

# '디지털 보물찾기' 팀의 '문서 타입 분류' 경연 도전기

'디지털 보물찾기(Digital Treasure Quest)' 팀이 '문서 타입 분류' 경연에 도전한 여정을 소개합니다. 이 대회는 주어진 문서 이미지를 여러 클래스 중 하나로 분류하는 모델을 개발하는 것이 목표였습니다. 금융, 의료, 보험, 물류 등 다양한 산업에서 대량의 문서를 효율적으로 처리하고 자동화하는 데 중요한 역할을 하는 기술을 개발하는 과정에서 팀원들의 노력과 창의성이 빛을 발했습니다. 이 도전을 통해 팀의 전략, 사용한 기술, 그리고 얻은 교훈을 공유하고자 합니다.

**Team #5 :: 디지털 보물찾기 (Digital Treasure Quest)**

박석, 백경탁, 한아름, 위효연

# 대회 개요

## 목표

문서 타입 분류를 위한 이미지 분류 모델을 개발합니다.

## 제공 데이터셋

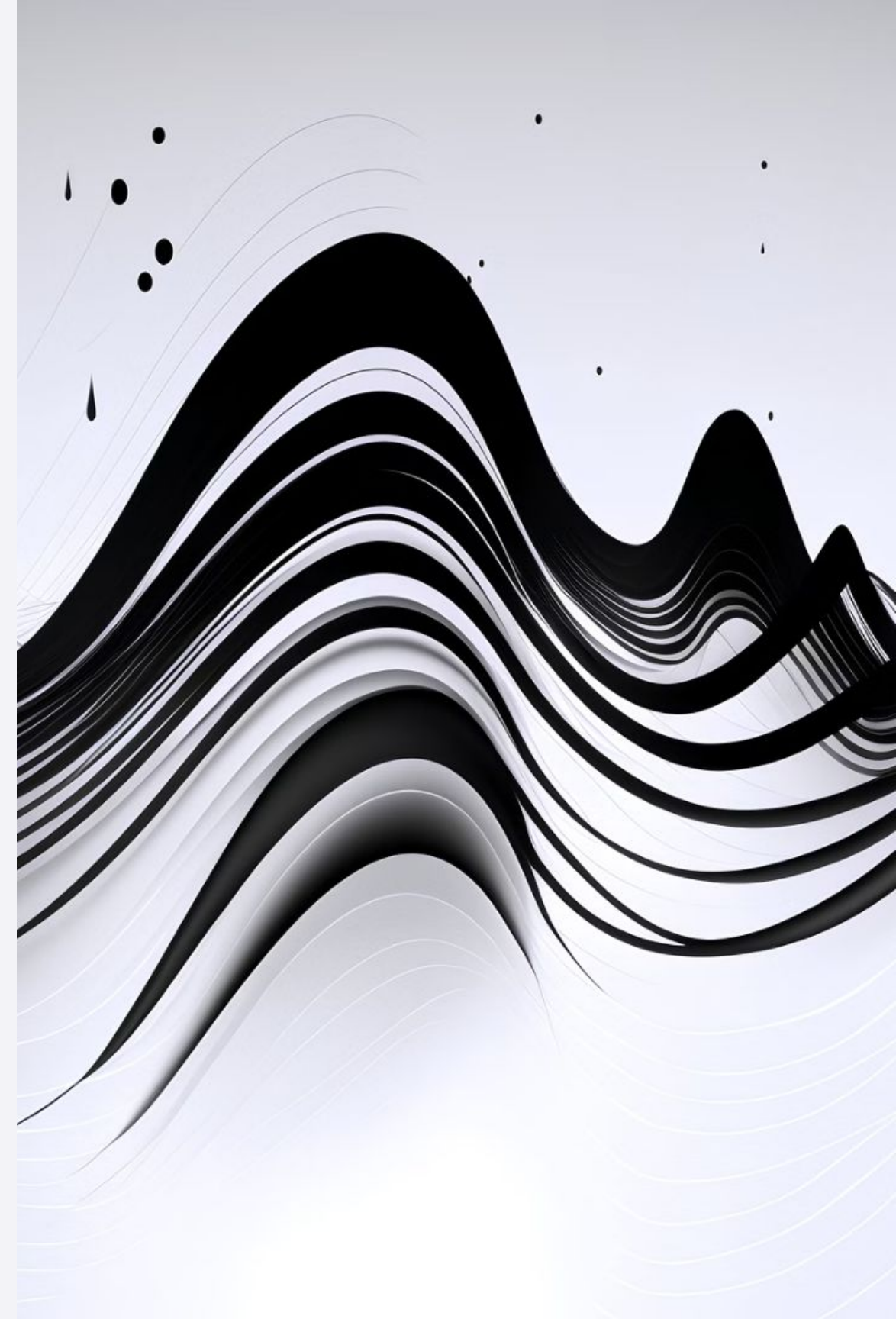
17개 클래스에 속하는 1,570장의 학습 이미지 데이터와 3,140장의 평가 이미지 데이터가 제공되었습니다.

## 사용 가능한

**알고리즘**  
다양한 딥러닝, 전이 학습, 컨볼루션 신경망(CNN) 등 다양한 이미지 분류 알고리즘을 활용할 수 있습니다.

## 평가 지표

모델의 성능은 Macro F1 score를 사용하여 평가됩니다.







# 대회 일정

1

## 대회 시작

2024년 7월 30일 (화) 10:00에 대회가 시작되어 데이터셋이 배포되었습니다.

2

## 팀 병합 마감

2024년 7월 31일 (수) 10:00까지 팀 병합이 가능했습니다.

3

## 개발 및 테스트 기간

2024년 7월 30일부터 8월 10일까지 모델 개발과 테스트가 진행되었습니다.

4

## 최종 모델 제출

2024년 8월 11일 (일) 19:00에 최종 모델 제출이 마감되었습니다.

# 평가 방법

1

## Macro F1 Score

모델의 성능은 Macro F1 Score를 사용하여 평가됩니다. 이는 모든 클래스에 대해 개별적으로 계산된 F1 Score의 단순 평균을 의미합니다.

2

## 클래스 불균형 대응

Macro F1 Score는 클래스 불균형이 존재하는 상황에서 모델의 성능을 더욱 정확하게 평가할 수 있는 지표입니다.

3

## Public 및 Private

평가용 Test 데이터 중 랜덤 샘플링 된 50%를 사용하여 Public 리더보드 점수가 산출되며, 나머지 50%로 최종 Private 리더보드 점수가 산출됩니다.



# DTQ 팀의 장단점

## 장점

- 다양한 경력과 경험을 가진 팀원들
- 평균 나이가 높음
- AI Assistant에 대한 수용력이 높음

## 단점

- Git을 활용한 팀단위의 R&D 경험 수준 낮음
- Python기반 R&D 경험 수준 낮음
- 머신러닝/딥러닝 R&D 경험 수준 낮음
- 경연 주제와 관련된 도메인 지식이 낮음
- Career Path에 대한 개인적인 목표가 모두 다름





# DTQ 팀의 전략적

## 적근

1

### AutoML 도구 활용

DataRobot과 같은 AutoML 도구를 적극 활용하여 Feature Engineering과 Model Selection의 방향성을 잡습니다.

2

### 개인별 모델링

팀원별로 서로 다른 머신러닝 모델링을 각 팀원의 수준에 맞게 진행합니다.

3

### AI Assistant 활용

AI Assistant를 적극적으로 활용하여 개인별 생산성을 극대화합니다.



# DTQ 팀의 문화와

## 정신

1

### 학습 중심

경연 참가의 목적은 개인별 학습을 통해 머신러닝 R&D에 필요한 지식과 경험을 얻는 것에 있습니다.

2

### 상호 존중

팀원 각각이 처한 상황을 서로 이해하고 인정하고 Respect 합니다.

3

### 균형 잡힌 접근

팀 전체 목표를 위해 팀원 개개인의 스케줄이나 리소스를 희생해서는 안 됩니다.

4

### 참여 독려

팀원별로 최소한 한번의 제출은 해 봅니다.

# 프로젝트 디렉토리

## 구조

1

### code 디렉토리

팀원별 실험 소스 코드 및 관련 문서를 포함합니다. tm1, tm2, tm3, tm4 서브디렉토리와 team 디렉토리로 구성됩니다.

2

### docs 디렉토리

팀 문서(발표자료, 참고자료 등)를 포함하며, presentation과 reference 서브디렉토리로 구성됩니다.

3

### images 디렉토리

프로젝트에 사용된 첨부 이미지들을 저장합니다.

4

### README.md

'디지털 보물찾기(Digital Treasure Quest)' 팀의 '문서 타입 분류' 경연 도전기에 대한 설명을 포함합니다.







# 데이터셋 개요

구분	학습 데이터	평가 데이터
이미지 수	1,570	3,140
파일명	train.csv	test.csv
컬럼	ID, target	ID, target
클래스 수	17	(예측용) -



# 클래스 분포

1

## 클래스 불균형

제공된 히스토그램을 통해 클래스 간 불균형이 존재함을 확인할 수 있습니다. 일부 클래스는 상대적으로 많은 샘플을 가지고 있는 반면, 다른 클래스는 적은 샘플을 가지고 있습니다.

2

## 모델링 고려사항

이러한 클래스 불균형은 모델 학습 시 고려해야 할 중요한 요소입니다. 클래스 가중치 조정이나 오버샘플링 등의 기법을 통해 불균형을 해소할 필요가 있습니다.

3

## 평가 지표 선택

클래스 불균형 상황에서 Macro F1 Score를 평가 지표로 사용하는 것은 적절한 선택으로 보입니다. 이를 통해 모든 클래스에 대해 균형 잡힌 성능 평가가 가능합니다.





# 학습 이미지

## 미리보기

1

### 이미지 품질

학습용 이미지들은 비교적 잘 정돈되어 있고 상태도 양호합니다. 문서의 내용이 뚜렷하게 보이며, 노이즈나 왜곡이 거의 없는 것으로 보입니다.

2

### 문서 다양성

다양한 유형의 문서가 포함되어 있음을 확인할 수 있습니다. 이는 모델이 다양한 문서 유형을 학습할 수 있게 해줍니다.

3

### 이미지 방향

대부분의 이미지가 올바르게 정렬되어 있어, 방향이 표준화된 것처럼 보입니다. 이는 모델 학습에 도움이 될 수 있습니다.





# 평가 이미지

## 미리보기

1

### 이미지 변형

평가 이미지들은 학습 이미지와 달리 다양한 각도에서 촬영되었으며, 회전, 기울기, 손상 또는 열화된 부분이 눈에 띕니다.

2

### 노이즈와 배경

평가 이미지에는 배경 노이즈와 변동이 더 많이 나타나며, 배경 텍스처, 색상, 또는 다른 물체들이 포함될 수 있습니다.

3

### 모델링 고려사항

이러한 차이점은 모델이 학습 데이터와 평가 데이터 간의 격차를 극복할 수 있도록 데이터 증강 및 전처리 전략을 세워야 함을 시사합니다.



# 데이터 증강 기법

1

## 회전 및 뒤집기

이미지를 90°/180°/270°/360° 중 랜덤하게 회전시키거나, 수평 또는 수직 방향으로 랜덤하게 뒤집습니다.

2

## 노이즈 및 블러

가우시안 노이즈를 추가하거나, 모션, Median, 일반 블러 중 선택하여 적용합니다.

3

## 왜곡 및 변형

이미지에 광학, 그리드, 어파인 변환 중 선택하여 왜곡을 적용합니다. 또한, 회전 후 남은 외곽을 흰색으로 채웁니다.



# 데이터 클렌징

1

## 라벨링 오류 수정

학습 데이터에서 잘못 레이블되어 있는 데이터를 찾아 수정했습니다. 그러나 오분류에 대한 기준이 명확하지 않아 일부만 수정되었습니다.

2

## 데이터 오염 사례

예를 들어, '납입확인서'로 레이블된 이미지에 "진료비 납입 영수증"이라고 적혀있는 경우가 발견되었습니다.

3

## 주최측의 의도

일부 데이터 오염은 대회 주최측의 의도된 것으로 판단되어 추가적인 수정을 하지 않았습니다.



# 모델 선택 전략

1

## AutoML 활용

DataRobot을 활용하여 Model Selection을 수행했습니다.  
이를 통해 Leaderboard의 상위권에 속한 모델들을  
식별했습니다.

2

## 선택된 모델

1. Regularized Logistic Regression L2
2. Keras Slim Residual Neural Network Classifier
3. Baseline Image Classifier

3

## 선택 기준

각 모델의 성능, 해석 가능성, 그리고 팀의 기술적 역량을  
고려하여 최종 모델을 선택했습니다.



# Regularized Logistic Regression L2

1

## 모델 구성

Train-Time Image Augmentation, Pretrained MobileNetV3-Small-Pruned Multi-Level Global Average Pooling Image Featurizer, 그리고 Regularized Logistic Regression (L2)로 구성됩니다.

2

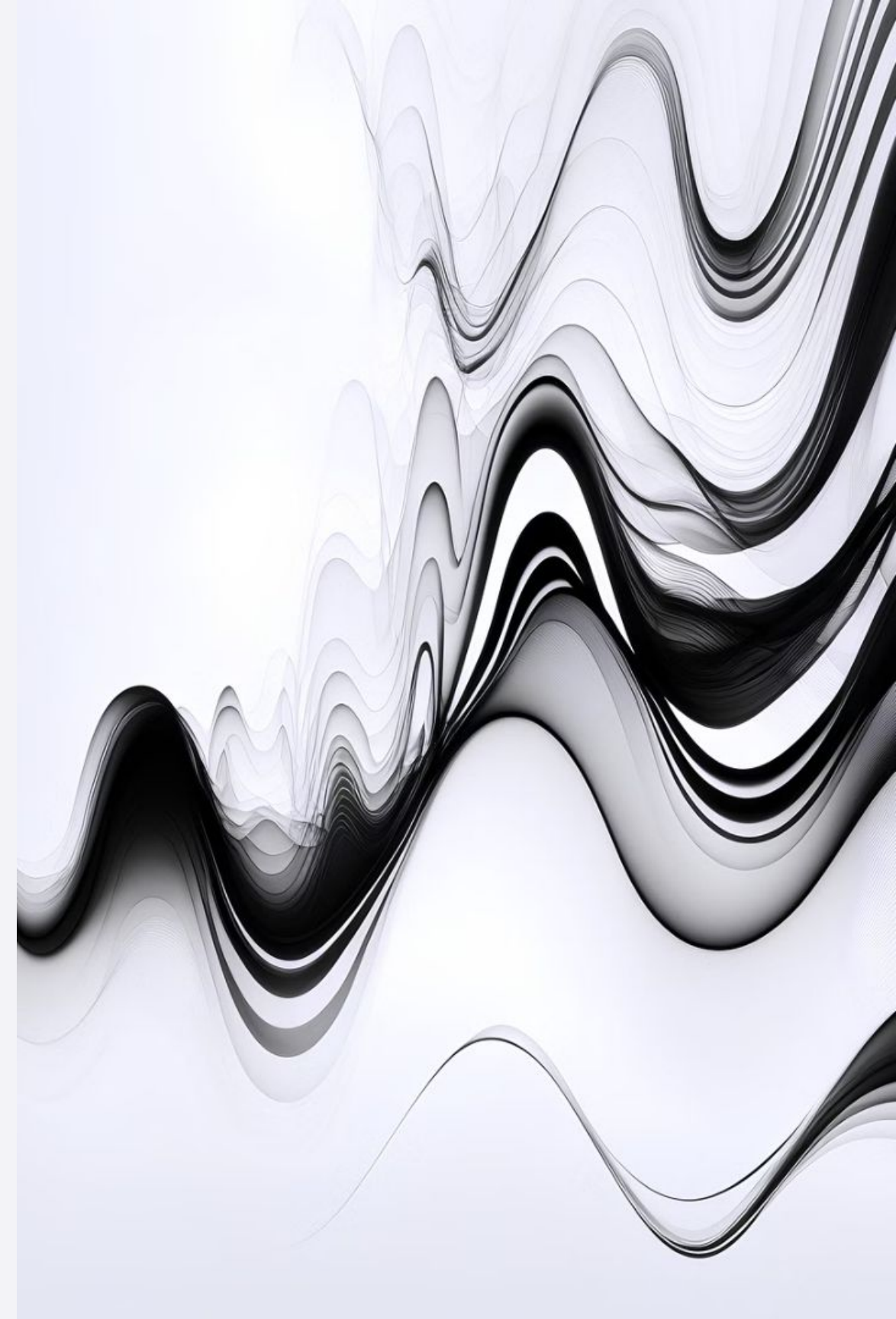
## 특징

L2 정규화를 통해 과적합을 방지하고, 사전 훈련된 CNN을 사용하여 이미지 특징을 추출합니다.

3

## 장점

해석 가능성이 높고, 잘 보정된 분류 확률을 생성하는 경향이 있습니다.





# Keras Slim Residual Neural Network Classifier

1

## 모델 구성

Train-Time Image  
Augmentation, Pretrained  
MobileNetV3-Small-Pruned  
Multi-Level Global Average  
Pooling Image Featurizer,  
그리고 Keras Slim Residual  
Neural Network Classifier를  
포함합니다.

2

## 특징

Residual 연결을 통해 깊은  
네트워크의 학습을 용이하게  
하고, 기울기 소실 문제를  
해결합니다.

3

## 장점

복잡한 패턴을 학습할 수 있는  
능력이 뛰어나며, 전이 학습을  
통해 적은 데이터로도 좋은  
성능을 낼 수 있습니다.





# Baseline Image Classifier

1

## 모델 구성

Train-Time Image Augmentation, Grayscale Downscaled Image Featurizer, 그리고 Regularized Logistic Regression (L2)으로 구성됩니다.

2

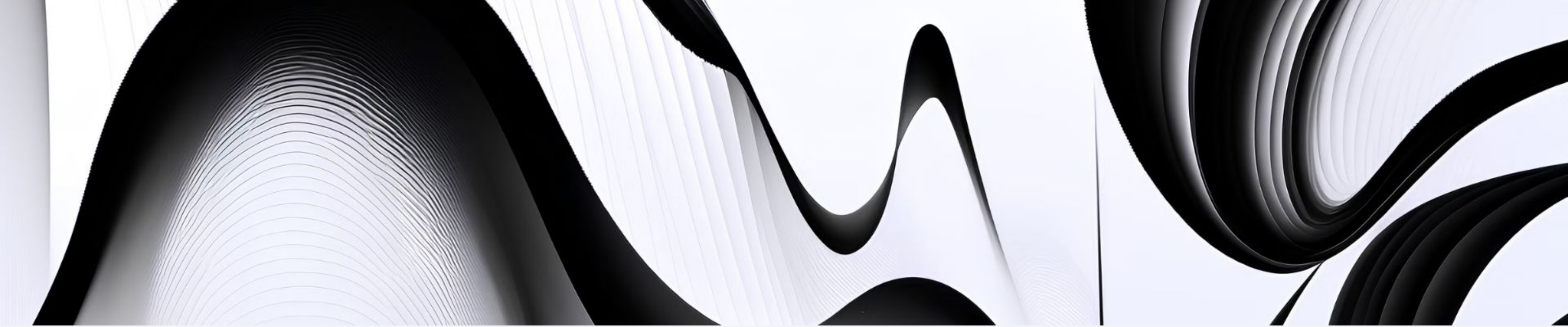
## 특징

이미지를 그레이스케일로 변환하고 크기를 줄여 간단한 특징 벡터로 만듭니다.

3

## 용도

주로 베이스라인 모델로 사용되며, 데이터의 기본적인 패턴을 파악하는 데 도움이 됩니다.



# 단위 모델 실험

## (백경탁)

1

사용 모델

ResNet34, ResNet50,  
EfficientNet (B4와 B5 버전)을  
주로 사용했습니다.

2

실험 방법


학습률, 배치 크기, 에포크,  
조기 중단, 가중치 감쇠 등의  
하이퍼파라미터를 변경해가며  
실험을 진행했습니다.

3

주요 결과

EfficientNet B4/B5 모델이  
뛰어난 성능을 보였으며,  
적절한 하이퍼파라미터  
설정과 고급 기법의 도입이  
성능 향상에 크게  
기여했습니다.





# 앙상블 모델 실험

## (백경탁)

1

### 앙상블 방법

여러 단위 모델의 예측 결과를 종합하여 최종 예측을 생성했습니다.

2

### 성능 향상

앙상블 기법을 통해 최종적으로 0.9118의 F1 점수를 기록하는 등, 단일 모델보다 높은 성능을 달성했습니다.

3

### 장점

앙상블은 개별 모델의 약점을 보완하고, 예측의 안정성을 높이는 데 효과적이었습니다.





# 단위 및 앙상블 모델 실험 (한아름)

다양한 모델 테스트 WanDB 모니터링 활용

1

## 1차 실험

Augmentation 효과 확인

2

## 2차 실험

과적합 방지 기법 적용, 데이터 수 변화

3

## 3차 실험

Voting 방식 시도, 클래스별 성능 분석

# 단위 및 앙상블 모델 실험 (위효연)

1

## 오분류 분석

잘못 분류된 이미지에 대한 상세한 분석을 수행했습니다. 이를 통해 모델의 약점과 개선 포인트를 파악했습니다.

2

## 앙상블 전략

여러 모델의 예측 결과를 종합하여 최종 예측을 생성하는 앙상블 기법을 적용했습니다.

3

## 성능 개선

앙상블 기법을 통해 개별 모델의 한계를 극복하고 전체적인 예측 성능을 향상시켰습니다.



# 모델 성능 테스트

## 비교

각 모델의 성능을 Macro F1 으로 비교했습니다. ???(위효연) 이 가장 우수한 성능을 보였습니다.

모델

Macro F1

??? (백경탁)

???

??? (위효연)

???

??? (한아름)

???

??? (박석)

???





# 최종 리더보드 결과

최종 순위

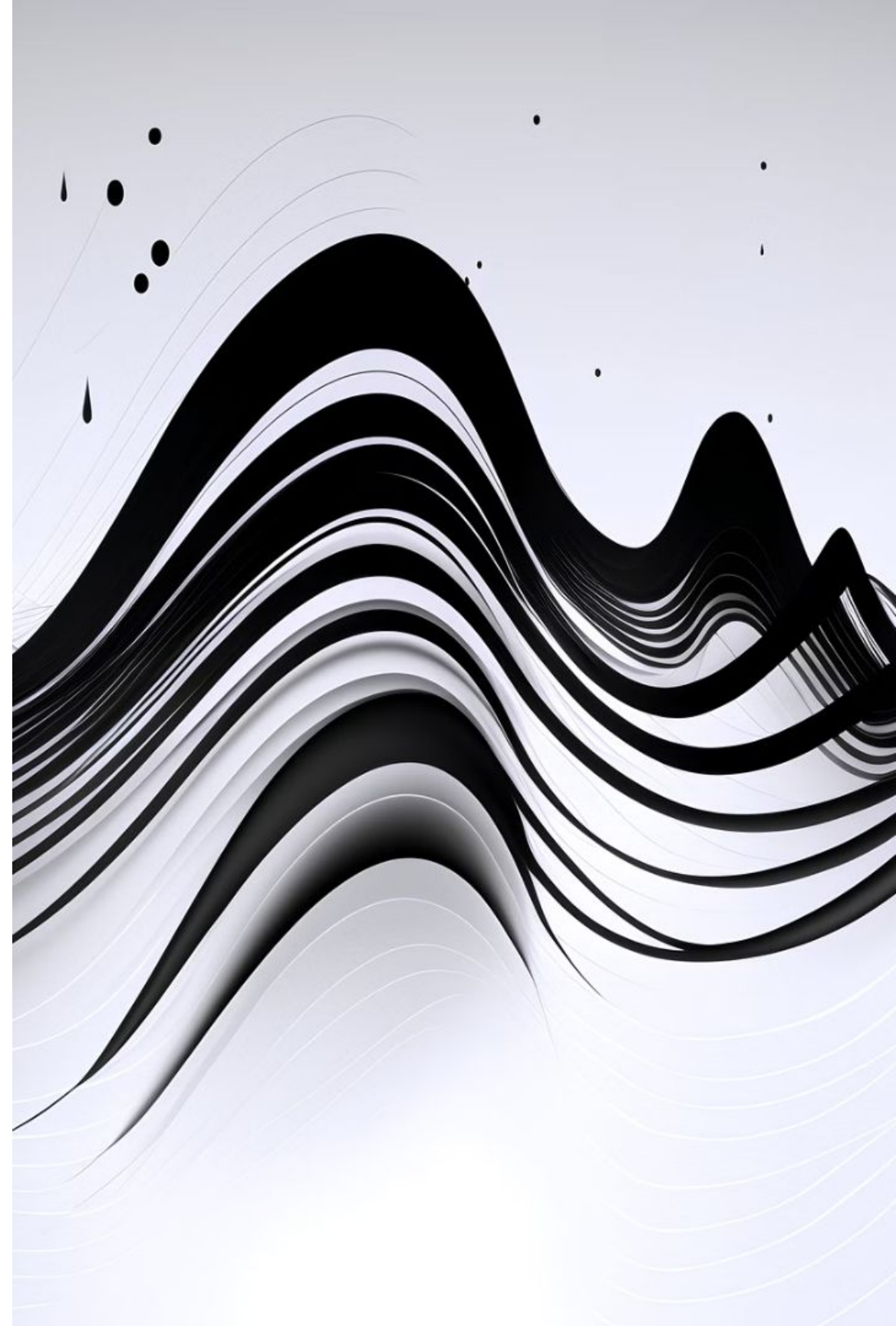
10위

최고 성능 모델

???

최종 Macro F1

???



# 문제 해결 방향

1

## 데이터 변형 대응

Rotate, Crop, Flip 등의 기법을 기본적으로 적용하고,  
테스트 데이터와 유사한 노이즈 패턴을 찾아 적용했습니다.

2

## 모델 선택

변형된 데이터를 잘 학습할 수 있는 BackBone Model을  
찾고, 다양한 기법을 통해 모델의 성능을 높였습니다.

3

## 데이터 증강 효과

초기 2.5만 개의 학습 데이터로 0.6206에서 0.8692로  
성능이 향상되었고, 최종 5만 개 학습 후 0.9340으로 더욱  
개선되었습니다.



# 추가 팁

1

## 테스트 데이터 EDA 중요성

이번 대회에서는 테스트 데이터에 대한 EDA(탐색적 데이터 분석)를 많이 신경 써야 합니다.

2

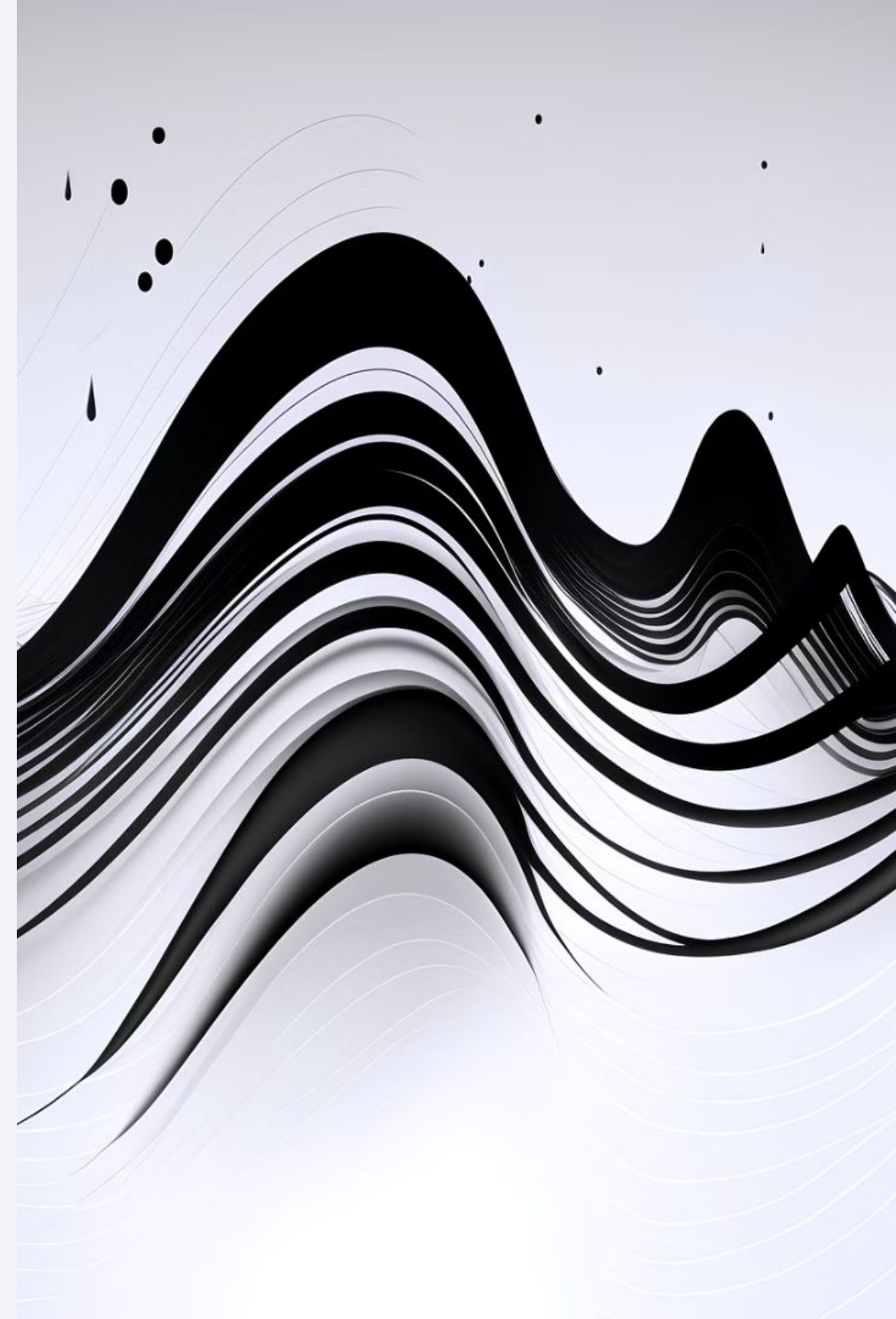
## Augmentation 분석

어떤 Augmentation을 적용했는지, 또한 이것의 특징이 카테고리별로 잘 묶이는지 EDA와 탐지를 해볼 필요가 있습니다.

3

## 전략적 접근

테스트 데이터의 특성을 잘 파악하고, 그에 맞는 데이터 증강 및 모델링 전략을 수립하는 것이 중요합니다.





# 대회 참가 소감

팀원들의 성장과 협업 경험 획득

1

## 학습 효과

머신러닝 R&D 지식과 경험 획득

2

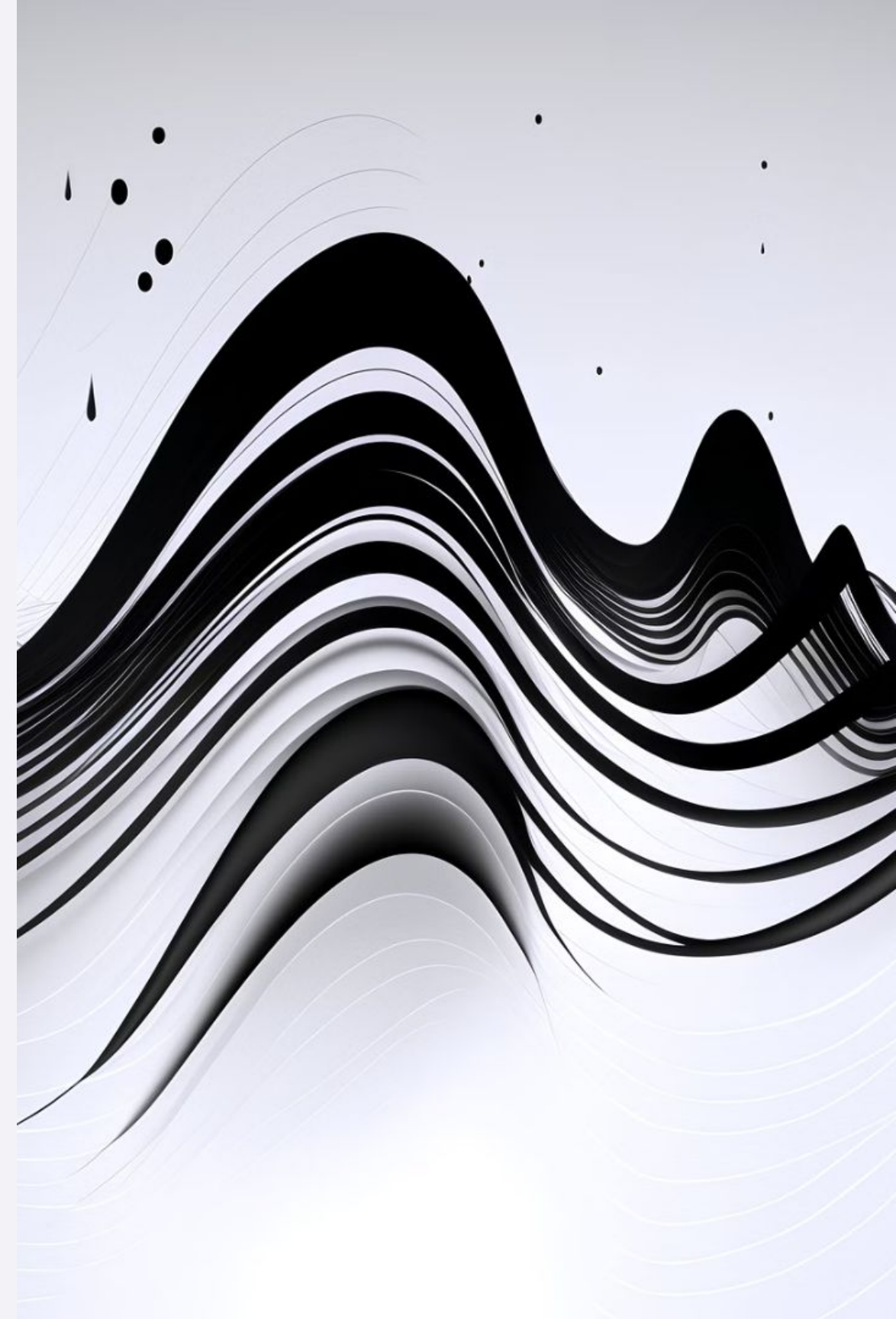
## 협업 역량

다양한 배경의 팀원들과 협업 경험

3

## 도전 정신

새로운 기술과 방법론에 도전



# 향후 개선 방향

경험을 바탕으로 한 future work 제안

1

## 데이터 확보

더 다양한 문서 유형 데이터 수집

2

## 모델 개선

최신 아키텍처 및 전이학습 기법 적용

3

## 실용화

실제 업무 환경에서의 적용 및 최적화

