وزارة العدل
Ministry of Justice

رؤية VISION 2030
المملكة العربية السعودية
KINGDOM OF SAUDI ARABIA

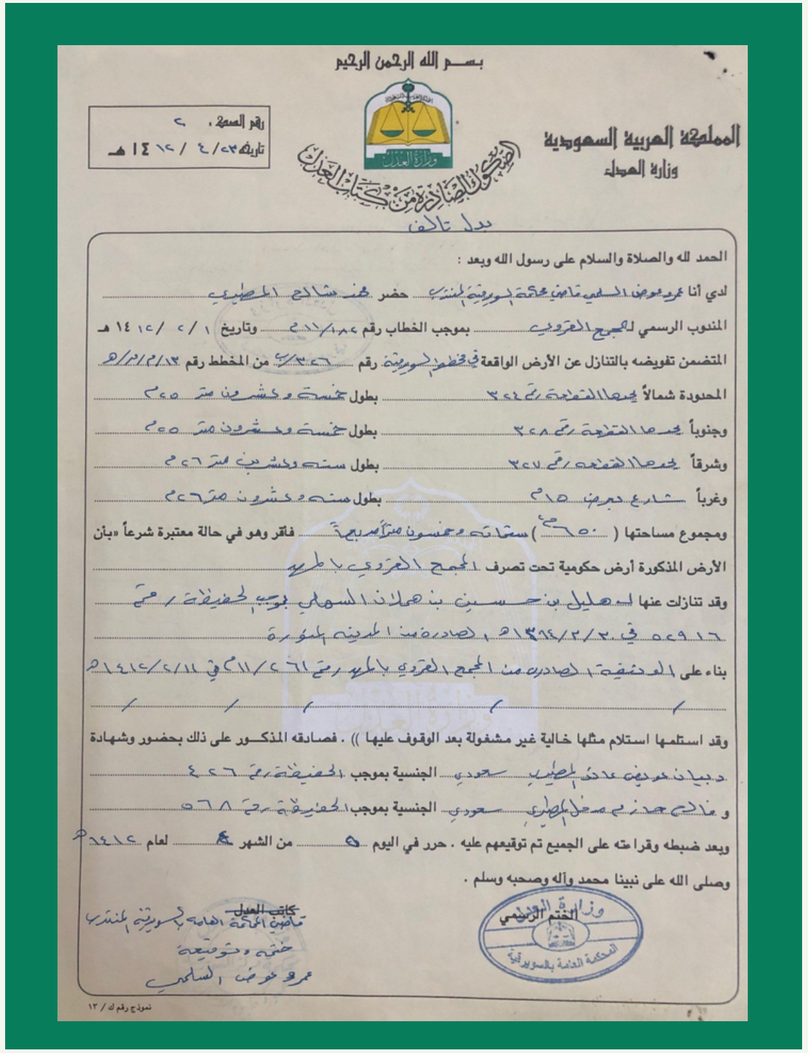
Arabic Handwritten Digit Recognition

01

Problem Definition

Handwritten archived documents:

- Forgery.
- Deterioration.
- Misinterpretation.



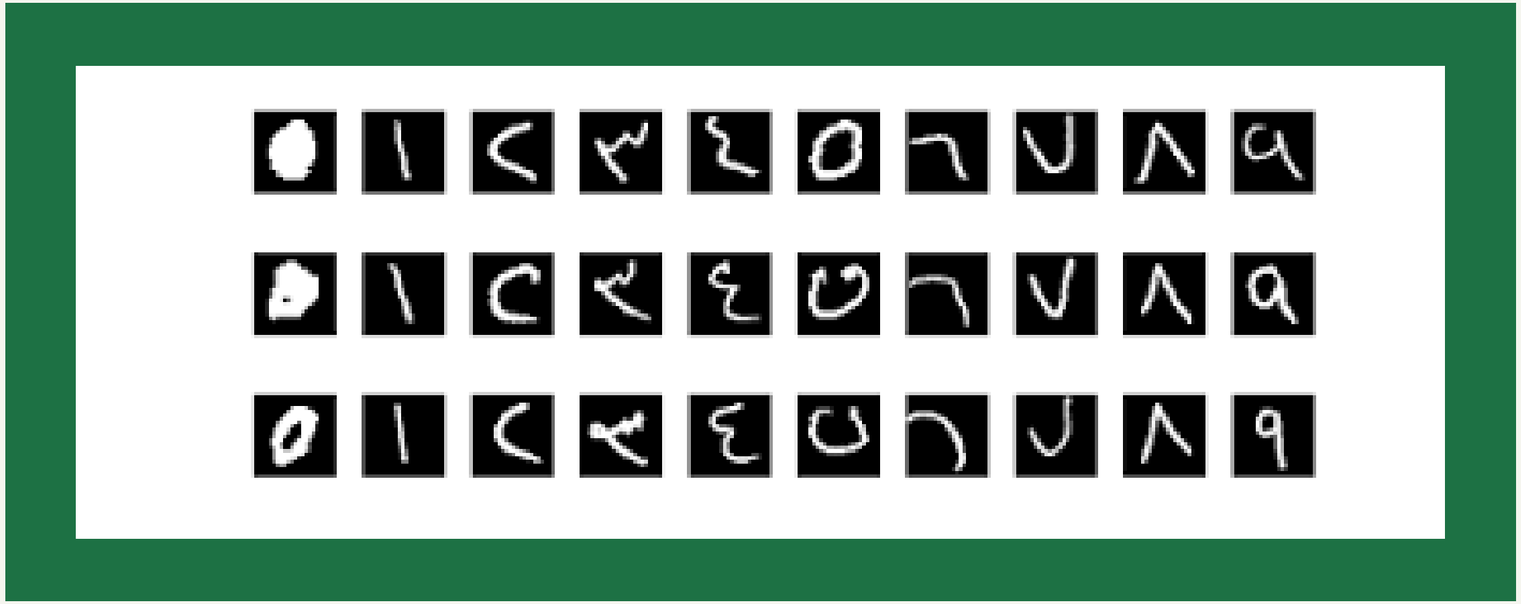
Solution

Build a tool for the ministry to digitize handwritten documents.

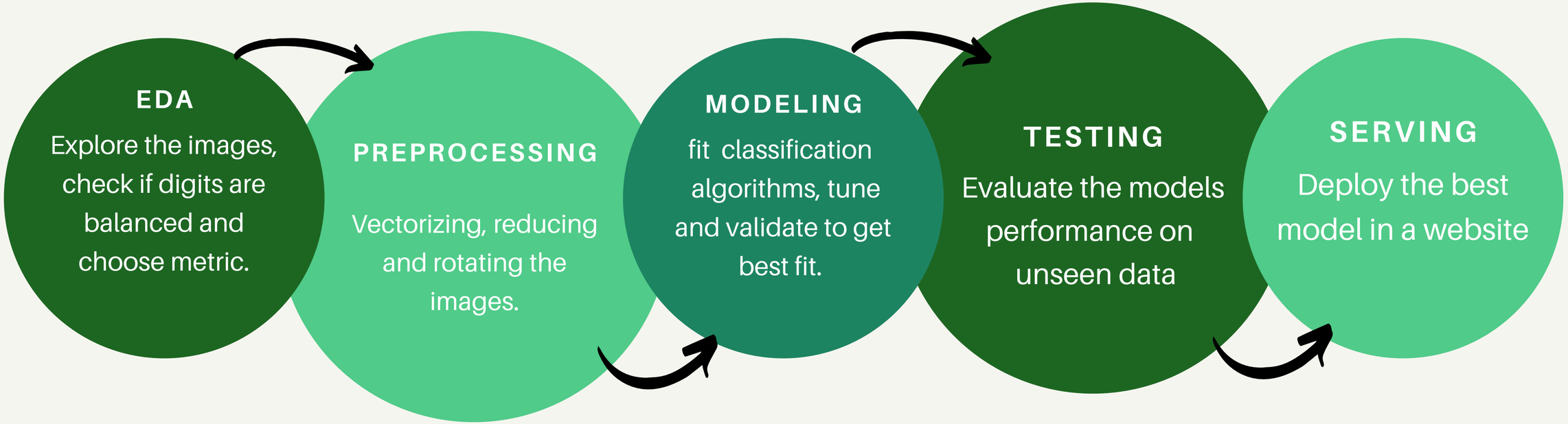
The Dataset

MADBase from The American University in Cairo:

- Different writing styles from all age groups.
- 28x28 Binary images.
- 70 k observations.



Methodology



Tools

Environment

- Python
- Jupyter Notebook
- git

Data Manipulation

- pandas
- numpy
- pickle
- open cv

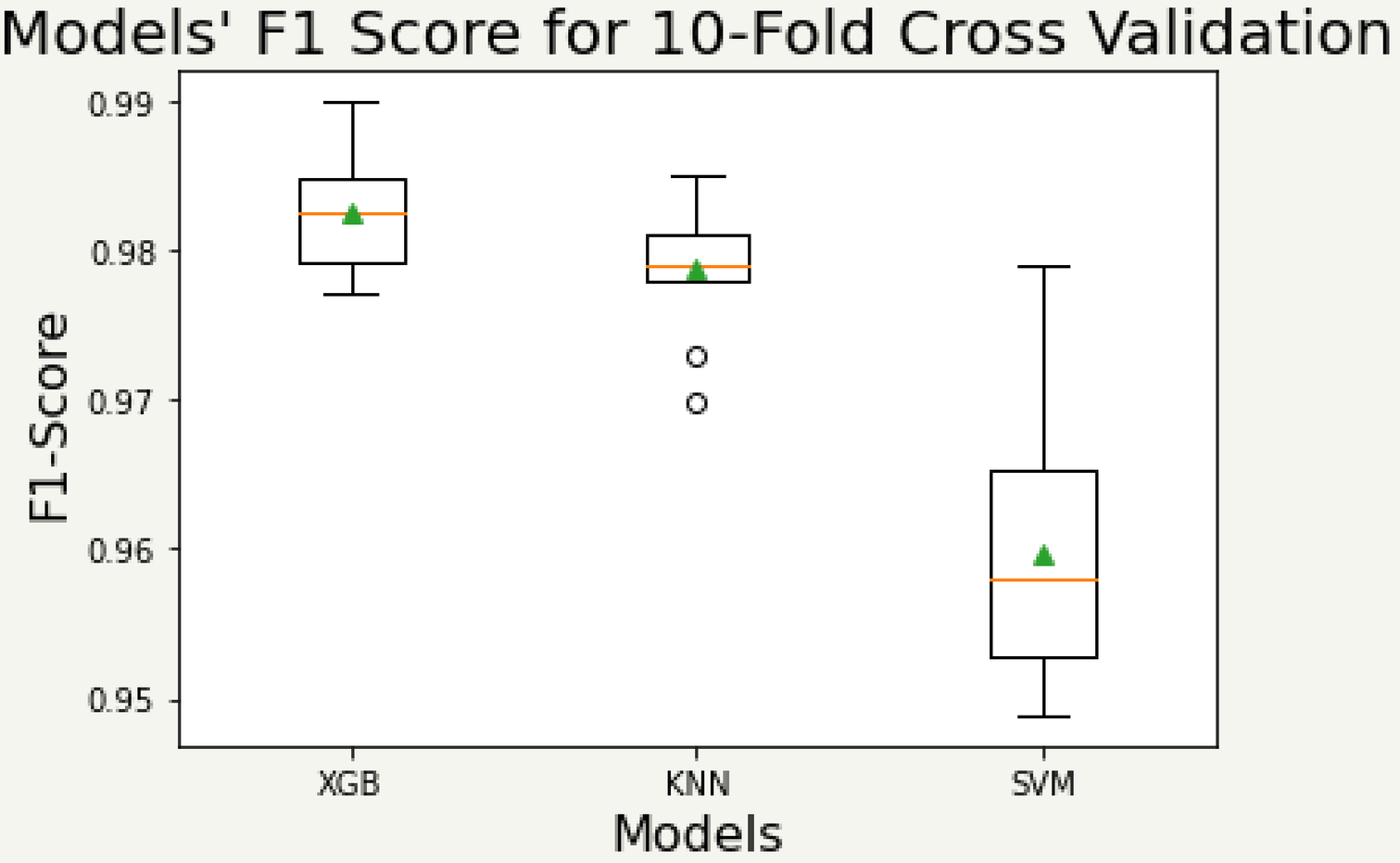
Modeling

- Sci-kit learn
- XGBoost
- matplotlib
- seaborn

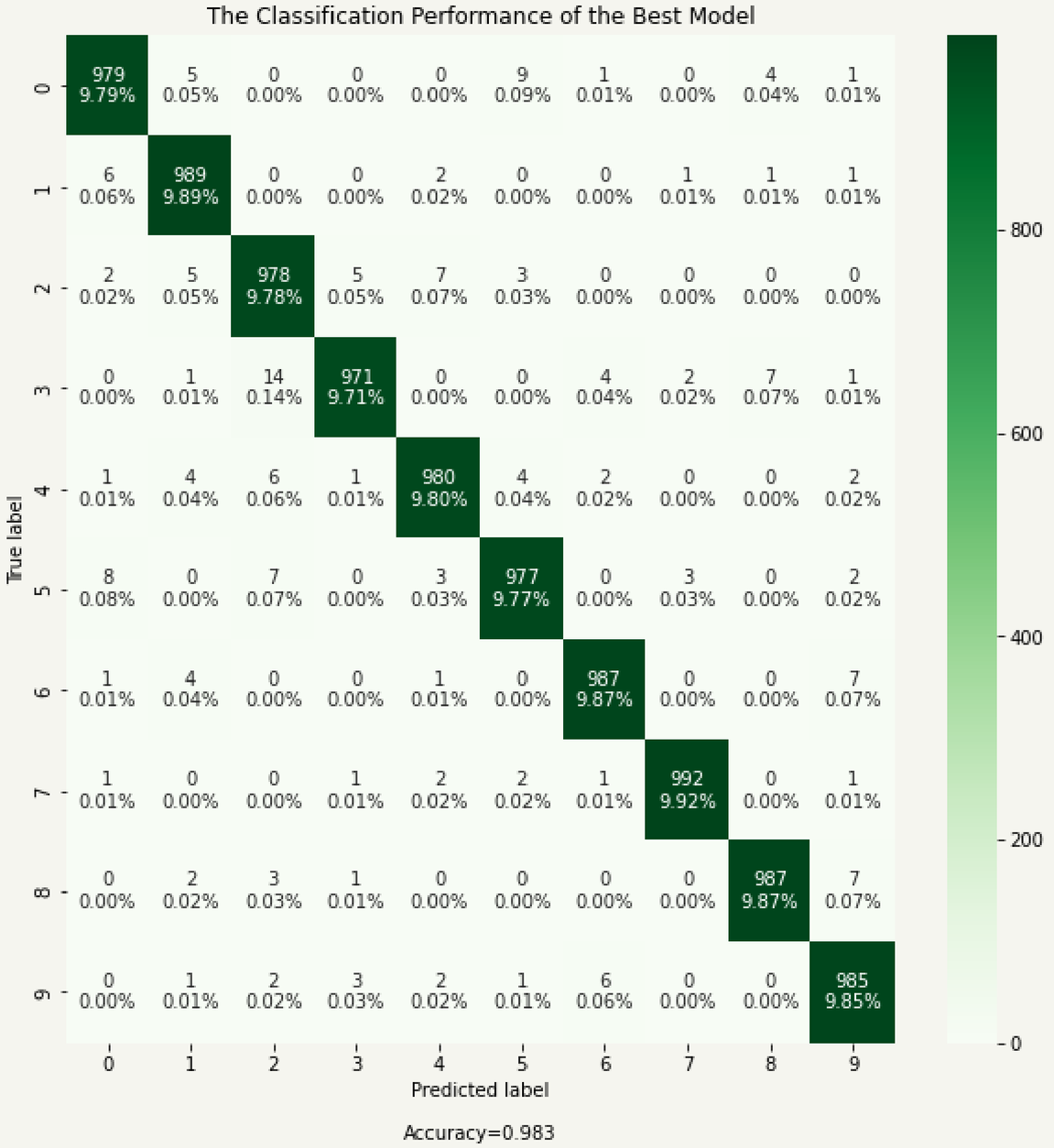
Website

- Straemlit

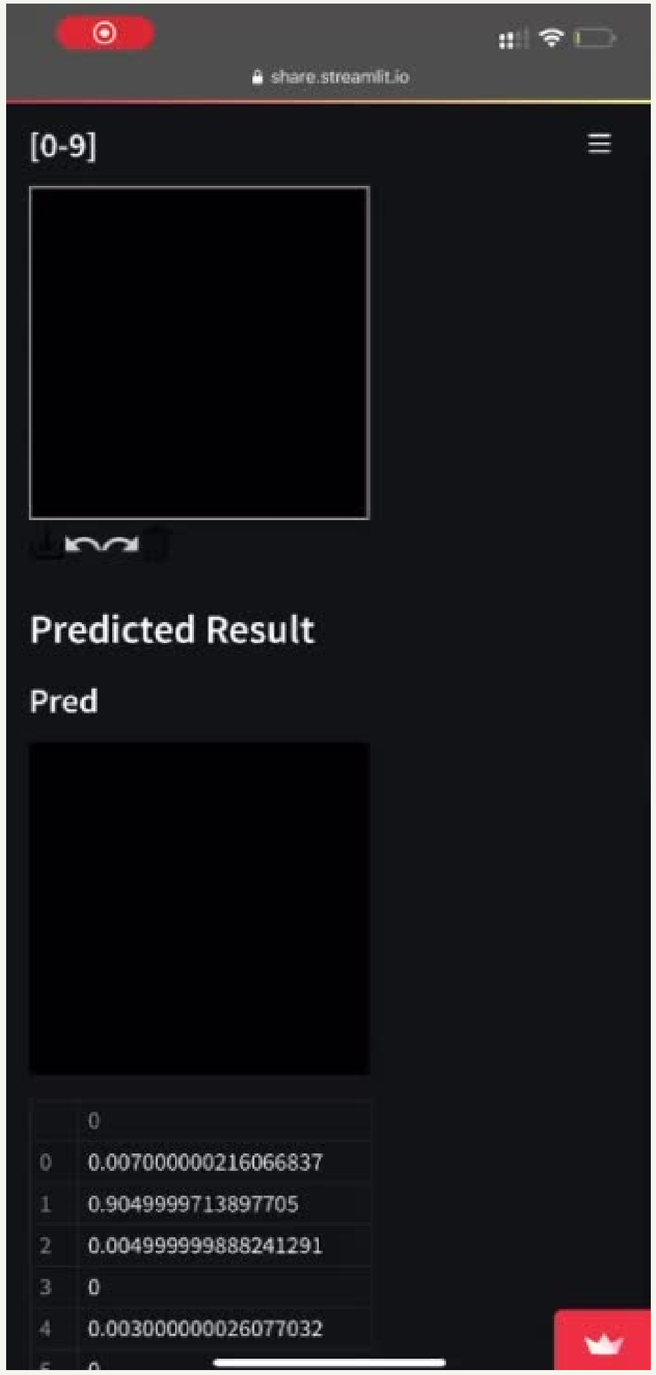
Models


Models' F1 Score for 10-Fold Cross Validation
F1-Score
0.99
0.98
0.97
0.96
0.95
XGB
KNN
SVM
Models

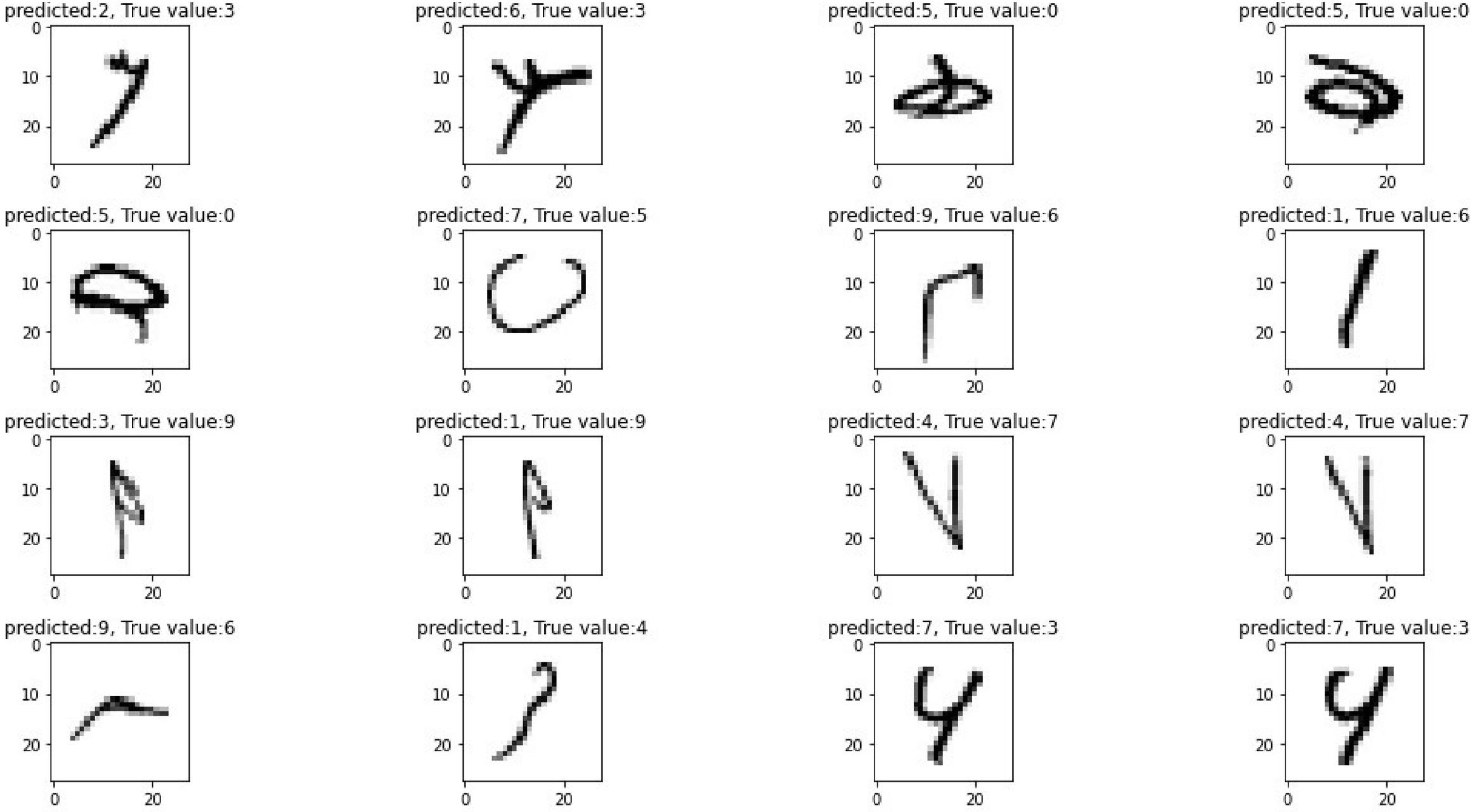
Best Model


The Classification Performance of the Best Model
True label
Predicted label
Accuracy=0.983
0 1 2 3 4 5 6 7 8 9
979 9.79% | 5 0.05% | 0 0.00% | 0 0.00% | 0 0.00% | 9 0.09% | 1 0.01% | 0 0.00% | 4 0.04% | 1 0.01%
6 0.06% | 989 9.89% | 0 0.00% | 0 0.00% | 2 0.02% | 0 0.00% | 0 0.00% | 1 0.01% | 1 0.01% | 1 0.01%
2 0.02% | 5 0.05% | 978 9.78% | 5 0.05% | 7 0.07% | 3 0.03% | 0 0.00% | 0 0.00% | 0 0.00% | 0 0.00%
0 0.00% | 1 0.01% | 14 0.14% | 971 9.71% | 0 0.00% | 0 0.00% | 4 0.04% | 2 0.02% | 7 0.07% | 1 0.01%
1 0.01% | 4 0.04% | 6 0.06% | 1 0.01% | 980 9.80% | 4 0.04% | 2 0.02% | 0 0.00% | 0 0.00% | 2 0.02%
8 0.08% | 0 0.00% | 7 0.07% | 0 0.00% | 3 0.03% | 977 9.77% | 0 0.00% | 3 0.03% | 0 0.00% | 2 0.02%
1 0.01% | 4 0.04% | 0 0.00% | 0 0.00% | 1 0.01% | 0 0.00% | 987 9.87% | 0 0.00% | 0 0.00% | 7 0.07%
1 0.01% | 0 0.00% | 0 0.00% | 1 0.01% | 2 0.02% | 2 0.02% | 1 0.01% | 992 9.92% | 0 0.00% | 1 0.01%
0 0.00% | 2 0.02% | 3 0.03% | 1 0.01% | 0 0.00% | 0 0.00% | 0 0.00% | 0 0.00% | 987 9.87% | 7 0.07%
0 0.00% | 1 0.01% | 2 0.02% | 3 0.03% | 2 0.02% | 1 0.01% | 6 0.06% | 0 0.00% | 0 0.00% | 985 9.85%
800
600
400
200
0

Demo


share.streamlit.io
[0-9]
Predicted Result
Pred
0
0 0.007000000216066837
1 0.9049999713897705
2 0.004999999888241291
3 0
4 0.003000000026077032

Total misclassified images by XGB# 174 out of 10000. Some examples:



Conclusion

- Supervised classification algorithms performs well for handwritten image classification problems.
- Image classification requires heavy computational power and long training hours.

Future Work

- Improve the model's accuracy.
- Solve for the outlier cases e.g. misclassifying the Arabic digits 0 and 5.
- Train the model to recognize Arabic letters and segment words.
- Add a new feature in the interface to accept uploaded images.

Thank You For Listening