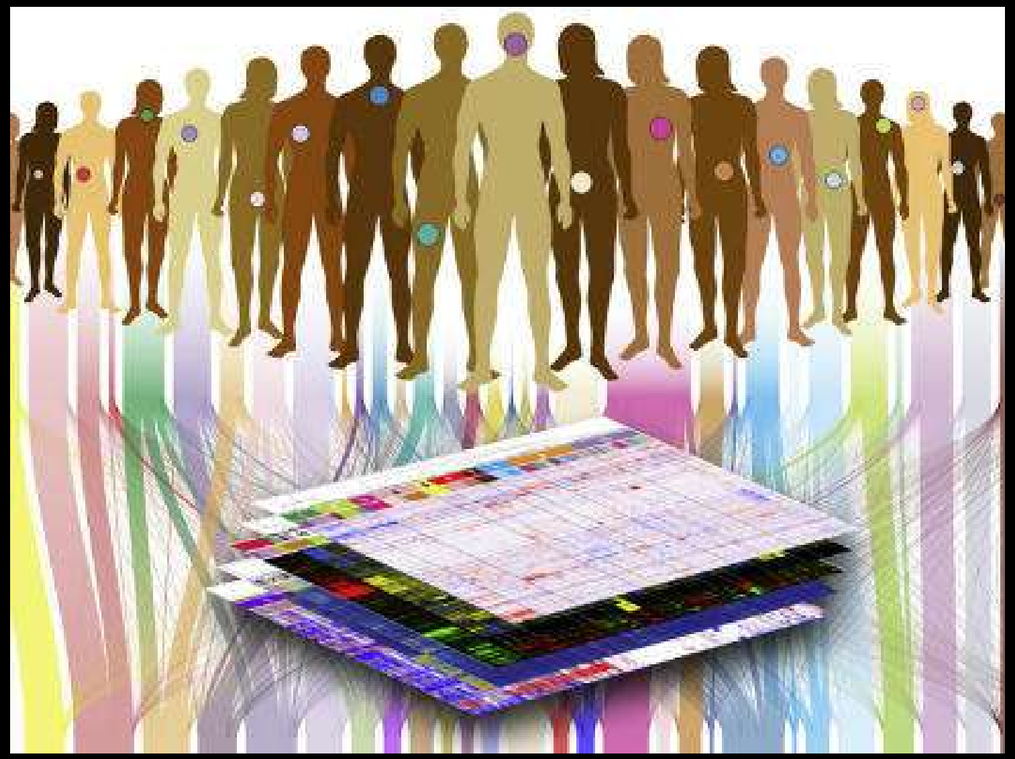
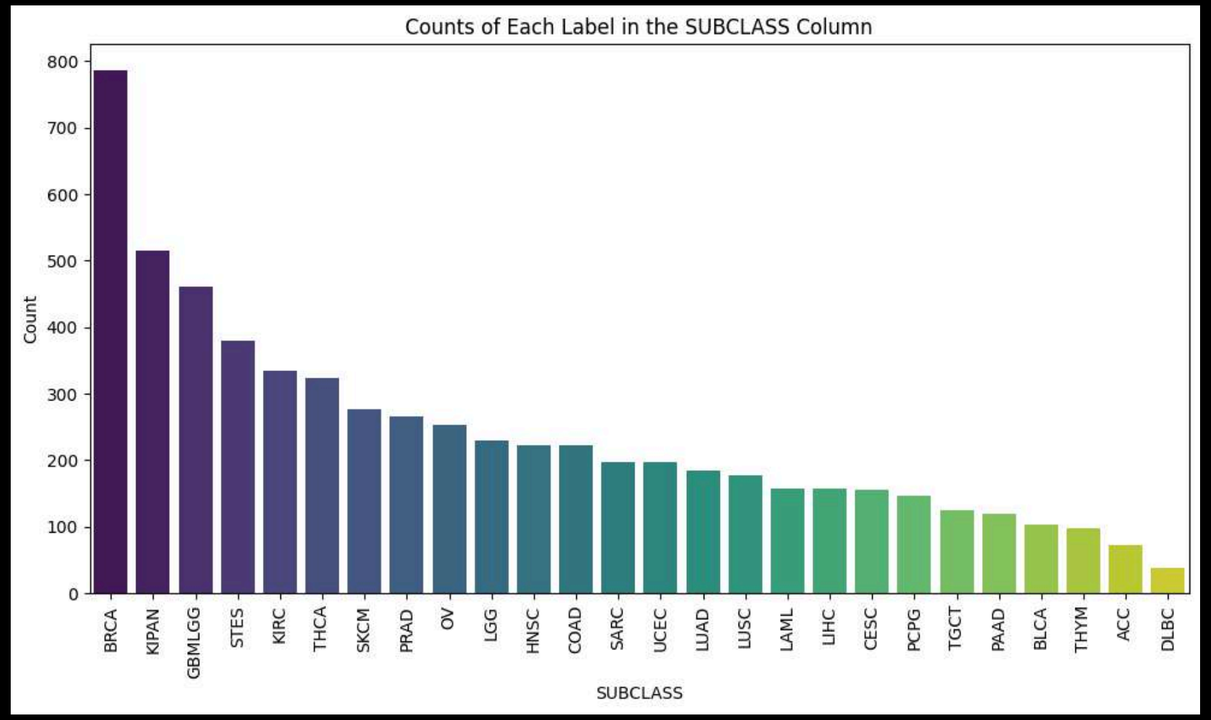
2024 생명연구자원 AI 활용 경진대회

Team GIGO

Datu, Stay, 상준.

PROBLEM DEFINITION

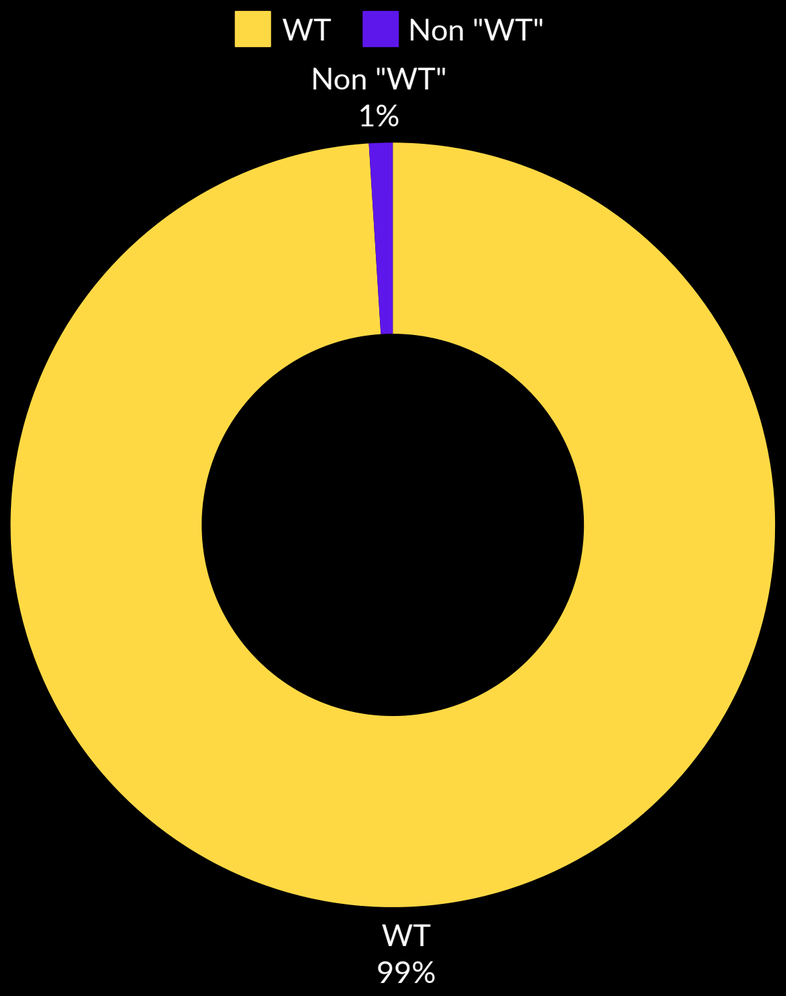




DOMAIN SPECIFIC TASK

CLASS IMBALANCE

DOMAIN SPECIFIC TASK



데이터 내에는 RNA 변이 정보는 없고,
오직 단백질 변이만 존재함

데이터 내 한 환자 샘플에 대해 WT(변이없음)에 대한 불필요한 정보로 인해
데이터 노이즈 발생 + 연산량 증가

자연어 기반 접근

변이가 있는 열만 사용하여 중요한 정보를 추출해내고
노이즈를 제거하여 필요한 정보만 학습

NATURAL LANGUAGE PROCESSING

동의성 변이, 비동의성 변이

{gene} -> {a}{pos}{a} or {a}{pos}{b}

In {gene}, {a} changes {a} {pos}
In {gene}, {a} changes {b} {pos}

종결 코돈 변이

{gene} -> {a}{pos}{*}

In {gene}, {a} {pos} changes to a stop codon.

프레임시프트 변이

{gene} -> {a}{pos}{b}fs

In {gene}, frameshift {a} {b} {pos}

삽입, 삭제 변이

{gene} -> {a}{pos}del

In {gene}, {pos} , {a} is del.

Sentence Example

In CR1, Arginine changes to Leucine on 2289; In KCNK5, Isoleucine changes to Threonine on 191; In LPCAT3, Valine changes to Valine on 157; In MMP8, Threonine changes to Lysine on 30; In PLAUR, frameshift on Tyrosine 171; In RFC4, Glutamine changes to Glutamine on 282; In TP53, Arginine changes to stop codon on 342.

EXCEPTION CASE

다중 변이 1

{gene} -> {a}{pos1}_{a}{pos2}del

In {gene}, {a} del in range
{pos1} to {pos2}

다중 변이 2

{gene} -> {a}{pos1}_{b}{pos2}>{c}{d}

In {gene}, {a} changes to {c} on {pos1}
and {b} changes to {d} on {pos2}

AMINO ACID DICT

'A': 'Alanine'
'R': 'Arginine'
'N': 'Asparagine'
'D': 'Aspartic acid'
'C': 'Cysteine'
'E': 'Glutamic acid'
'Q': 'Glutamine'
'G': 'Glycine'
'H': 'Histidine'
'I': 'Isoleucine'
'L': 'Leucine'
'K': 'Lysine'
'M': 'Methionine'
'F': 'Phenylalanine'
'P': 'Proline'
'S': 'Serine'
'T': 'Threonine'
'W': 'Tryptophan'
'Y': 'Tyrosine'
'V': 'Valine'

WHY TF-IDF?

중요도가 높은 변이에 대한 가중치

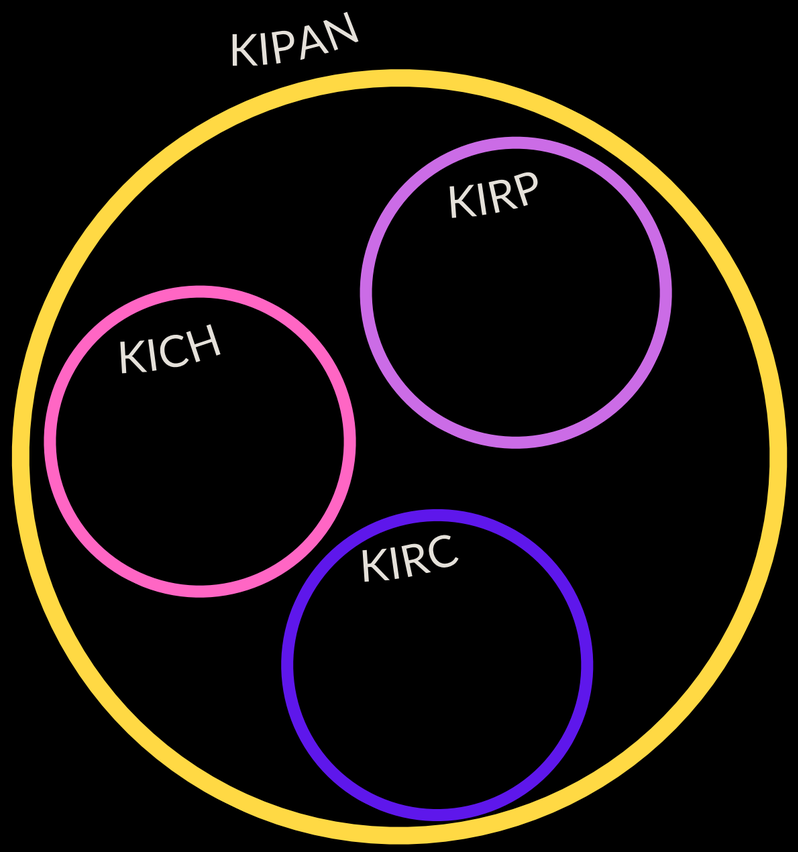
- TF-IDF는 자주 나온 단어에 대한 가중치가 적고, 적게 나온 단어에 대한 가중치가 큼.
- 모든 암종류 에서 많은 변이가 일어나는 여러 유전체들에 대한 가중치가 적고, 특정 암종류에만 일어나는 특정 변이에 대한 가중치가 크므로 TF-IDF vectorization 방법을 사용.
- 이는 딥러닝 모델, 특히 Transformer 계열의 모델의 input data로 활용하기에 적절하다고 판단.
- 라벨 인코딩을 사용 시, 유전학적인 정보 마스킹으로 인해 손실이 발생.

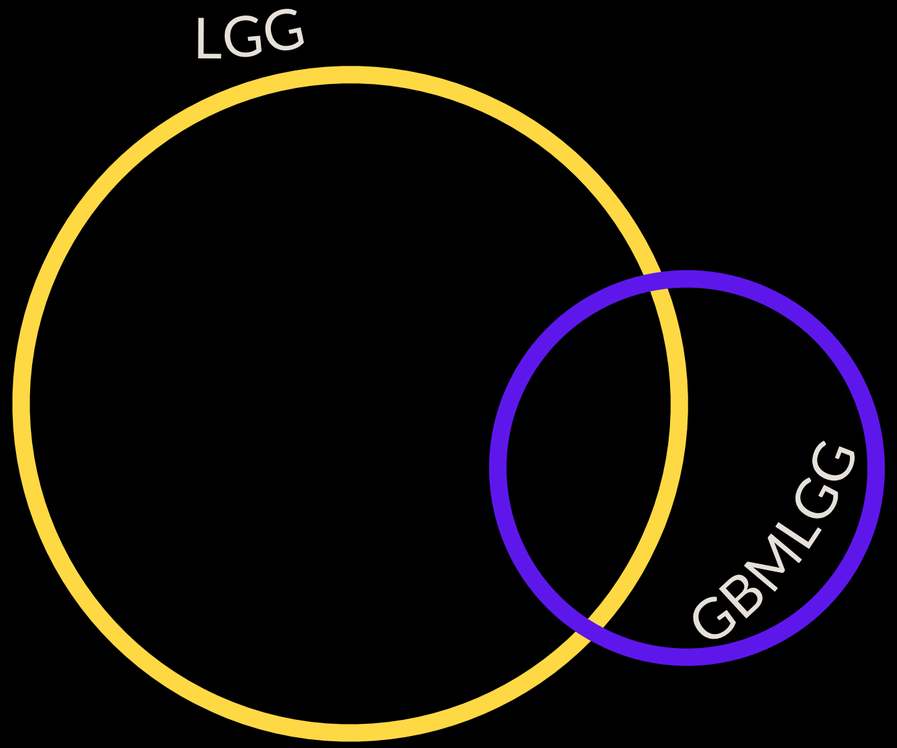
In TP53, Arginine changes to stop codon on 342.

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t,D) = log\frac{N}{|\{d \in D : t \in d\}|}$$

CLASS IMBALACE



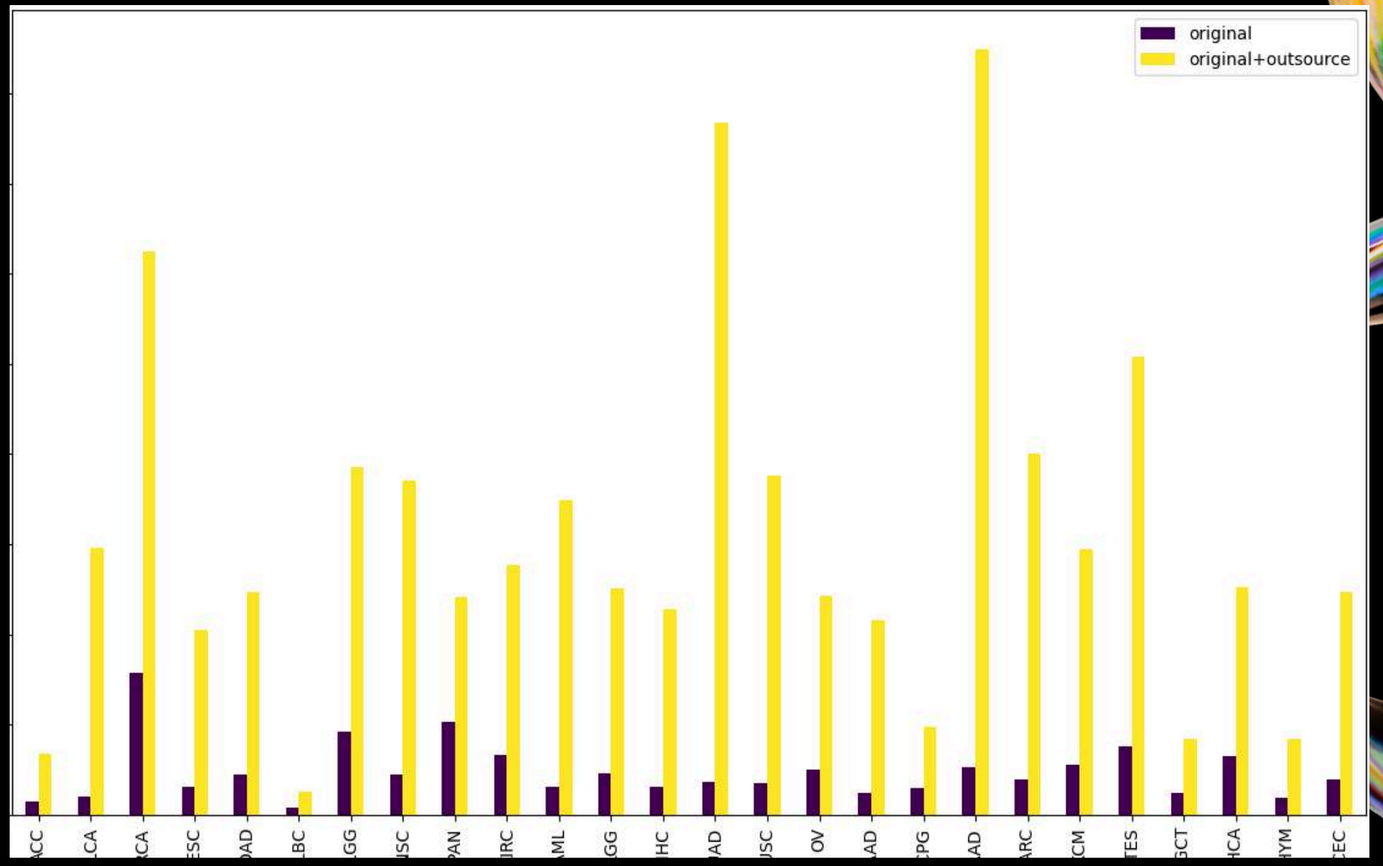


KIPAN은 KICH, KIRP, KIRC 세개의 신장암 종류로 구성.
허나 이번 데이터셋은 KIRC 와 KIPAN 암종을 다르게 두었을 뿐 아니라
LGG 와 GBMLGG 도 같은 뇌종양 이지만 차등 암종일 뿐이라 모델이 예측하는데 혼란을 야기함

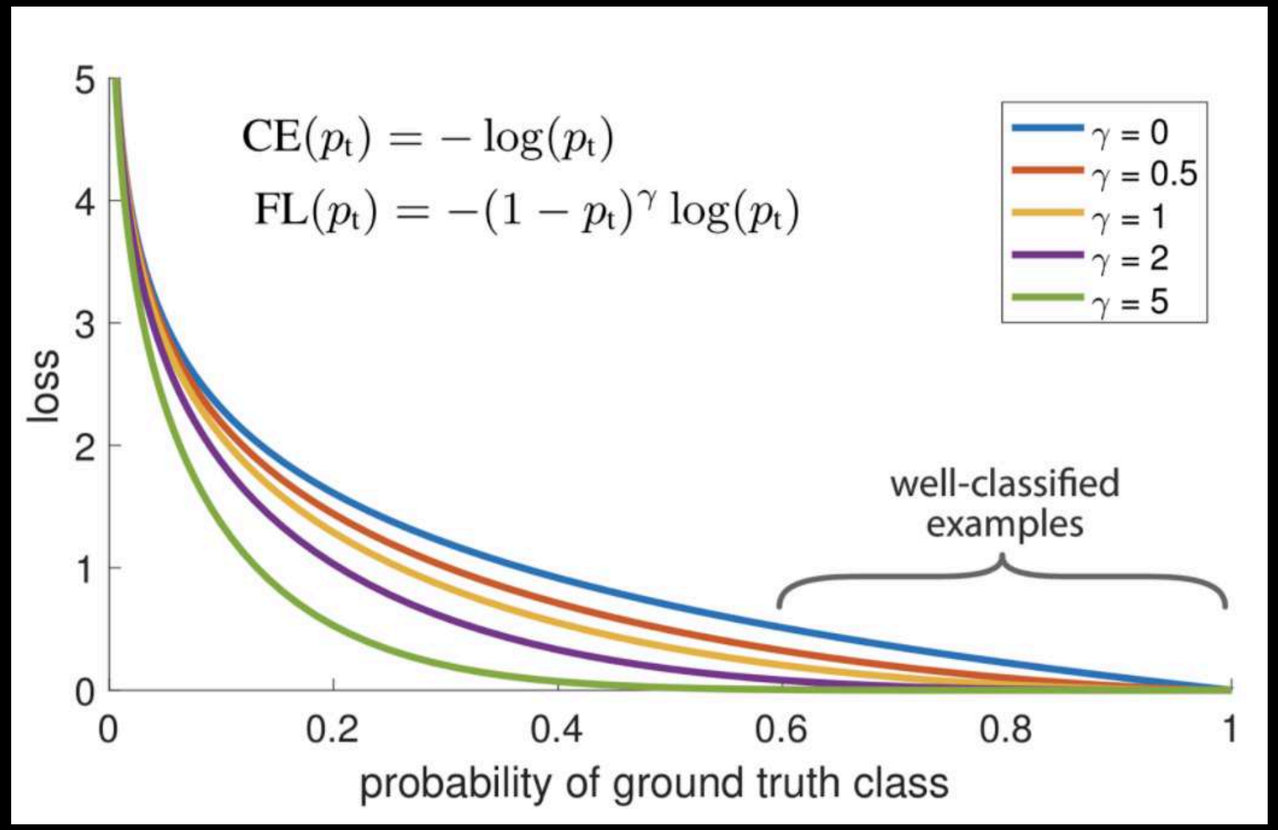
외부데이터 수집

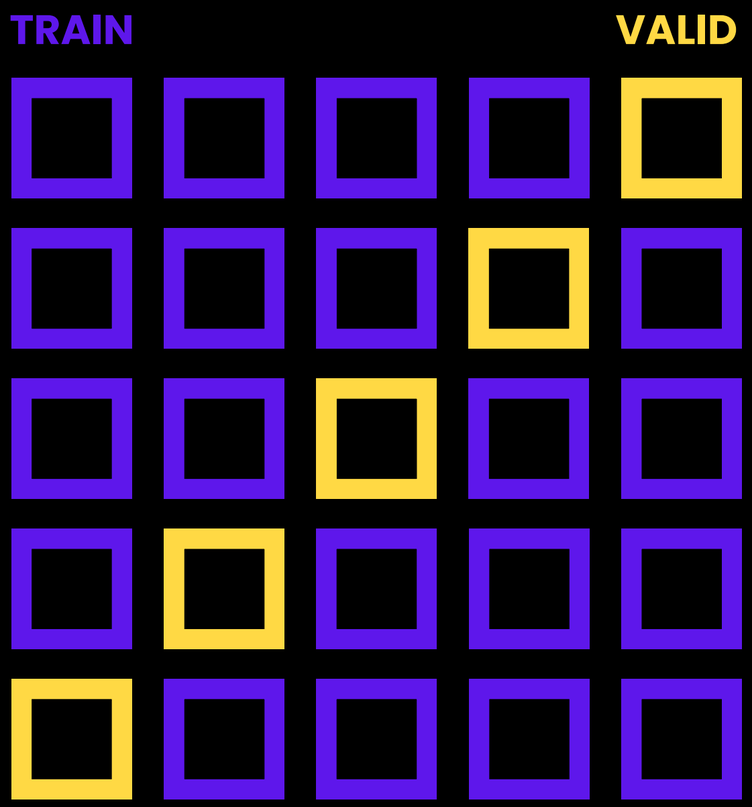
각 CLASS별 외부 데이터 수집

	Tumor_Sample_Barcode	Hugo_Symbol	Consequence	HGVSp_Short
0	TCGA-OR-A5L9-01A-11D-A29I-10	FBLN2	missense_variant	p.V368A
1	TCGA-OR-A5L9-01A-11D-A29I-10	WWP1	synonymous_variant	p.G363G
2	TCGA-OR-A5L9-01A-11D-A29I-10	NDRG2	missense_variant	p.A147P
3	TCGA-OR-A5L9-01A-11D-A29I-10	HDAC6	missense_variant	p.H651Q
4	TCGA-OR-A5L9-01A-11D-A29I-10	G6PD	missense_variant	p.V340I



VALIDATION STRATEGY





FOCAL LOSS

STRATIFIED KFOLD

MODEL PIPELINE

LightGBM

Boosting ML

Transformer Based Model

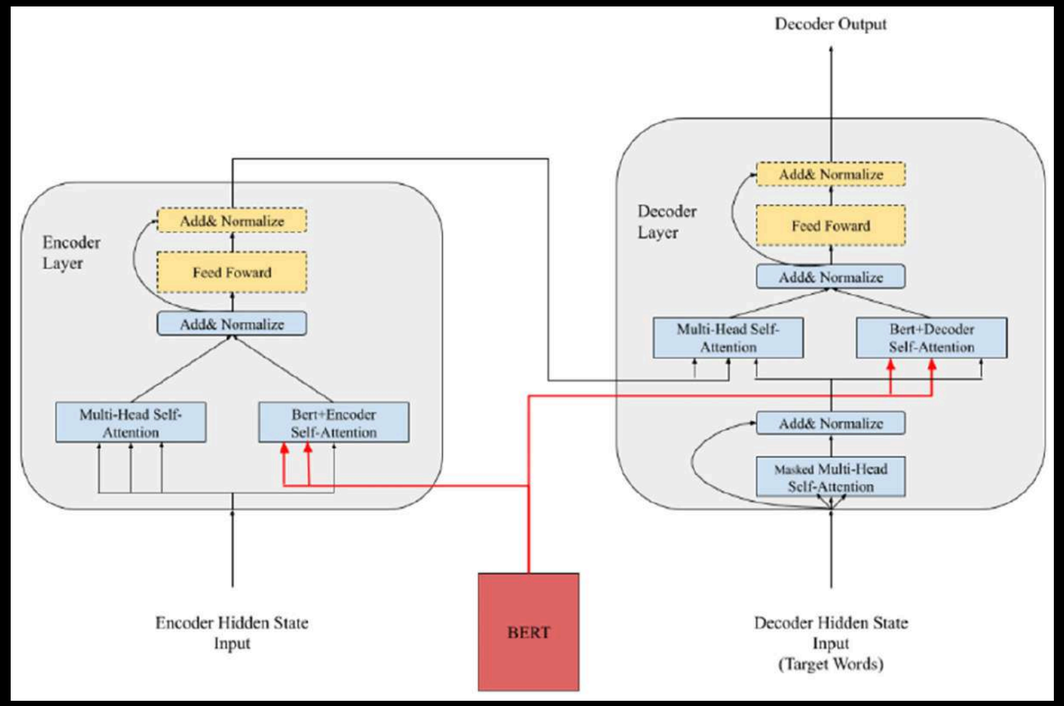
XTransformer

SAINT

Transformer Ensemble

Final Prediction

TRANSFORMER



XTransformer

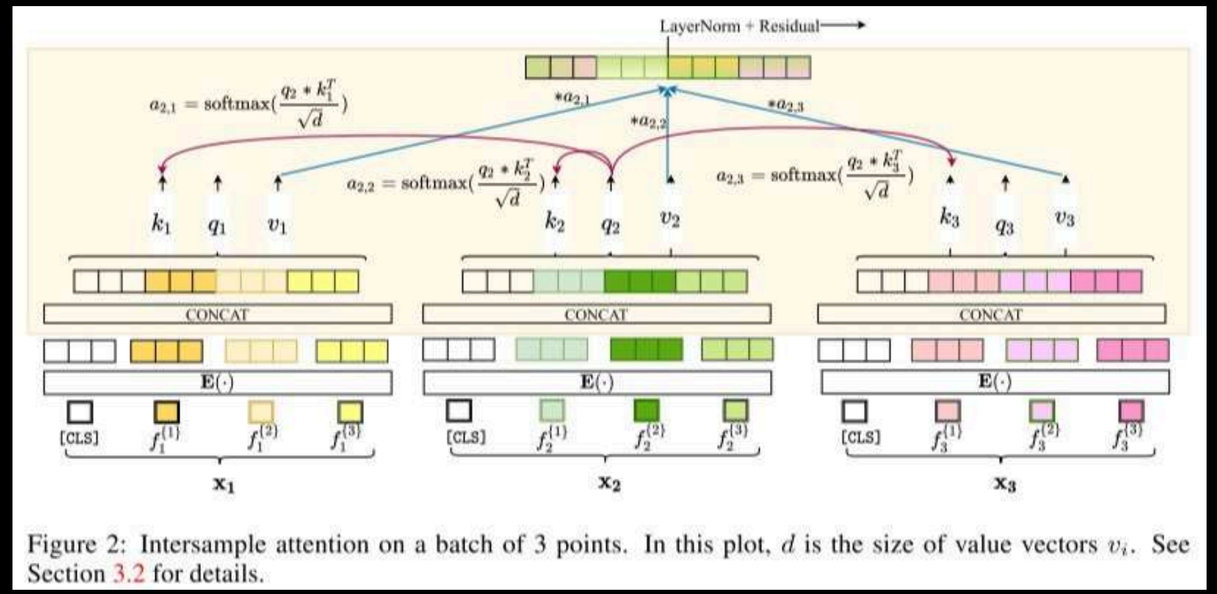


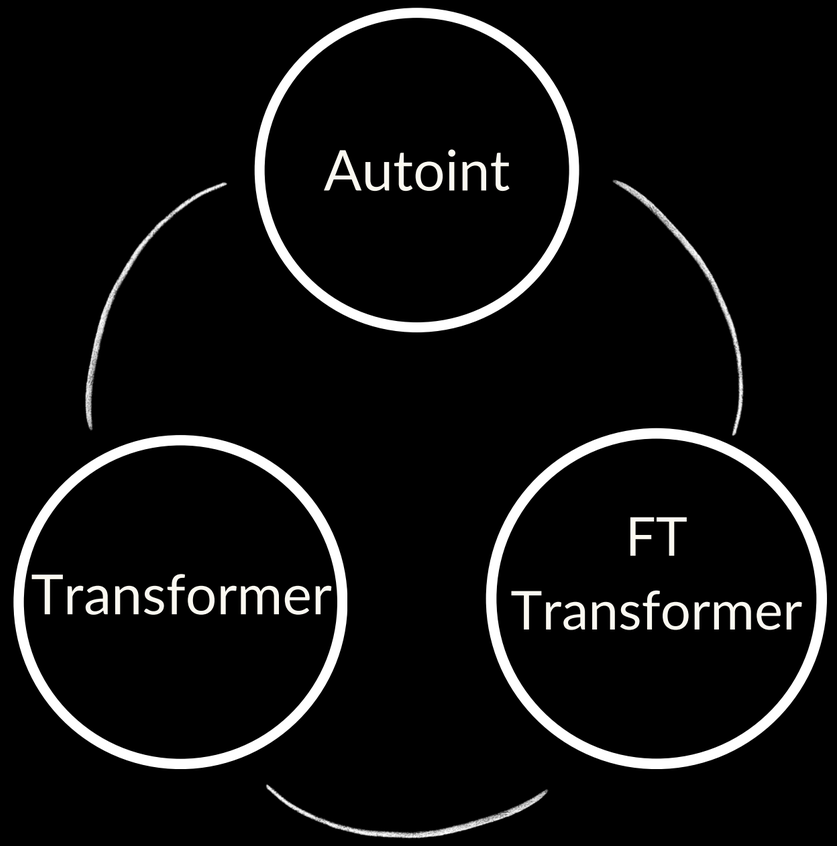
Figure 2: Intersample attention on a batch of 3 points. In this plot, d is the size of value vectors v_i. See Section 3.2 for details.

SAINT

MODEL PERFORMACE

	LightGBM	XTransformer	Saint
Single	0.61592	0.66349	0.65123
Ensemble Strategy 1	0.66578		-
Ensemble Strategy 2	-	0.65213	
Final Ensemble	0.67573		

Trial Model



추가 실험

TRIAL 1

특정 단백질 변이는 특정 코돈 변이로 이루어짐
Ex) (N->AAT,AAC) , (H->CAT,CAC)

위 배경지식을 활용하여 train 의 단백질 변이로부터 코돈 변이 진행 및 외부데이터의 코돈 변이를 활용하여 진행해 보았지만 성능 소폭 하락

TRIAL 2

외부데이터의 polyphen , sift score(변이의 위험정도) 를 사용

-> 비동의성 변이에만 polyphen score가 등재되어 있어서 학습시에 과적합이나 클래스가 줄어드는 현상을 발견하여 사용하지 못함

TRIAL 3

단백질 변이를 자연어처리로 학습하기 위해
과학 논문들의 내용으로 학습된 BERT 계열의
PUBMEDBERT 나 SCIBERT 를 사용

-> 문장의 길이가 너무 길어져 BERT 의 한계점 도달

TRIAL 4

대/중분류([KIPAN, KIRC], [LGG, GBMLGG]) 등에
대해서 따로 분류하여 구분하게 하거나
contrastive learing을 사용

-> 성능 소폭 하락

범용성

정밀한 병리 현상 분석과 문제 해결

암종 분류 작업에서 단백질 변이뿐만 아니라
코돈 변이 정보를 포함하여 모델링하면 더 세부적인 학습이 가능.
이를 통해 보다 정확하고 정밀한 암종 분류 모델을 구축.

질병 진단 및 예측

단백질 변이와 코돈 변이 분석은 특정 암에만 국한되지 않고,
다양한 유전 질환 및 희귀 질환의 유전적 변이 분석에도 활용.
특히 자연어 처리 기술을 활용하면 더욱 효과적인 질병 진단 모델 구축

MULTIMODAL

단백질 변이 정보를 활용해 암종을 분류하는 것처럼,
의료 영상이나 생물학적 이미지 데이터를 분석하는 모델로 확장 가능.
병리학적 이미지, CT/MRI 영상 분석을 통해 다양한 의료 영상에서
병변을 자동으로 감지하는 데 적용.

대형 언어 모델(LLM)을 통한 확장성

질병별로 개별 모델을 구축하다 보면
데이터의 부족으로 어려움을 겪을 수 있음.
이를 해결하기 위해 **각 질병에 대한 지속적인 학습(continual learning)을 통해 모델의 성능을 향상**.

THANK YOU

for your time and attention