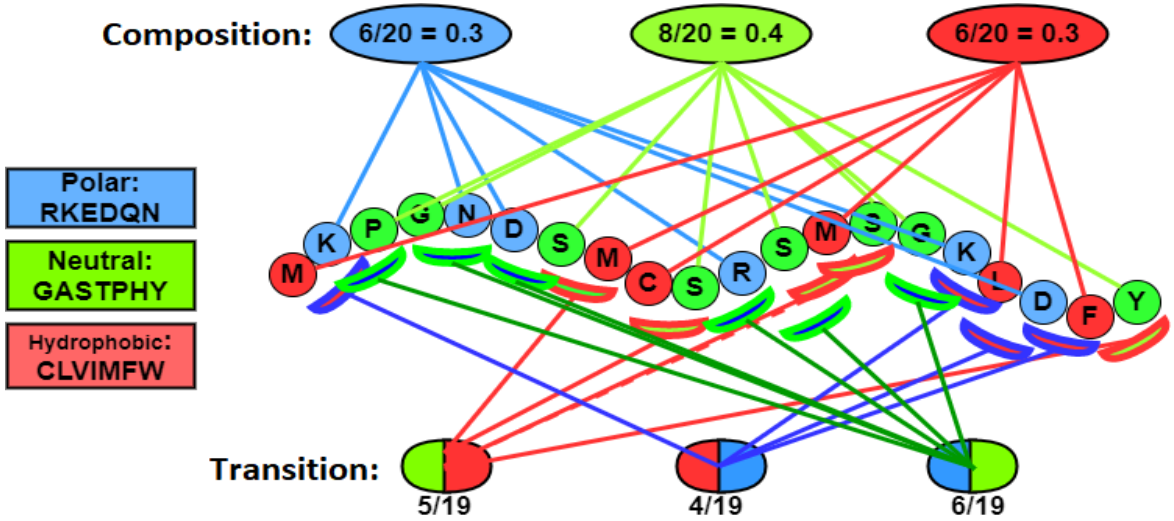
- **CTD features : Composition, Transition, Distribution feature 로 추출**
- **13개 구조적, 물리화학적 특성으로 각 특성마다 3개의 그룹(1, 2, 3)으로 분류**

Attributes	Division		
	Group1	Group2	Group3
hydrophobicity_PRAM900101	RKEDQN	GASTPHY	CLVIMFW
hydrophobicity_ARGP820101	QSTNGDE	RAHCKMV	LYPFIW
Hydrophobicity_ZIMJ680101	QNGSWTDERA	HMCKV	LPFYI
hydrophobicity_PONP930101	KPDESNQT	GRHA	YMFWLCVI
hydrophobicity_CASG920101	KDEQPSRNTG	AHYMLV	FIWC
hydrophobicity_ENGD860101	RDKENQHYP	SGTAW	CVLIMF
hydrophobicity_FASG890101	KERSQD	NTPG	AYHWVMFLIC
norm waals volume	GASCTPD	NVEQIL	MHKFRYW
polarity	LIFWCMVY	PATGS	HQRKNED
polarizability	GASDT	CPNVEQIL	KMHFRYW
charge	KR	ANCQGHILMFPSTWYV	DE
secondary struct	EALMQKRH	VIYCWFT	GNPSD
solvent access	ALFCGIVW	RKQEND	MPSTHY

Preprocessing(Feature Engineering)

② CT-CTD features

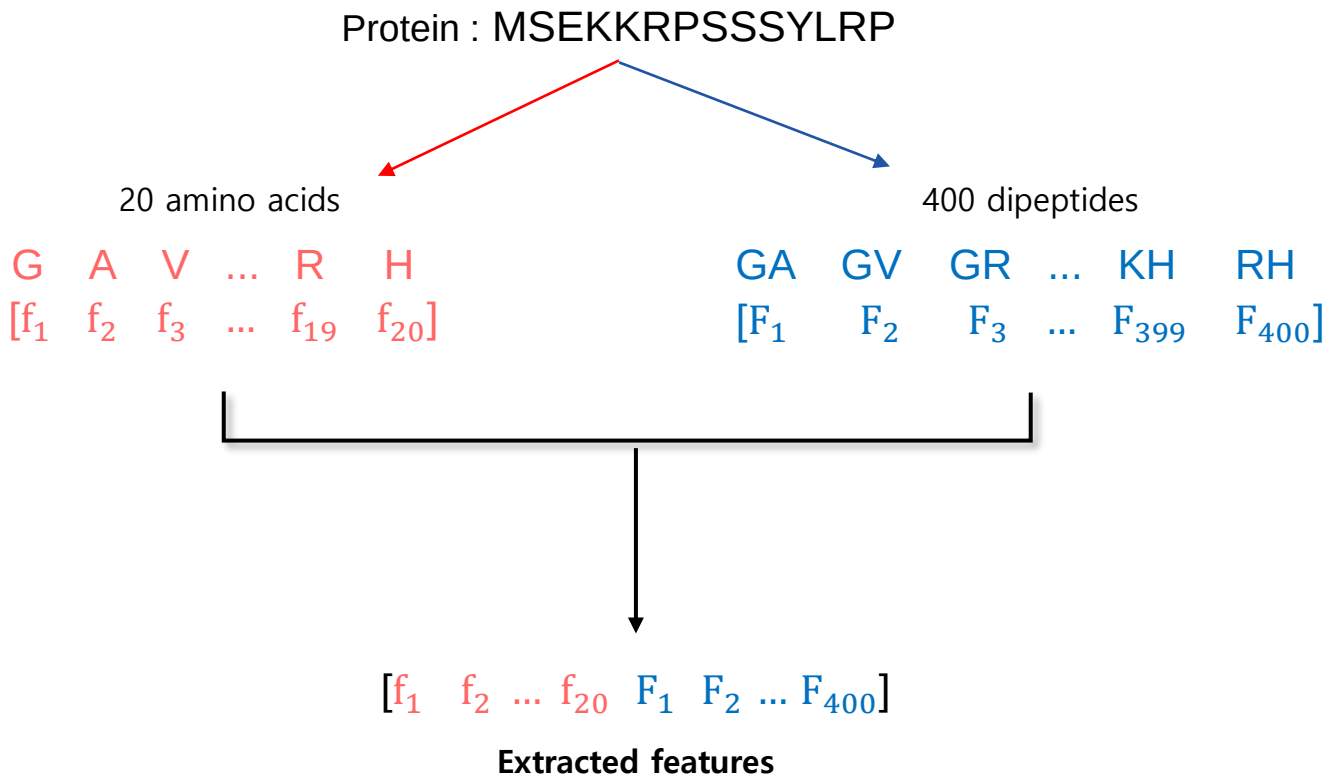
- 각 그룹 번호에 맞게 특성마다 새로운 시퀀스 부여
- Composition feature는 각 attribute division의 비율들 -> 13x3 features
- Transition feature는 그룹간 transition 비율을 의미 -> 13x3 features
- Distribution feature는 각 그룹의 position 위치(1, 25%, 50%, 75%, 100%) -> 13x3x5 features



Preprocessing(Feature Engineering)

③ Counting Feature extraction

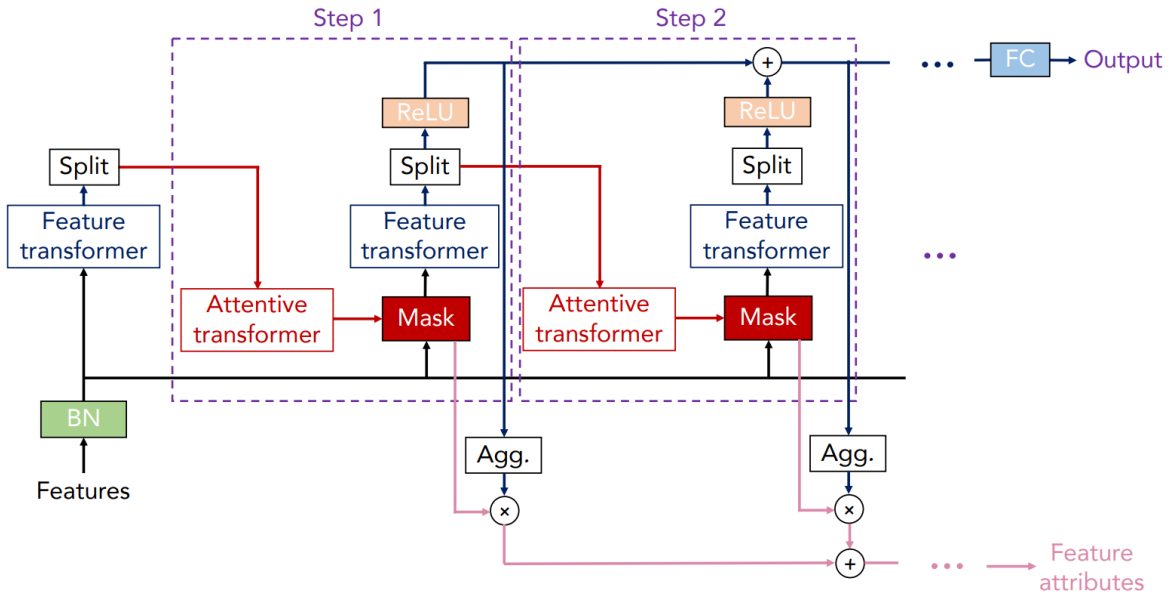
- **Protein sequence에서 20개의 amino acids 와 400개의 dipeptides frequency를 추출**



Model

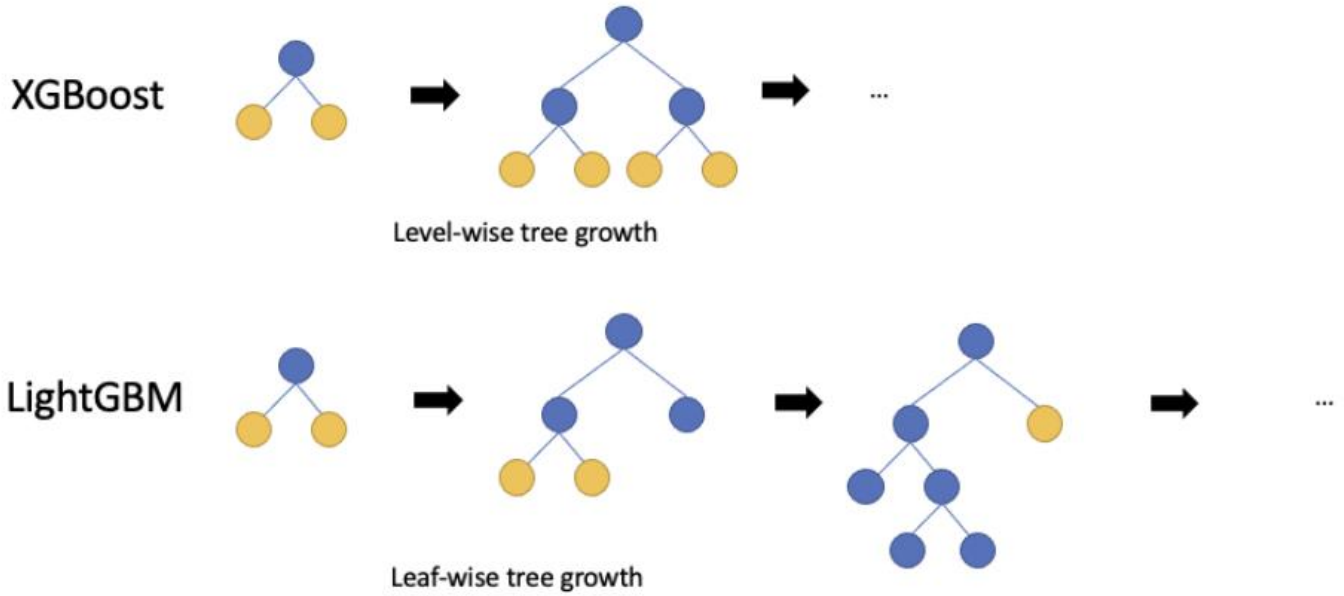
① Tabnet[4]

- AAAI에 발표된 Tabular 데이터 학습 모델로 최근 여러 대회에서 RF, XGBoost, LGBM 등의 트리 기반 모델들보다 좋은 성능을 보여주고 있음
- Feature Transformer 와 Attentive transformer 기반의 모델 구조를 통해 Tabular data에서 딥러닝 모델의 성능을 크게 끌어올린 모델
- Sparse matrix를 이용해 masking하고 학습하며 중요 feature를 스스로 학습할 수 있는 모델



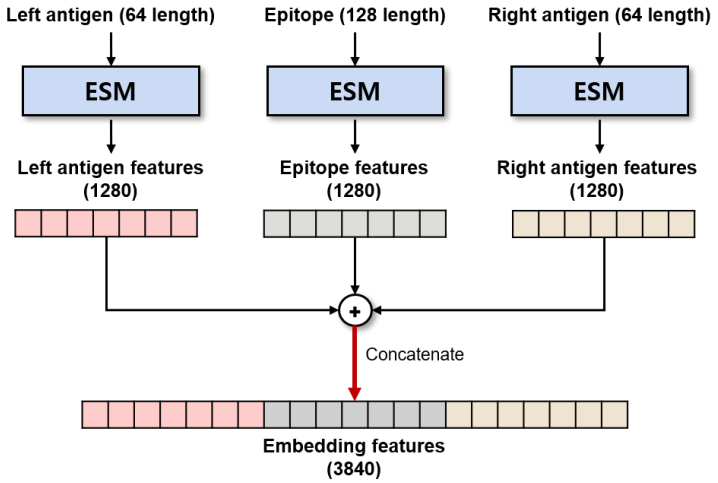
② LightGBM

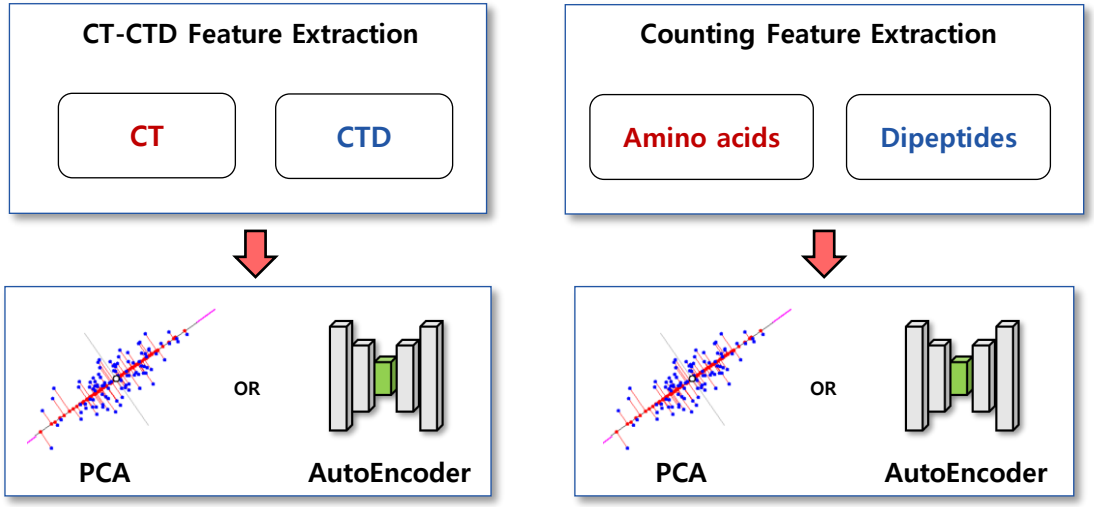
- 다른 트리 기반 학습 모델보다 빠르게 학습되고, 많은 데이터에 대해 강력함
- Leaf-wise 확장을 통해 level-wise 확장 모델보다 빠르게 loss를 수렴시킴
- 데이터가 LGBM이 학습하기에 충분히 많기(190k)때문에 Overfitting될 확률이 낮고 정확도에 초점을 맞출 수 있음



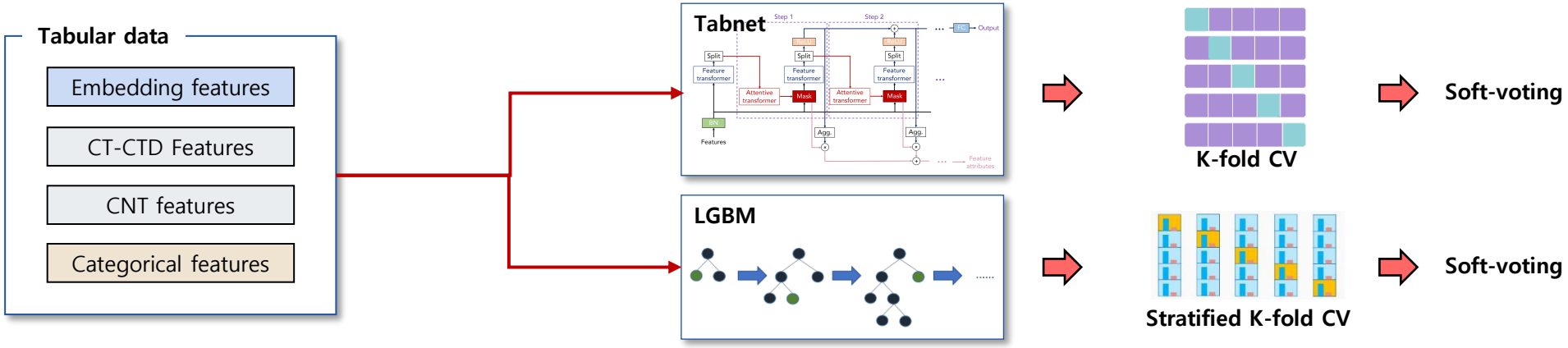
Framework

Preprocessing





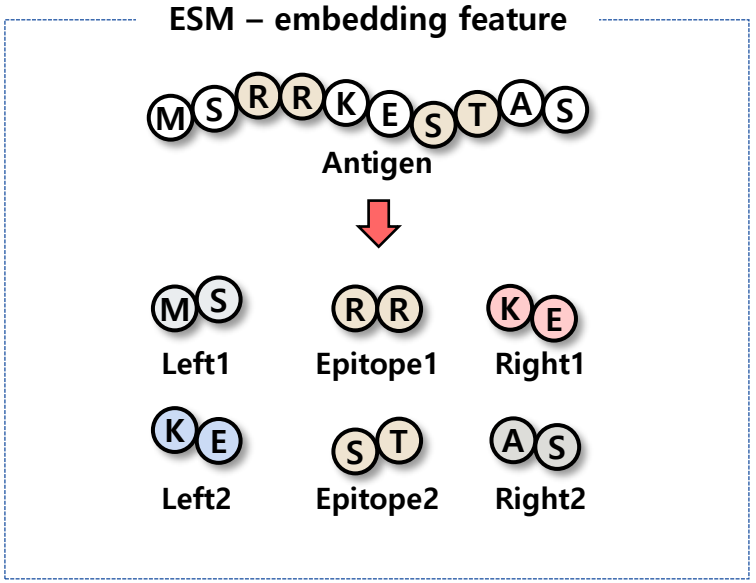
Model Training

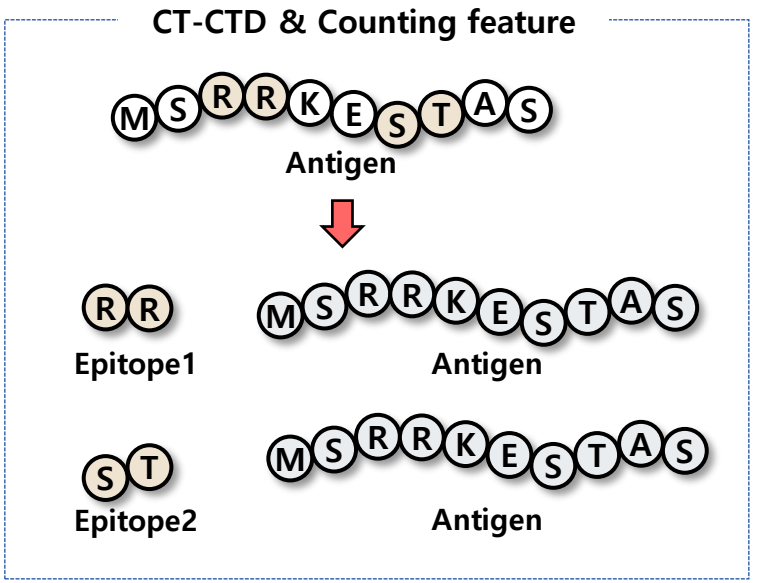


Experimental setup

✓ Feature Combination

- ESM-embedding feature는 epitope를 기준으로 antigen을 slice 하기 때문에 **같은 antigen이더라도 다른 정보**를 가짐
- CT-CTD feature와 Counting Feature는 epitope 전체와 antigen 전체 서열에 대해 적용하기 때문에 **같은 antigen에서는 feature가 같음**
- ESM-embedding을 기본적으로 적용하되, CT-CTD feature 혹은 Counting Feature의 조합으로 결과를 확인하고 **최종 모델을 결정**



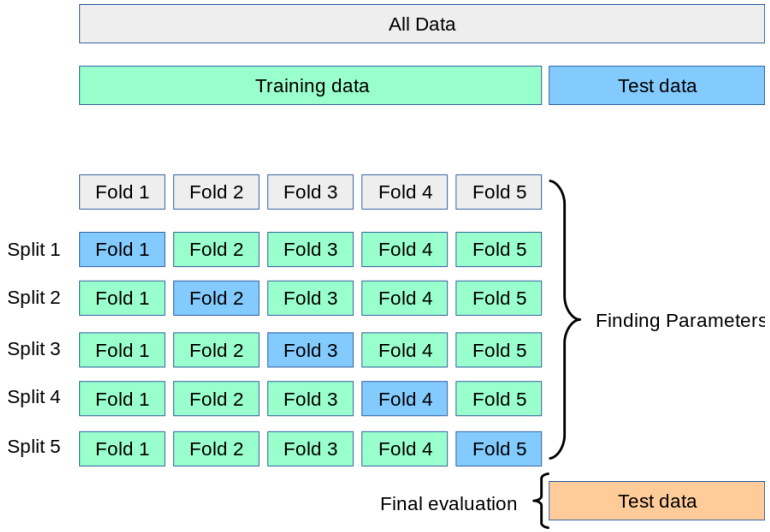


Experimental setup

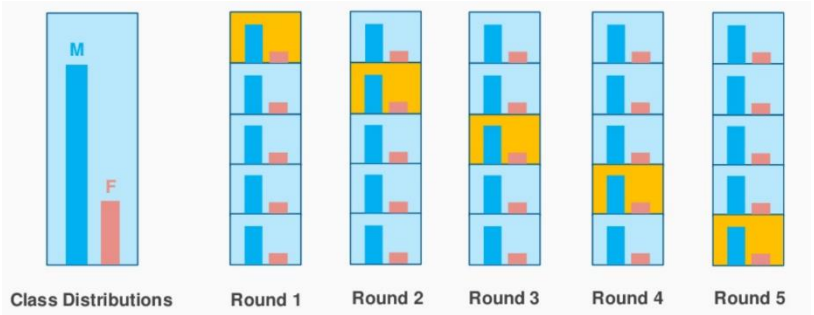
✓ Model validation

- K-fold cross validation을 이용해 전체 train data를 활용함과 동시에 robust한 모델 구축
- Tabnet 모델은 5-fold cross validation을 적용하고 soft-voting으로 결과 예측
- LGBM 모델은 5-fold stratified cross validation을 적용해 fold 마다 label 비율을 유지시켜 validation 신뢰성을 높임

K-fold cross validation



K-fold stratified cross validation



Experimental result

- ✓ **ESM embedding feature**와 **Tabnet**을 이용한 모델이 가장 높은 성능을 보임
- ✓ **CT-CTD feature**와 **CNT feature**와 **PCA/AE**를 적절히 사용하면 좋은 모델을 구성할 수 있을 것이라고 생각함
- ✓ **Tabnet**과 **LGBM**의 **ensemble**과 **threshold**를 조절하면 성능 향상을 기대해볼 수 있음

Result table

Features	Model	PCA(CT-CTD)	PCA(CNT)	K-folds	CV	Public LB	Private LB
ESM + Cat	**Tabnet**	**N/A**	**N/A**	**5**	**0.8560**	**0.7537**	**0.7603**
	LGBM	N/A	N/A	5 (stratified)	0.8171	0.7076	0.7094
ESM + CTCTD + Cat	Tabnet	0	N/A	5	0.8601	0.7356	0.7379
	LGBM	300	N/A	5 (stratified)	0.8155	0.7066	0.7090
ESM + CNT + Cat	Tabnet	N/A	0	5	0.8610	0.6686	0.6729
	LGBM	N/A	300	5 (stratified)	0.8200	0.7058	0.7087

Thank you

References

[1] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, *118*(15), e2016239118.

[2] Sharma, A., & Singh, B. (2020). AE-LGBM: Sequence-based novel approach to detect interacting protein pairs via ensemble of autoencoder and LightGBM. *Computers in Biology and Medicine*, *125*, 103964.

[3] Mu, Z., Yu, T., Liu, X., Zheng, H., Wei, L., & Liu, J. (2021). FEGS: a novel feature extraction model for protein sequences and its applications. *BMC bioinformatics*, *22*(1), 1-15.

[4] Arik, S. Ö., & Pfister, T. (2021, May). Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 8, pp. 6679-6687).