최종 발표

7조

김희균, 송준규, 정승연, 정하연

목차

1. Goal of the Project

“ 높은 점수보다 Overfitting을 막는 최대한 내성이 강한 모델을 만들자”

(노이즈, 테두리, shifted)

2. 데이터 분석 및 전처리

(1) 15개(0~9, +, -, /, x, =) 클래스 이외의 데이터 → **제거**

	숫자	기호
Training	(15119, 28, 28)	(15329, 28, 28)
Test	(2160, 28, 28)	(2190, 28, 28)

〈Data Cleaning 이전 dataset의 shape〉

	숫자 + 기호
Training	(30448, 28, 28)
Test	(4350, 28, 28)

〈15개 클래스 분류기를 위한 데이터 통합〉

최종 데이터셋 shape	
Training	(26249, 28, 28)
Test	(3730, 28, 28)

〈최종 데이터셋〉

제거된 데이터셋 개수	
Training	4199 (13.8%)
Test	620 (14.3%)

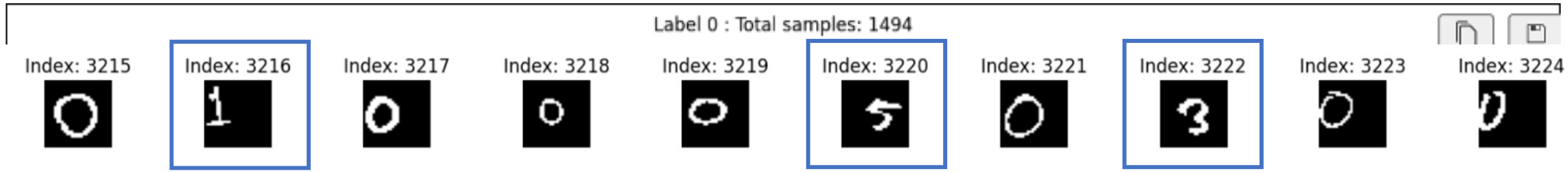
〈15개 이외 라벨 제거〉

2. 데이터 분석 및 전처리

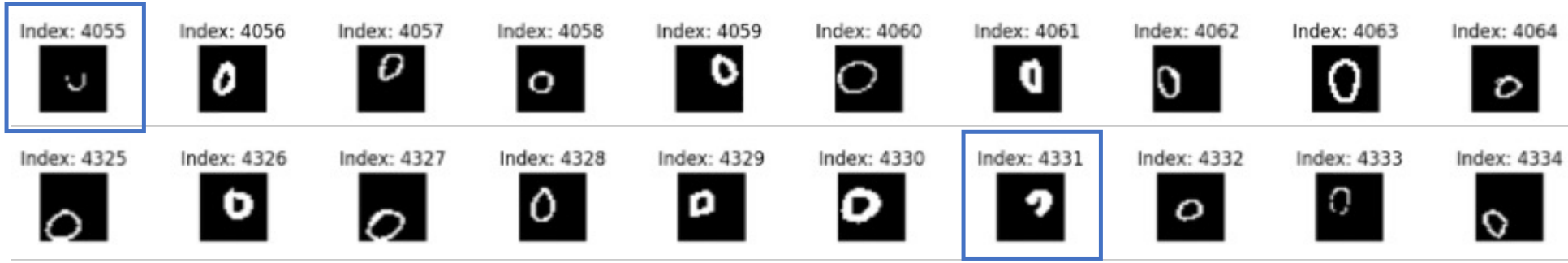
- Discover and Visualize the data

(1) 15개(0~9, +, -, /, x, =) 클래스 이외의 데이터 → **제거**

(2) 잘못된 라벨링 → **육안으로 판별 후 제거**



(3) 학습에 애매한 데이터 → **육안으로 판별 후 제거** (2, 3번 과정에서 1910(train)+209(test)개 삭제됨)



2. 데이터 분석 및 전처리

- Discover and Visualize the data

(4) Handmade dataset 데이터 수 : 약 30000개 → **Combined Dataset 구성 (총 60,000개, trainset : 각 4200개, testset : 각 700개)**

(5) 픽셀값(feature 값)의 차이 → **파이프라인에 Normalizer( ) 추가 (Normalizer에 대한 분석 및 결정 과정 추가 필요)**

- Original Dataset : 0~255
- Handmade Dataset : 0~1

(6) 노이즈가 있는 데이터 → **노이즈 추가하여 trainset, testset 구축 + 파이프라인에 denoising 함수 추가**

- 픽셀값이 0에서 1이라고 하면, 0에서 0.4 사이의 랜덤 값을 픽셀에 추가
- 임계값을 0.4로 설정하여 픽셀값이 0.4보다 크면 유효한 값(숫자나 기호)으로 판별

2. 데이터 분석 및 전처리

- Discover and Visualize the data

shift된 데이터, 테두리가 남은 데이터 → **파이프라인에 중앙화, 최대화 과정 추가**

(7) 비슷한 숫자 및 기호의 존재 → **삭제, 향후 Confusion Matrix 분석 계획**

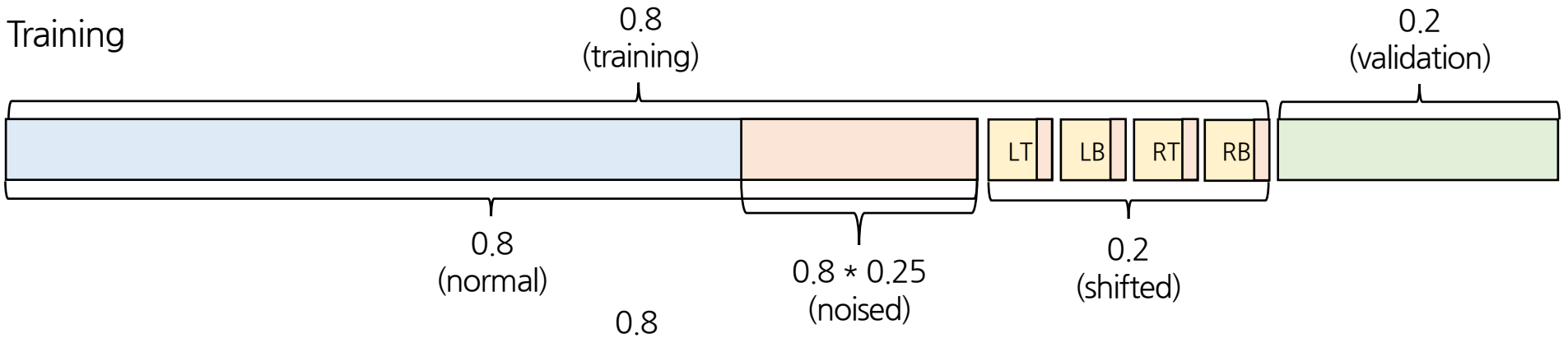
(8) 데이터 크기 : 28*28 데이터만 있음 → **현재 가진 데이터를 확대, 축소해서 새로운 데이터를 만드는 데 한계가 있음**

2. 데이터 분석 및 전처리

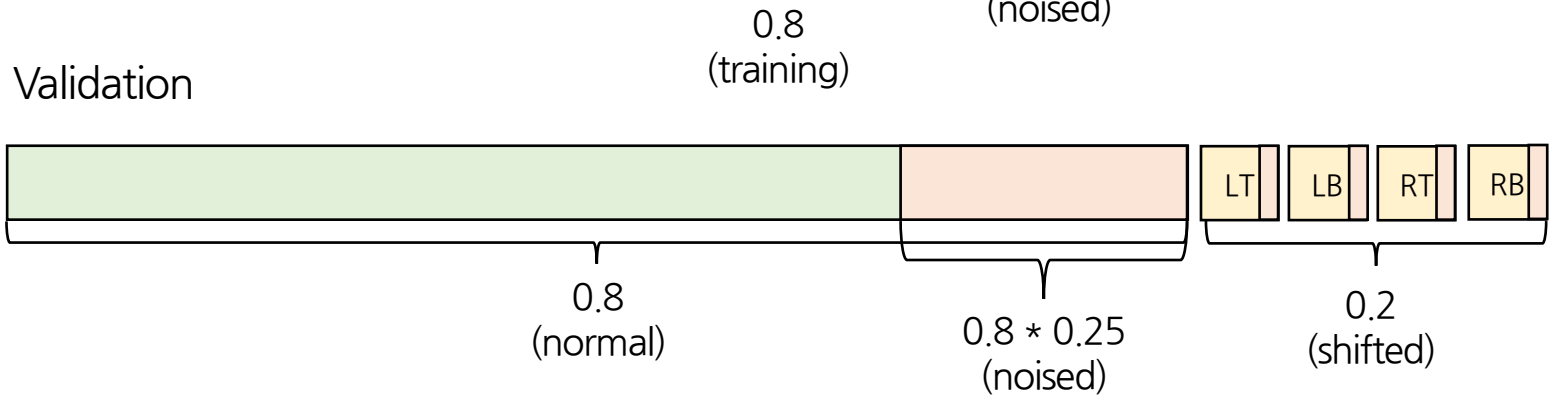
- Prepare the data

Handmade Dataset

- Training



- Validation

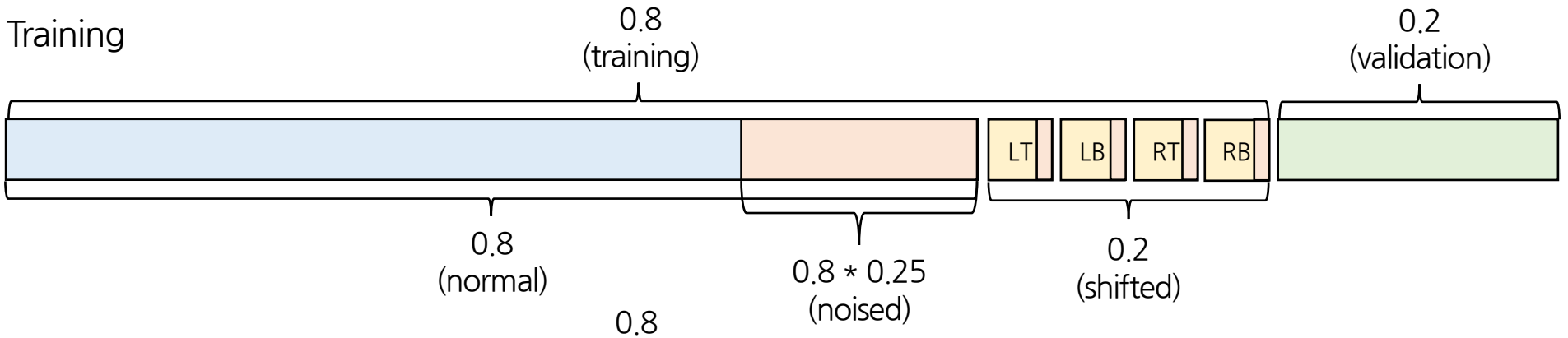


2. 데이터 분석 및 전처리

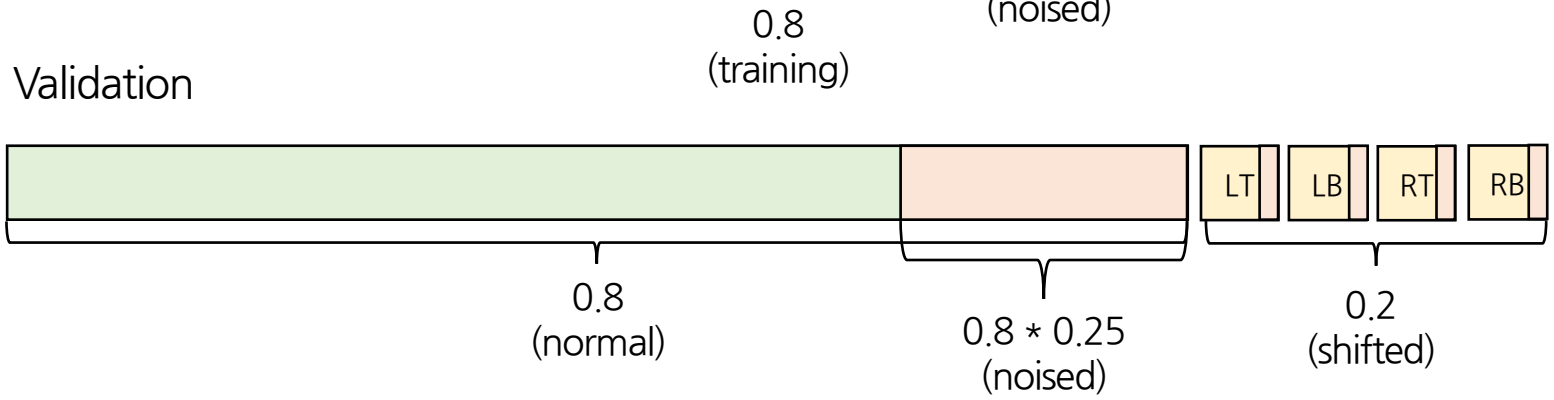
- Prepare the data

Combined Dataset (Training)

- Training



- Validation



2. 데이터 분석 및 전처리

- Prepare the data

- **Final Test Dataset**
  - Training

(Let's Train with the Original dataset)

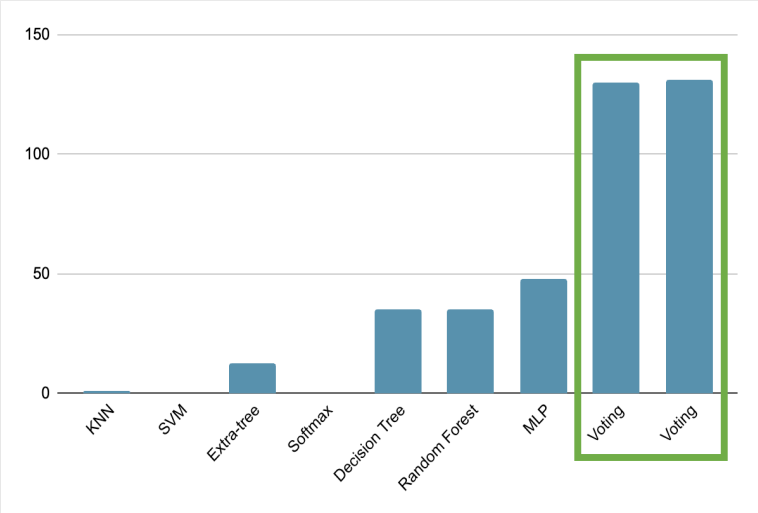
(1) Original Dataset → 학습, 예측 시간이 너무 긴 SVM, VotingClassifier를 사용하지 말자.

Model	KNN	SVM	Extra-tree	Softmax	Decision Tree	Random Forest	MLP	Voting (hard)	Voting (soft)
Training Time (sec)	0.9169		12.76156	0.05638	35.14291	35.14291	47.88176	~~129.84410~~	~~131.20013~~
Original Testset 예측시간	0.97564		0.29157	0.05638	0.02232	0.22734	0.05838	~~8.99141~~	~~9.08884~~
Accuracy on Original Testset	0.97564		0.97000	0.91979	0.87364	0.96636	0.97400	0.97336	0.97479
Team Testset 예측시간	0.23000		0.01019	0.00128	0.16250	0.00857	0.20750	0.77231	0.76504
Accuracy on Team Testset	0.23000		0.18000	0.18500	0.16250	0.19500	0.20750	0.19500	0.19750

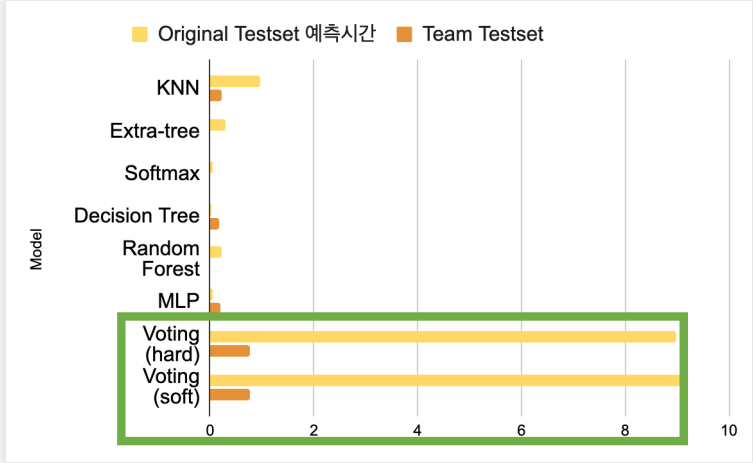
(Let's Train with the Original dataset)

(1) Original Dataset → 학습, 예측 시간이 너무 긴 SVM, VotingClassifier를 사용하지 말자.

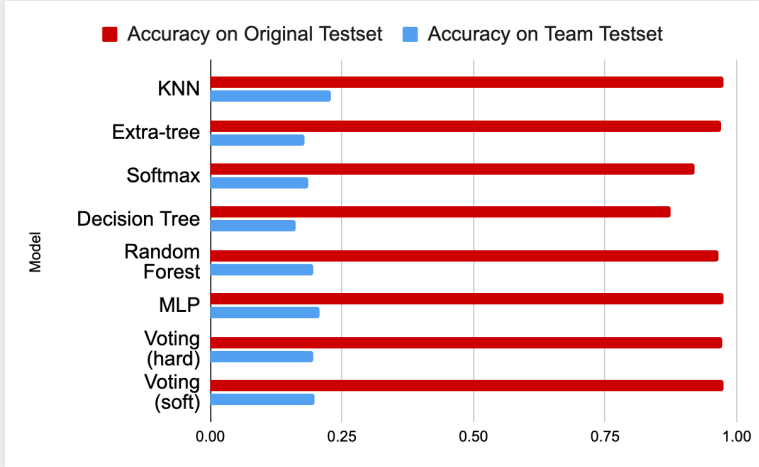
- Training Time (sec)



- Inference Time (Original, Team)



- Accuracy (Original, Team)



Original Testset : 예측이 잘 되었다. (약 90%) ←

Team Testset : 거의 예측이 안됨 (약 20%) ←

3. Select and Train Models

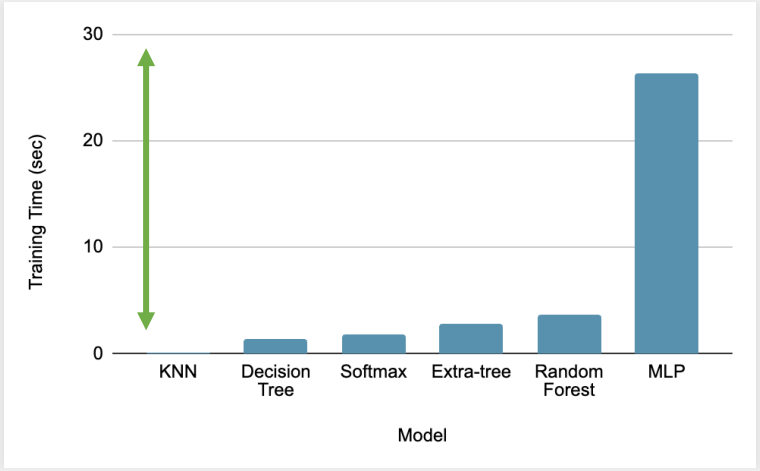
(2) Handmade Dataset → Softmax, DecisionTree의 성능이 낮으니 사용하지 말자.

Model	KNN	SVM	Extra-tree	Softmax	Decision Tree	Random Forest	MLP	Voting (hard)	Voting (soft)
Training Time (sec)	0.09063		2.82671	1.77935	1.44450	3.69965	26.31937		
Original Testset 예측시간	3.95527		0.23225	0.04901	0.02034	0.21984	0.08072		
Accuracy on Original Testset	0.68907		0.68571	~~0.35757~~	~~0.32886~~	0.62143	0.71271		
Team Testset 예측시간	0.05734		0.01308	0.00144	0.00124	0.01171	0.00197		
Accuracy on Team Testset	0.41500		0.37250	~~0.21000~~	~~0.22750~~	0.33000	0.38250		

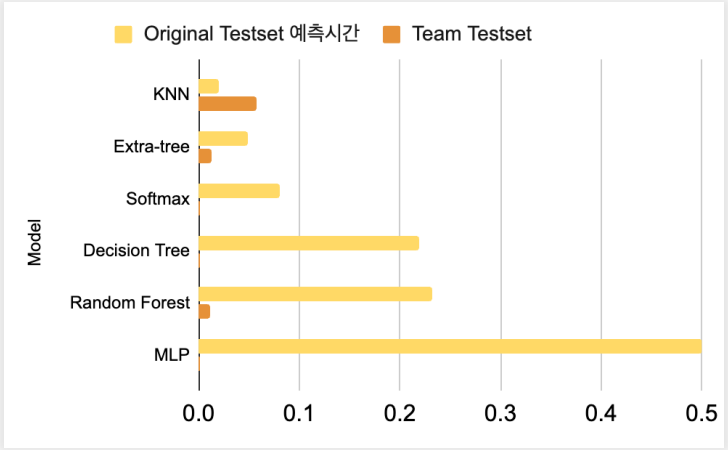
3. Select and Train Models

(2) Handmade Dataset → Softmax, DecisionTree의 성능이 낮으니 사용하지 말자.

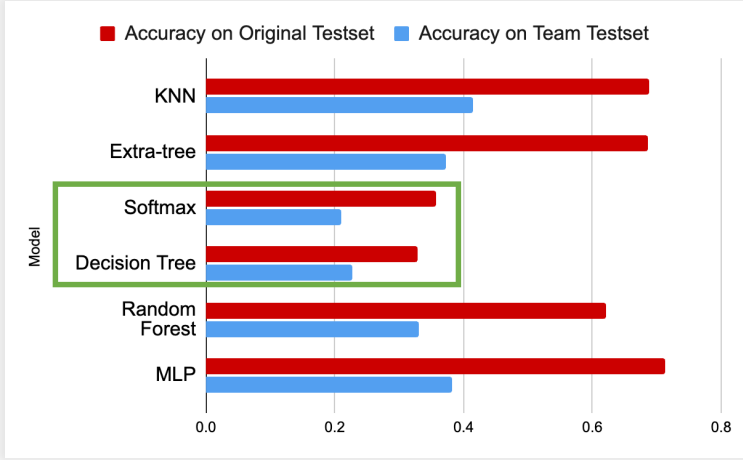
- Training Time (sec)



- Inference Time (Original, Team)



- Accuracy (Original, Team)



Original Testset : 예측 잘 안됨 (약 40% ~ 70%) ←

Team Testset : 예측 잘 안되지만 이전보다 성능 향상됨 (약 30~40%) ←

3. Select and Train Models

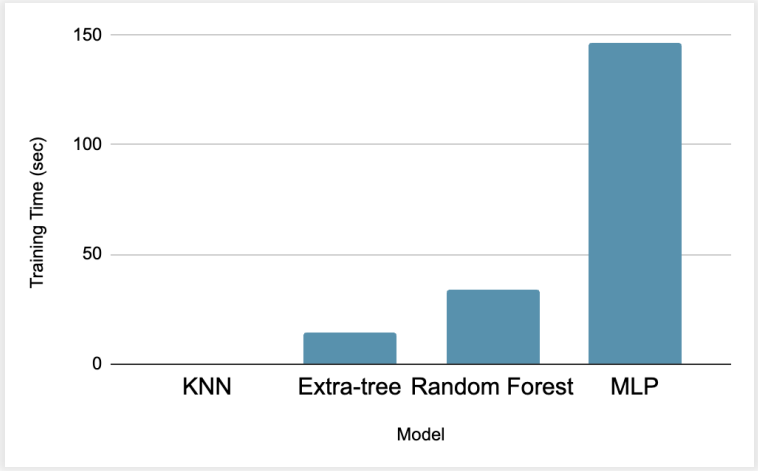
- **Combined Dataset (10 classes)** → KNN은 예측 시간이 길고, Random은 성능이 낮으니 사용하지 말자.

Model	KNN	SVM	Extra-tree	Softmax	Decision Tree	Random Forest	MLP	Voting (hard)	Voting (soft)
Training Time (sec)	0.29248		14.58727			33.62962	146.48267		
Original Testset 예측시간	~~19.26764~~		0.31900			0.24504	0.0561		
Accuracy on Original Testset	0.97514		0.96914			0.96543	0.9764		
Team Testset 예측시간	0.24254		0.01468			0.01233	0.00418		
Accuracy on Team Testset	0.43000		0.39000			~~0.34500~~	0.38250		

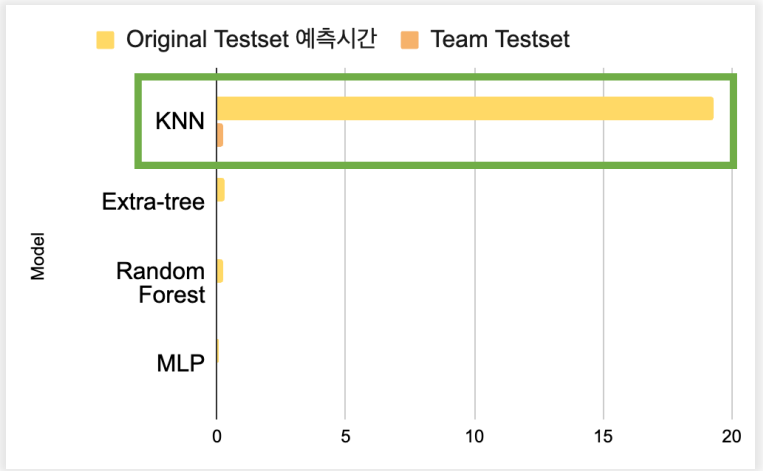
3. Select and Train Models

- Combined Dataset (10 classes) → KNN은 예측 시간이 길고, Random은 성능이 낮으니 사용하지 말자.

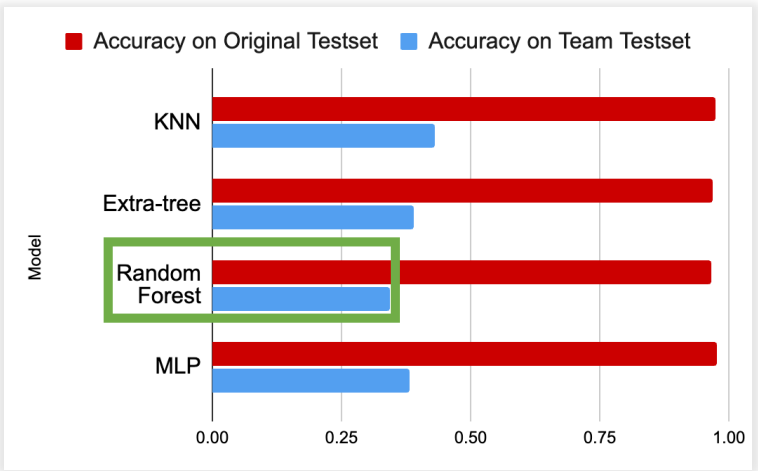
- Training Time (sec)



- Inference Time (Original, Team)



- Accuracy (Original, Team)



→ 후보 1 : Extra-Tree

→ 후보 2 : MLP

4. Fine-Tune the Model

- 15 classes 분류기 : GridSearch로 최적의 파라미터 찾기

(1) Extra-Tree

- **n_estimators** : [100, 200, **300**]
- **max_depth** : [10, **20**]

- Search (Training) Time : 415(sec)
- Validation Dataset - Accuracy : 0.9282
- Validation Dataset - Inference Time : 0.87(sec)

(2) MLP

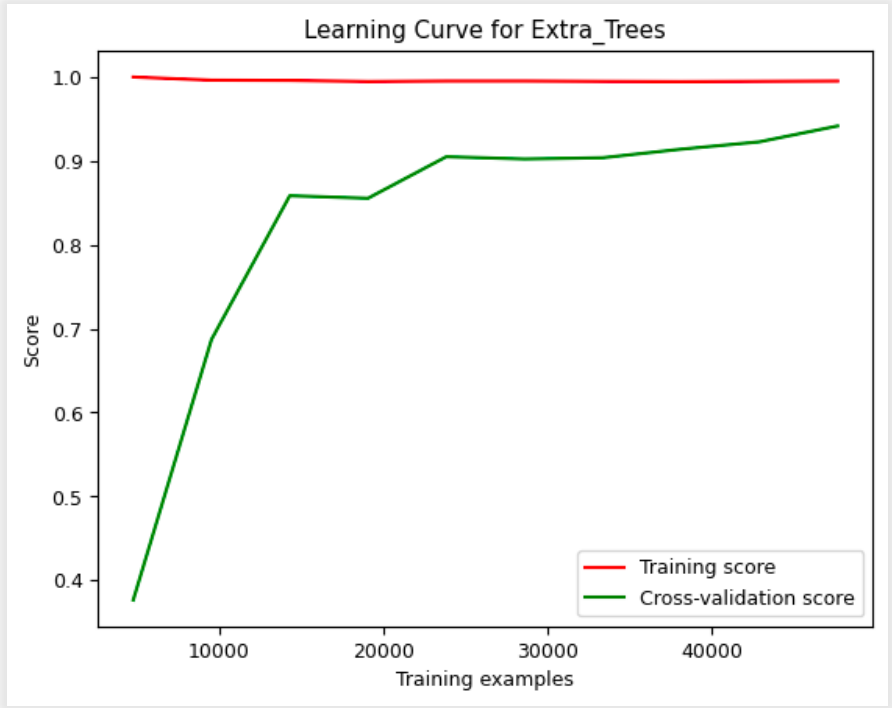
- max_iter : [500, 1000, **2000**]
- alpha : [0.0001, 0.001, **0.01**, 0.1]

- Search (Training) Time : 8017(sec)
- **Validation Dataset - Accuracy** : 0.9387
- **Validation Dataset - Inference Time** : 0.47(sec)

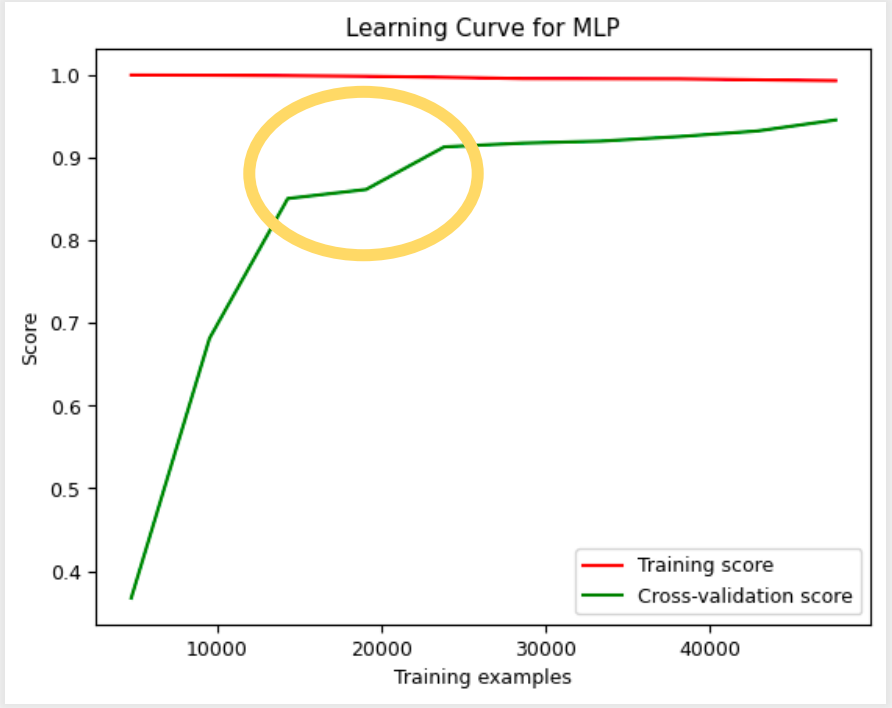
4. Fine-Tune the Model

- 15 classes 분류기 : Learning Curve

(1) Extra-Tree



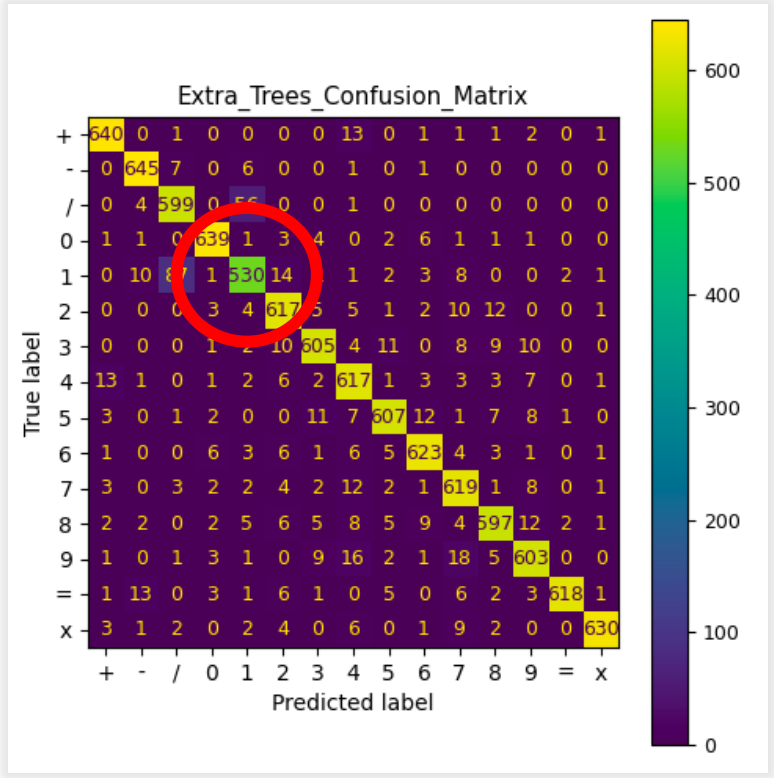
(2) MLP



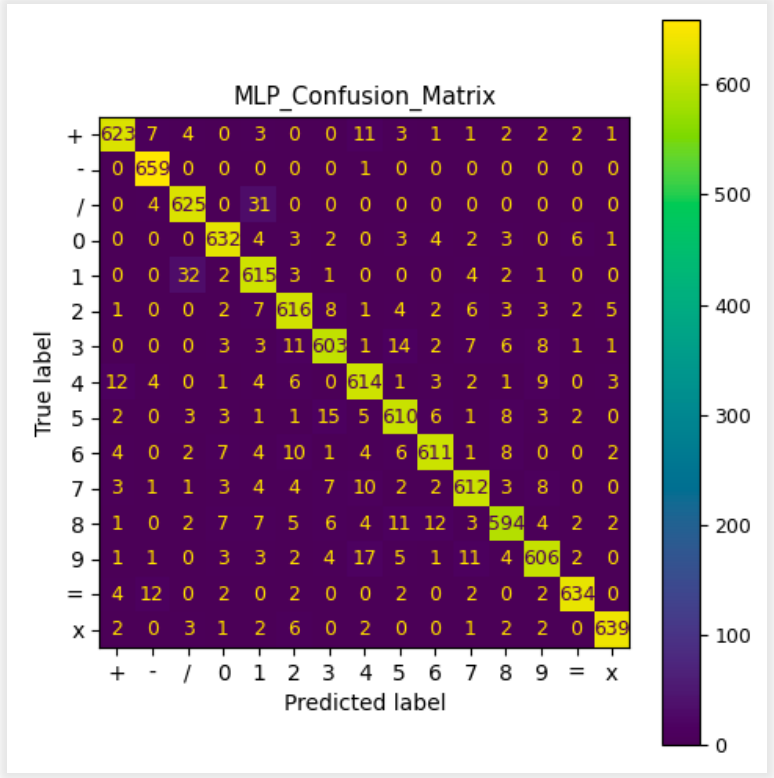
4. Fine-Tune the Model

- 15 classes 분류기 : **Confusion Matrix** (Validation Dataset으로 진행)

(1) Extra-Tree



(2) MLP



→ 최종 : **MLP**가 가장 좋은 성능을 보인다.

4. Fine-Tune the Model

- 15 classes 분류기 : **Voting Classifier**

(1) Voting Classifier (soft)

- Search (Training) Time : 197(sec)
- Validation Dataset - Accuracy : 0.9422
- Validation Dataset - Inference Time : 1.62(sec)

(2) Voting Classifier (hard)

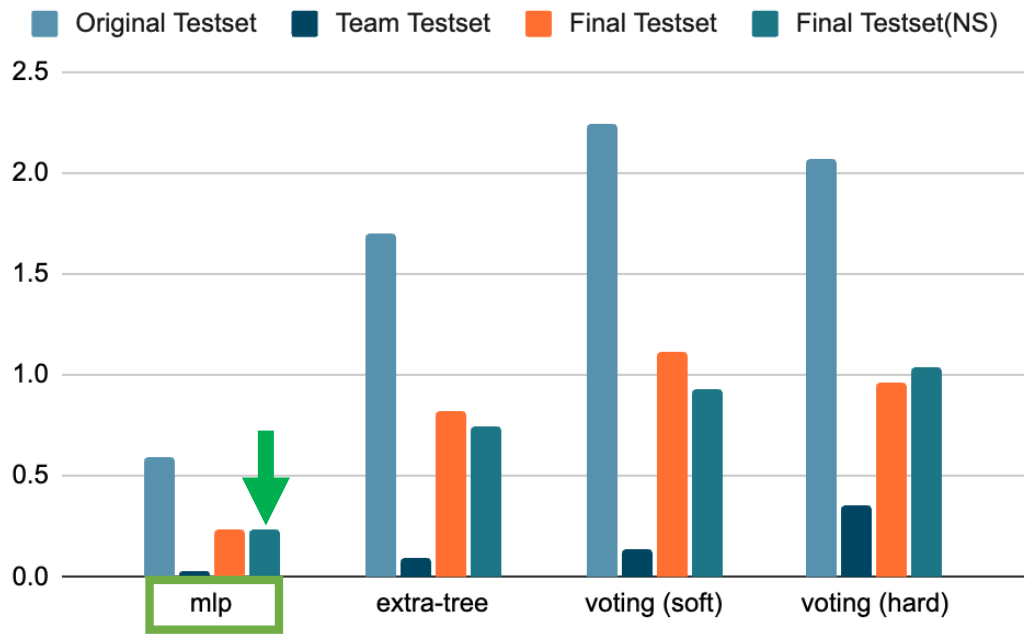
- Search (Training) Time : 233(sec)
- **Validation Dataset - Accuracy** : 0.9305
- Validation Dataset - Inference Time : 1.66(sec)

4. Final Test

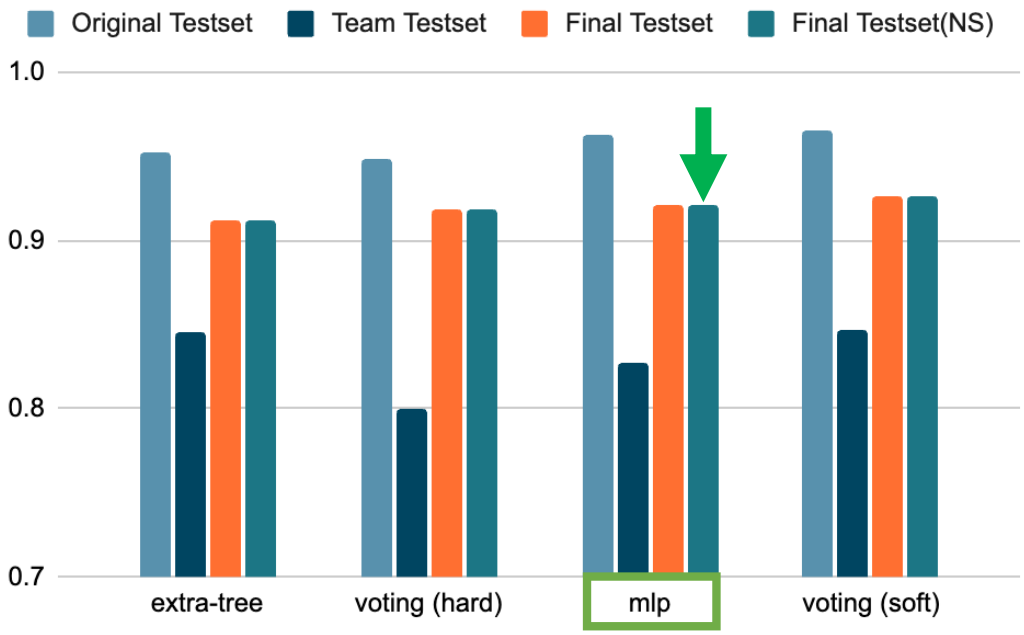
		Extra-Tree	MLP	Voting (hard)	Voting (soft)
Original Testset (10classes, 14,000)	Inference Time	1.69720	0.59328	2.06655	2.25069
	Accuracy	0.95264	0.96271	0.94907	0.96500
Team Testset (10classes, 400)	Inference Time	0.08964	0.02795	0.35179	0.13811
	Accuracy	0.84500	0.82750	0.80000	0.84750
Final Testset (15classes, 5,478)	Inference Time	0.82431	0.23394	0.96035	1.12011
	Accuracy	0.91201	0.92150	0.91822	0.92680
Final Testset (15classes, **Noised + shifted**, 5,478)	Inference Time	0.74779	0.23500	1.04238	0.92722
	Accuracy	0.91201	0.92150	0.91822	0.92680

4. Final Test

- **Inference Time** (Final Testset(NS) 기준 정렬)

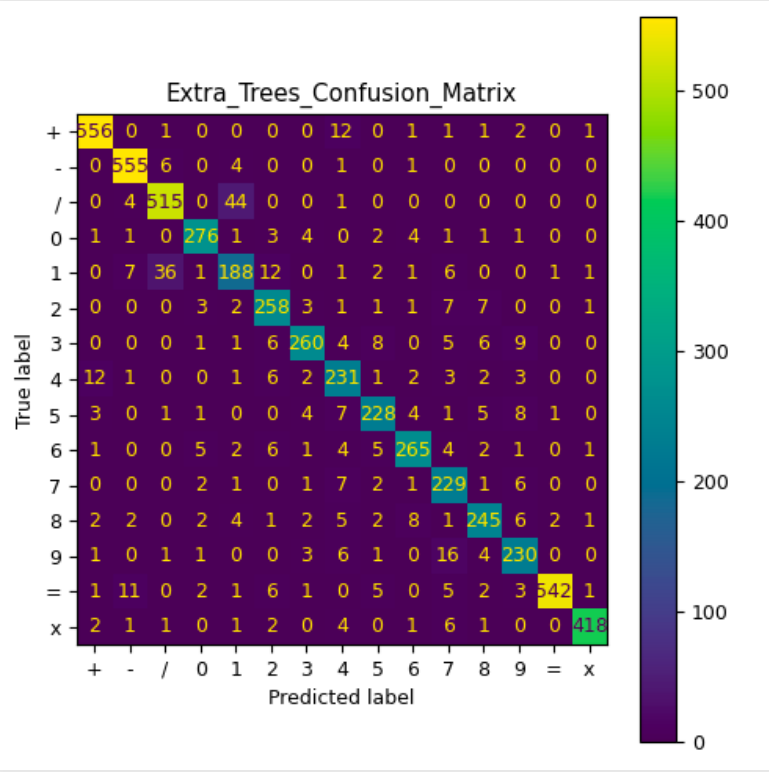


- **Accuracy** (Final Testset(NS) 기준 정렬)

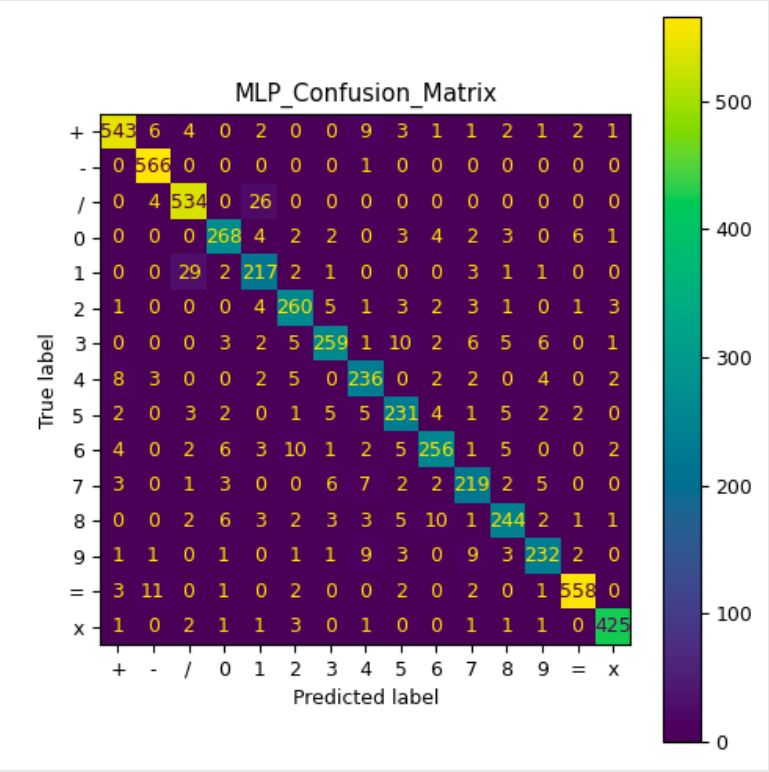


4. Final Test

(1) Extra-Tree



(2) MLP



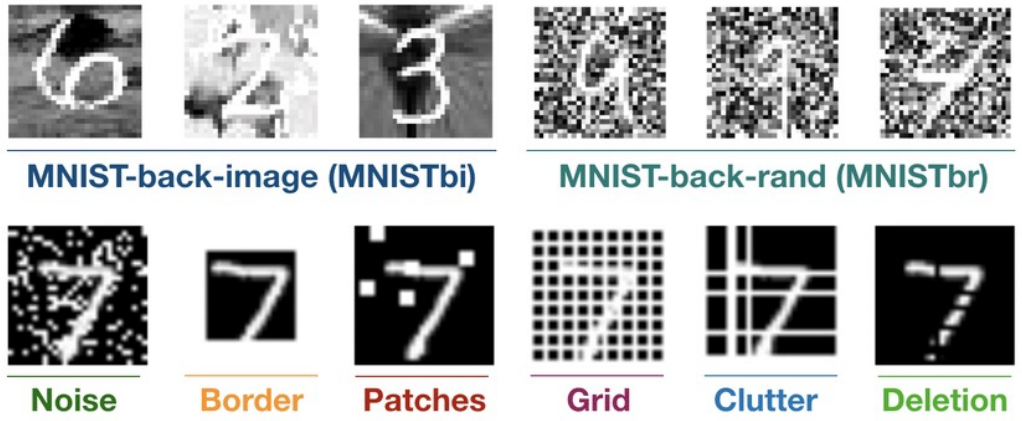
5. Feedback

- Shifted data 처리 방식 결정 과정

5. Feedback

- 한계 1 : Rotate, 테두리

- 한계 2 : 다양한 노이즈에 대한 해결책?



감사합니다