Deepfake Detector

Team Faux Fighters
Cho, Anna
Denq, Christopher
Nelson, Reid (Jackson)

Roadmap

Roadmap

Problem Statement

Roadmap

Problem Statement

Methodology

Roadmap

Problem Statement

Methodology

Analysis

Roadmap

Problem Statement

Methodology

Analysis

Conclusion

01

Problem Statement

“AI is spook.”

—Elon Musk, probably

PROBLEM

Generative AI makes it easy to deepfake.

PROBLEM

Generative AI makes it easy to deepfake.

Deepfake

Real

PROBLEM

Generative AI makes it easy to deepfake.

SOLUTION

We have “out-of-box” deepfake detector.

02

METHODOLOGY

01

“OpenForensics” dataset from a paper presented on computer vision

SOURCE

02

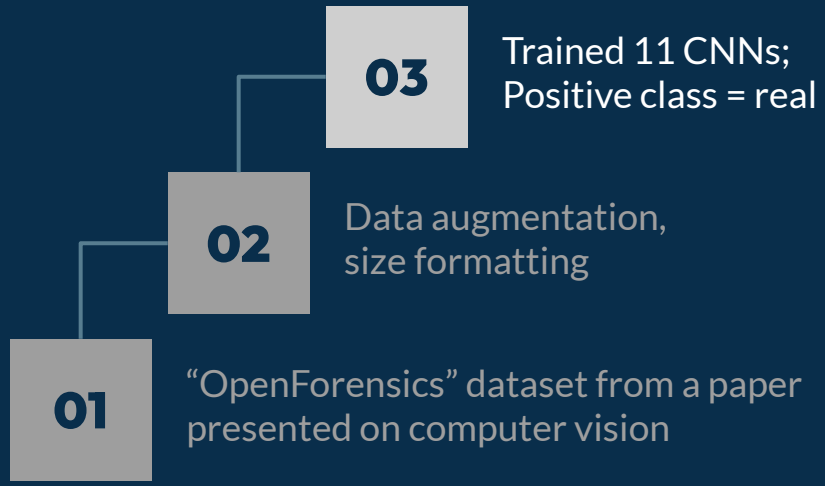
Data augmentation,
size formatting

01

“OpenForensics” dataset from a paper presented on computer vision

CLEAN

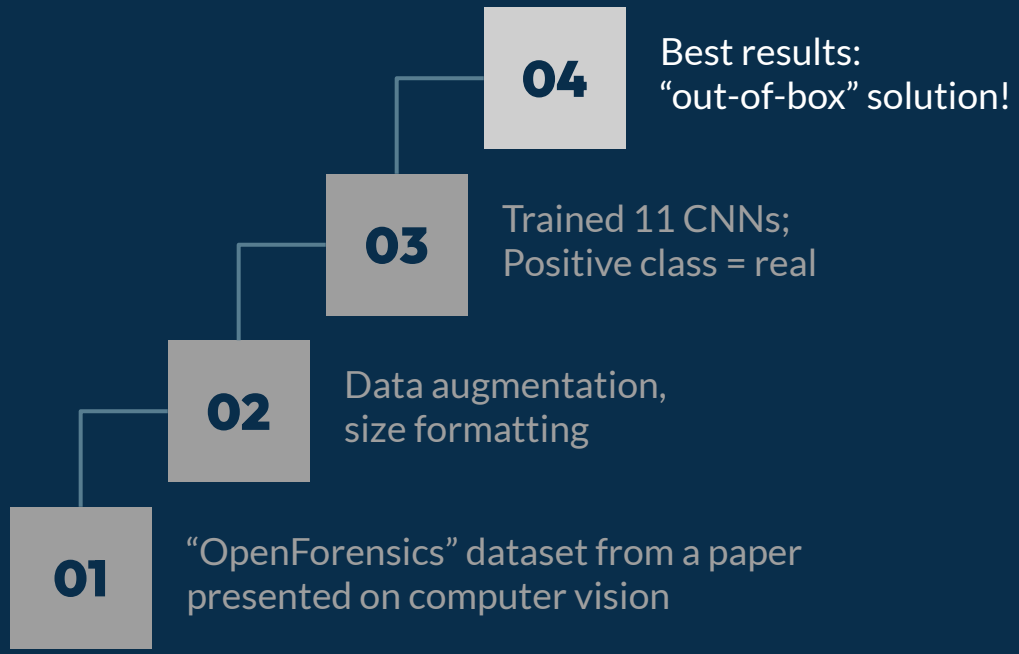
SOURCE



MODEL

CLEAN

SOURCE



RESULT

MODEL

CLEAN

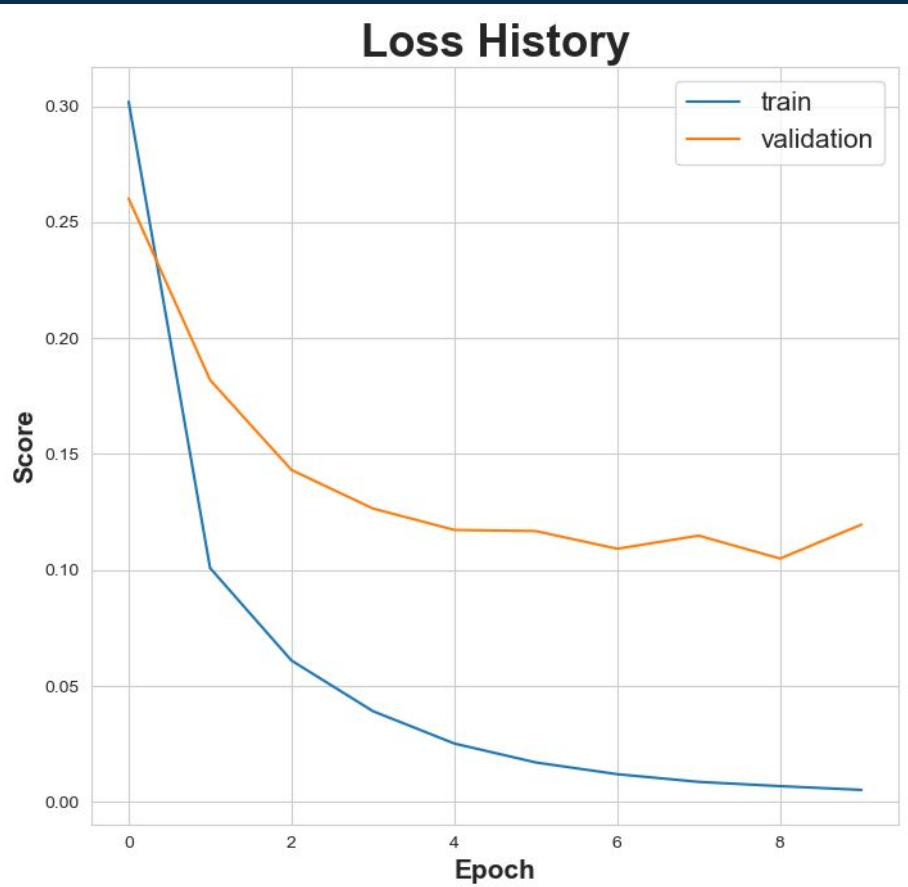
SOURCE

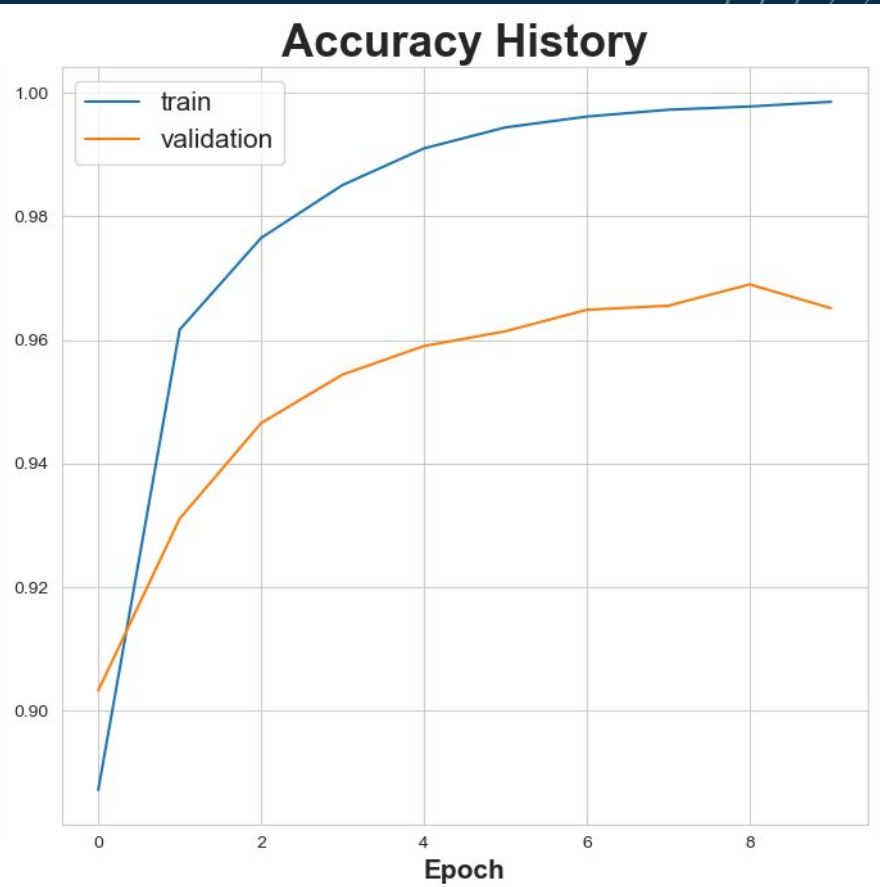
03

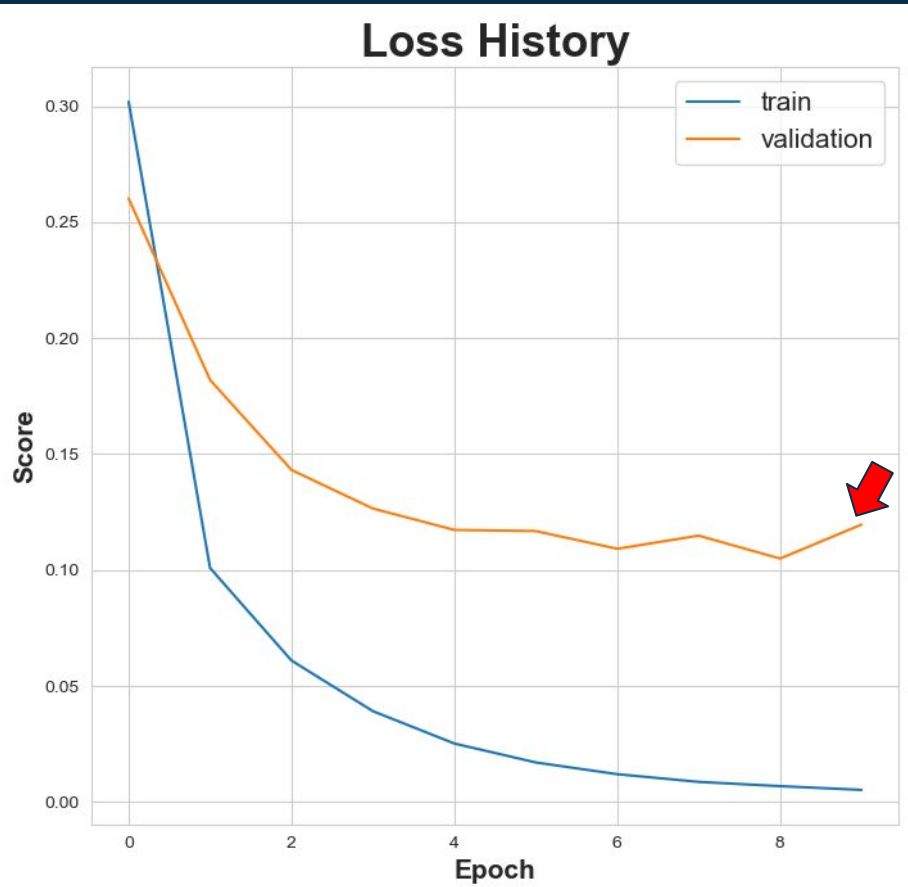
ANALYSIS

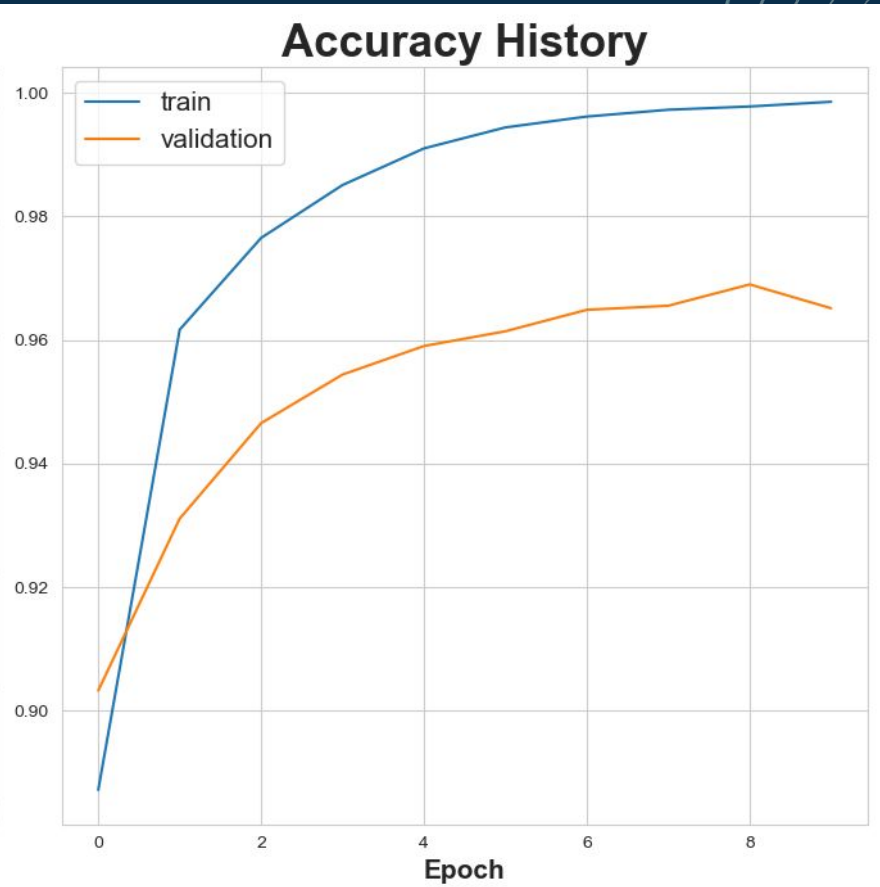
Best Model

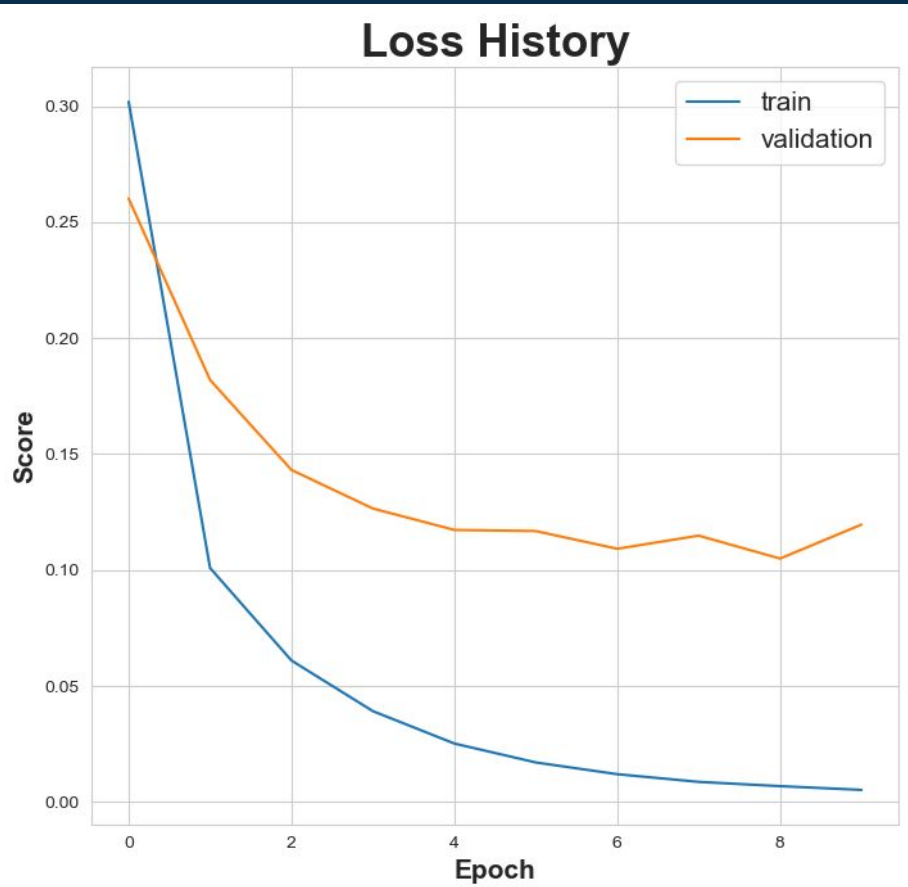
Eff.NetV2_B0

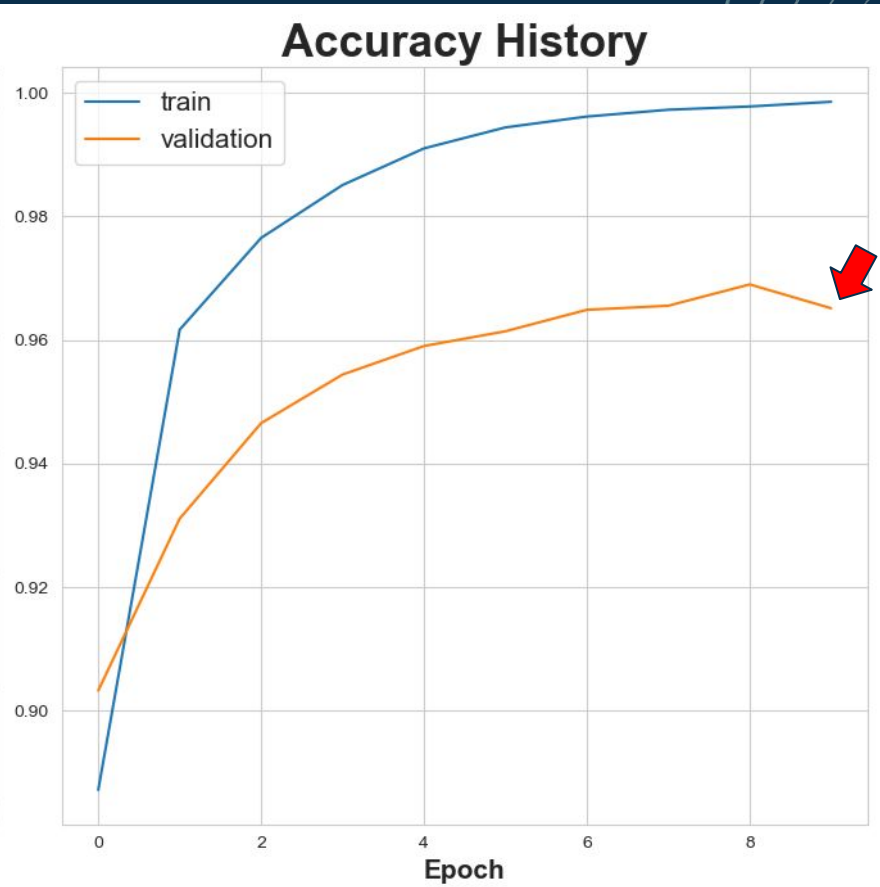

Loss History
train
validation
Score
Epoch
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8

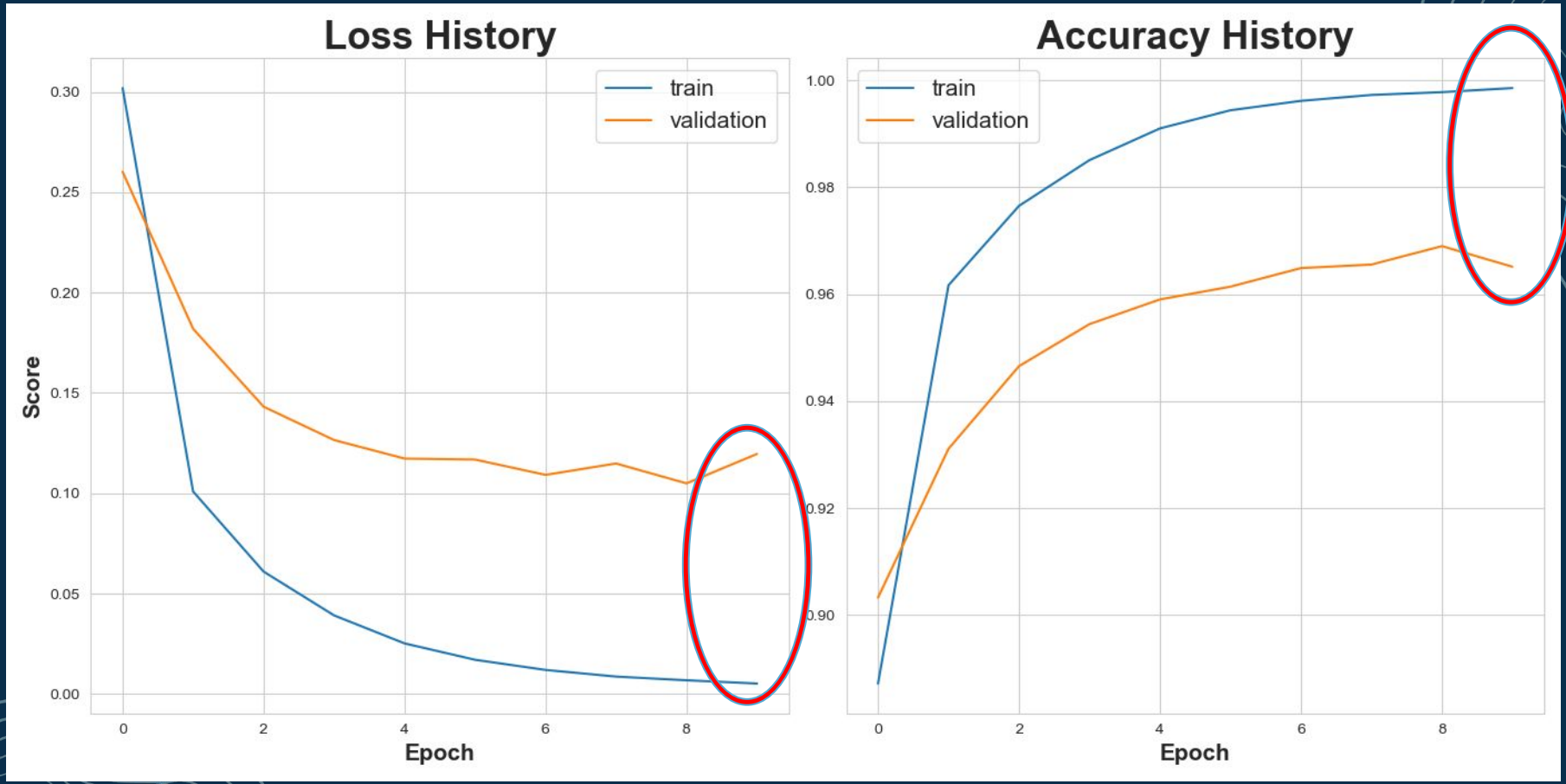

Accuracy History
train
validation
Epoch
1.00
0.98
0.96
0.94
0.92
0.90
0
2
4
6
8


Loss History
train
validation
Score
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
Epoch

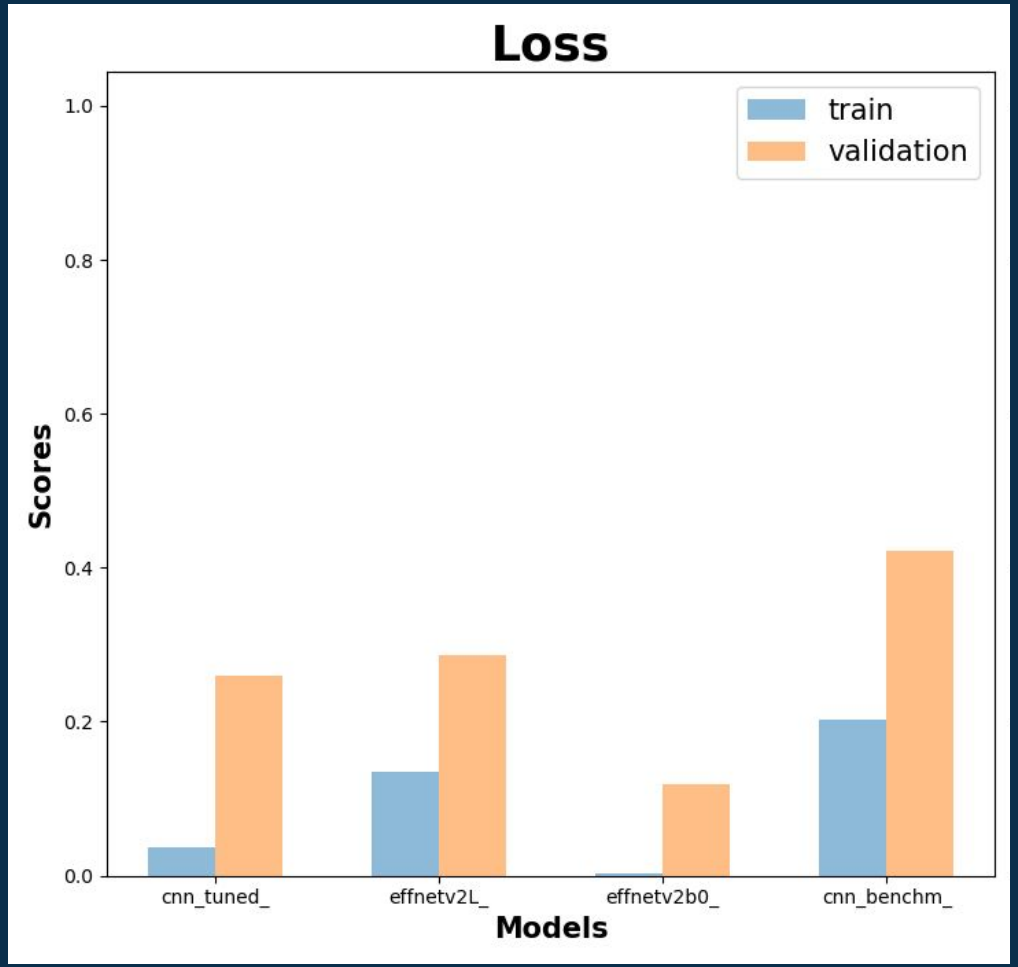

Accuracy History
train
validation
1.00
0.98
0.96
0.94
0.92
0.90
0
2
4
6
8
Epoch

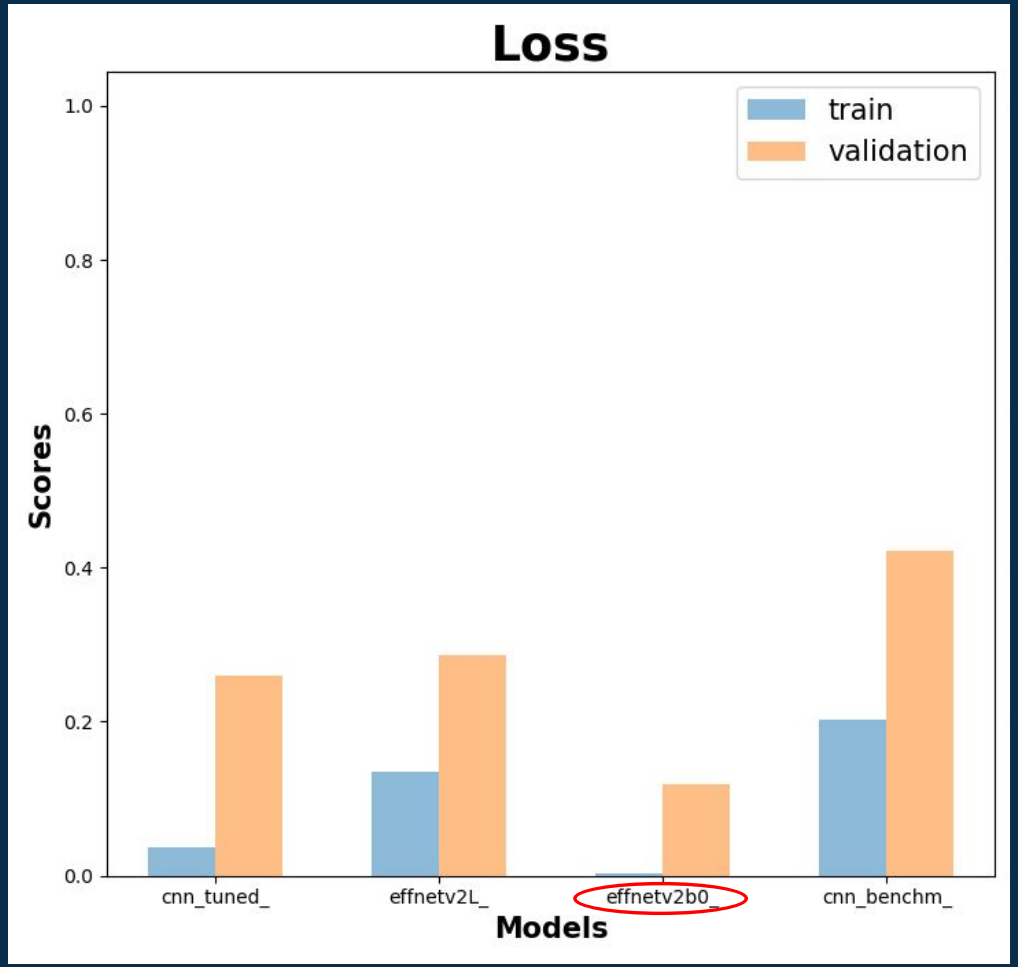

Loss History
train
validation
Score
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
Epoch

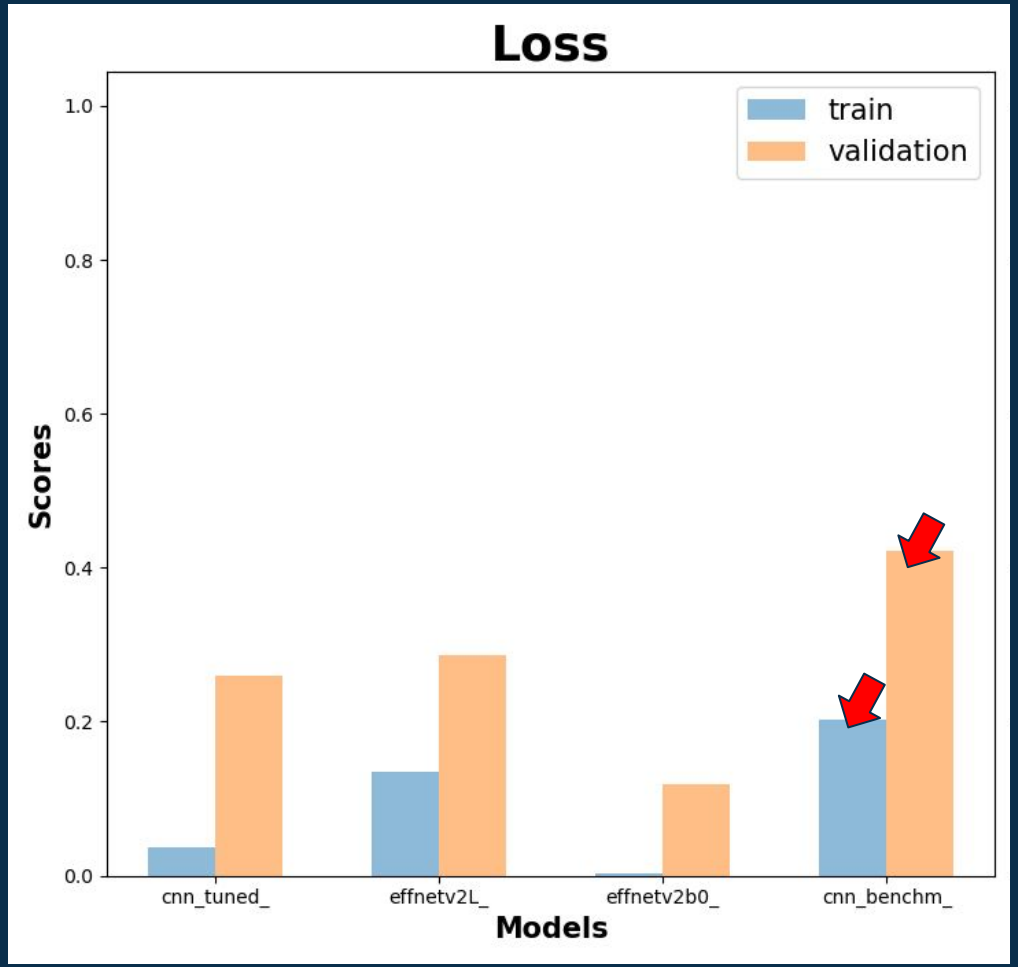

Accuracy History
train
validation
1.00
0.98
0.96
0.94
0.92
0.90
0
2
4
6
8
Epoch

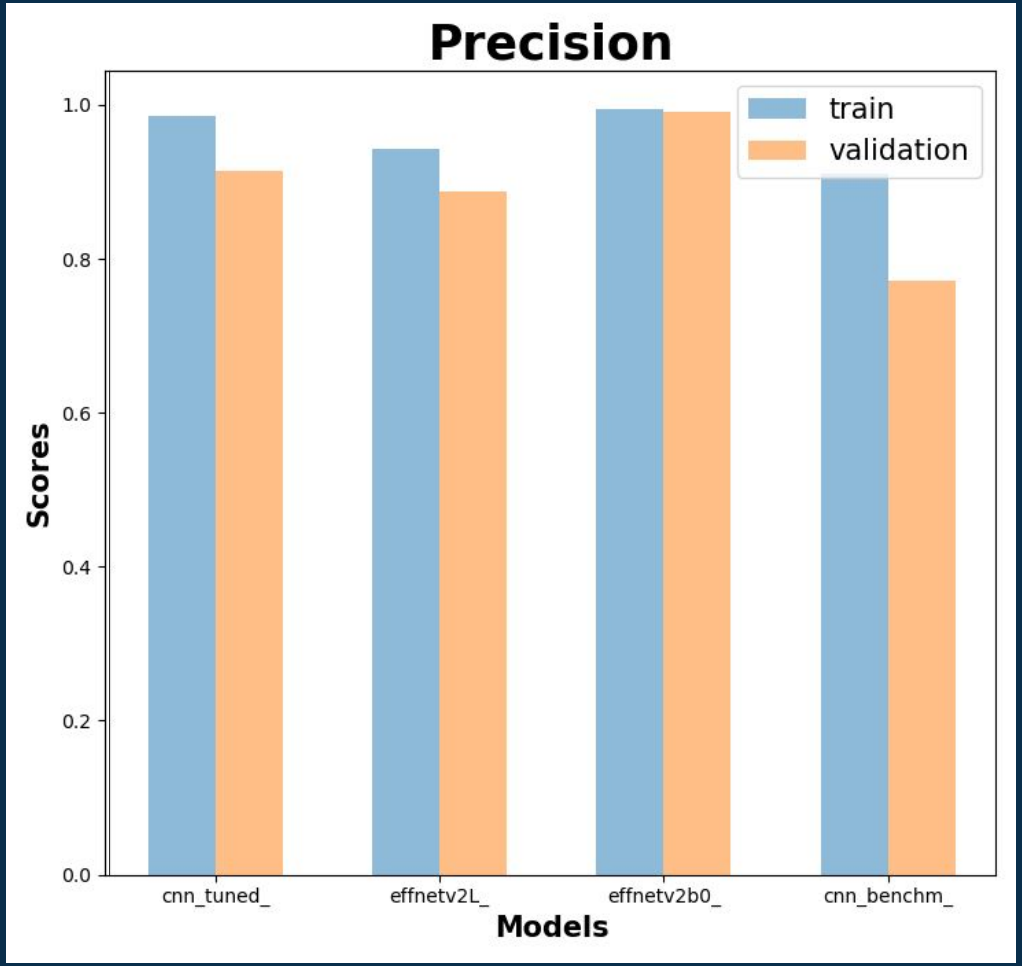

Loss History
train
validation
Score
Epoch
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
Accuracy History
train
validation
Epoch
1.00
0.98
0.96
0.94
0.92
0.90
0
2
4
6
8

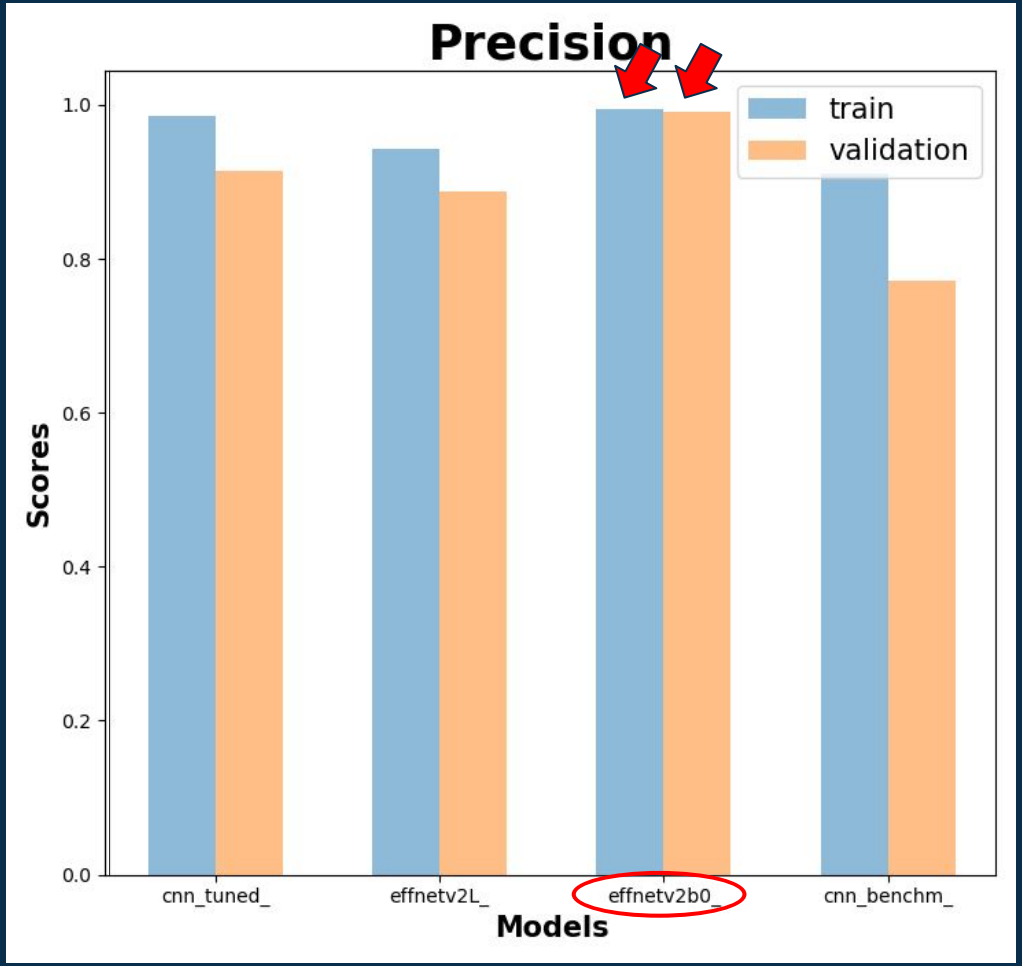
Overall Model Comparisons


Loss
train
validation
1.0
0.8
0.6
0.4
0.2
0.0
Scores
cnn_tuned_
effnetv2L_
effnetv2b0_
cnn_benchm_
Models


Loss
train
validation
1.0
0.8
0.6
0.4
0.2
0.0
Scores
cnn_tuned_
effnetv2L_
effnetv2b0_
cnn_benchm_
Models


Loss
train
validation
1.0
0.8
0.6
0.4
0.2
0.0
Scores
cnn_tuned_
effnetv2L_
effnetv2b0_
cnn_benchm_
Models


Precision
train
validation
1.0
0.8
0.6
0.4
0.2
0.0
Scores
cnn_tuned_
effnetv2L_
effnetv2b0_
cnn_benchm_
Models


Precision
train
validation
Scores
1.0
0.8
0.6
0.4
0.2
0.0
cnn_tuned_
effnetv2L_
effnetv2b0_
cnn_benchm_
Models

04

CONCLUSION

Best

Benchmark

Experimental

Best

Benchmark

Sequential CNN (Base)
Acc: 0.860 | **Prec**: 0.849

Experimental

Best

Benchmark

Sequential CNN (Base)
Acc: 0.860 | **Prec**: 0.849

Experimental

EfficientNetv2_L (TL)
Acc: 0.774 | **Prec**: 0.822

Best

EfficientNetv2_B0
Acc: 0.965 | **Prec**: 0.992

Benchmark

Sequential CNN (Base)
Acc: 0.860 | **Prec**: 0.849

Experimental

EfficientNetv2_L (TL)
Acc: 0.774 | **Prec**: 0.822

TAKEAWAYS

BEST MODEL

EfficientNetv2_B0 Precision: 0.992

TAKEAWAYS

BEST MODEL

EfficientNetv2_B0 Precision: 0.992

UNCERTAINTIES

- Generalize to other types of deepfake image generators?
- Susceptible to image manipulations (eg. blur)?

TAKEAWAYS

BEST MODEL

EfficientNetv2_B0 Precision: 0.992

UNCERTAINTIES

- Generalize to other types of deepfake image generators?
- Susceptible to image manipulations (eg. blur)?

FUTURE

- Retrain larger EfficientNetv2 Model (eg. more compute resources)
- Address overfitting: (eg. data augmentation)

THANK YOU!

Questions?