Marks2CSV

A simple solution to convert tabular mark fields to CSV file

Mini Project Presentation: First Review

Guided by:

Dr. Deepa V.

Presented by:

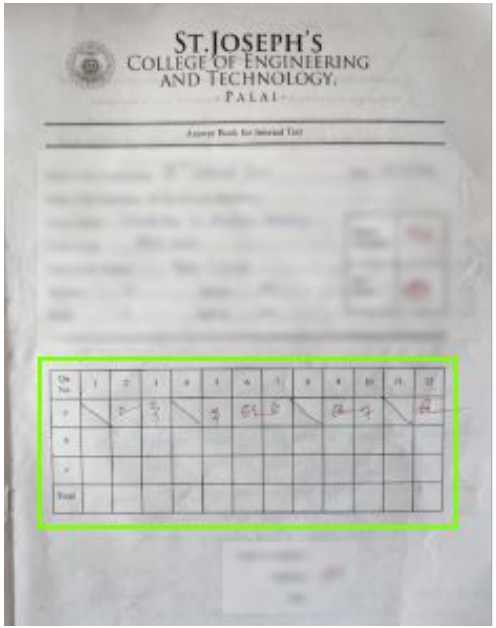
Ajay T Shaju, SJC20AD004
Emil Saj Abraham, SJC20AD028
Justin Thomas Jo, SJC20AD046
Vishnuprasad K G, SJC20AD063

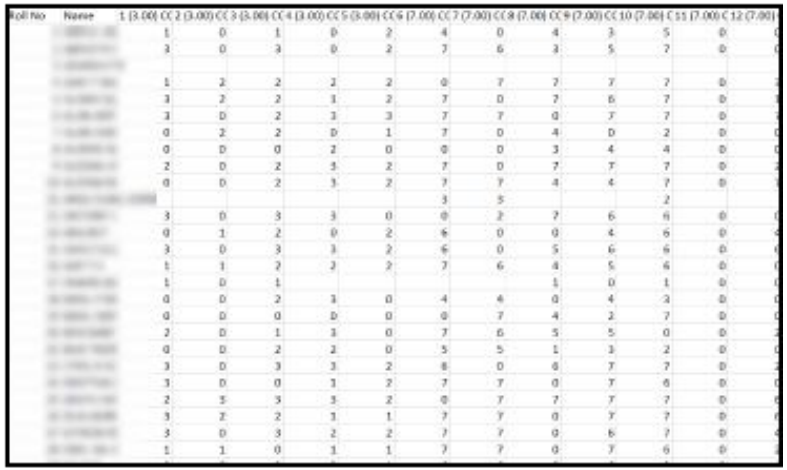
Outline

- Introduction
- Problem Statement
- Application
- Literature Survey
- Block Diagram
- Data Collection
- Work Done So Far
- Performance Evaluation
- Work To Be Done
- Conclusion
- References

Introduction

- Idea: Answer sheet **marks to CSV** File.
- Saves time and resources.
- Reduce errors.





Problem Statement

- Manual data entry is a **labor-intensive process** that requires significant time and effort.
  - Prone to errors
  - Inaccurate data
  - Loss of valuable time
- This project aims to develop an **automated solution** that streamlines the aforementioned problems.

Application

- **Detection of handwritten marks** from images and their **conversion into CSV** format.
- **Simplify the mark data entry process** by **avoiding manual input** of marks on each cell.
- Final **output as a CSV file** comprising all numbers extracted from the input images.
- **Automated generation** of mark list **saves time.**
- Designed **specifically for SJCET** Teachers.

Literature Review

[1] A.Raj, S.Sharma, J.Singh, A.Singh, “Revolutionizing Data Entry: An In-Depth Study of Optical Character Recognition Technology and Its Future Potential”, International Journal for Research in Applied Science & Engineering Technology, Vol. 11 No.2, pp: 645-653, Feb 2023.

- OCR,artificial intelligence,document scanning,machine learning,image recognition.
- Increased speed and efficiency, improved accuracy, reduced costs and increased accessibility.
- Gap - recognition accuracy of complex data structures is poor.
- Future scope - impact of OCR increases as technology advances, thereby giving more importance and emphasis to using it in businesses and organizations.

[2] Ömer Aydin, “Classification of Documents Extracted from Images with Optical Character Recognition Methods”, Anatolian Journal of Computer Sciences, Vol.6 No.2 pp:46-55, 01 Jun, 2021.

- OCR classification and image processing with use of Naive Bayes algorithm.
- Identification of text from handwritten documents, extracting features and training them.
- Gap - accuracy is only 53%, lack of implementation of better model.
- Future scope - apply same method with a neural network.

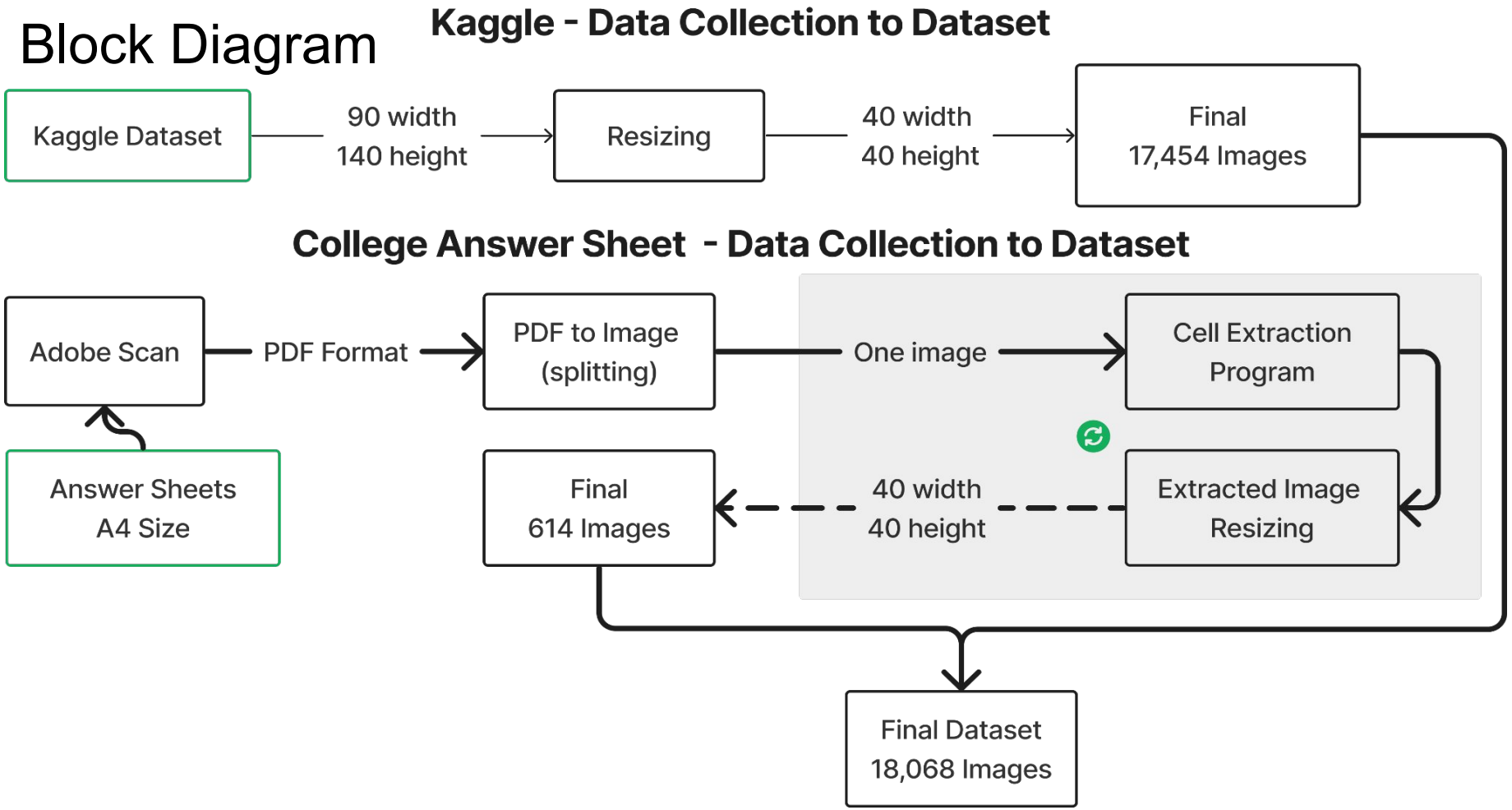
[3] Raajkumar G., Indumathi D., “Optical Character Recognition using Deep Neural Network”, International Journal of Computer Applications, Vol. 176 No. 41 pp:61-65, July 2020.

- Image processing, OCR model, long short term memory.
- Related work - text and image segmentation, CNN.
- Gap - cannot identify text set at a particular angle.
- Future scope - implies more usage of PyTesseract over SVM.

[4] J.Memon, R.Sami, Rizwan A.Khan, M.Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)”, IEEE Access, Vol. 8, pp:142642-142668, 2020.

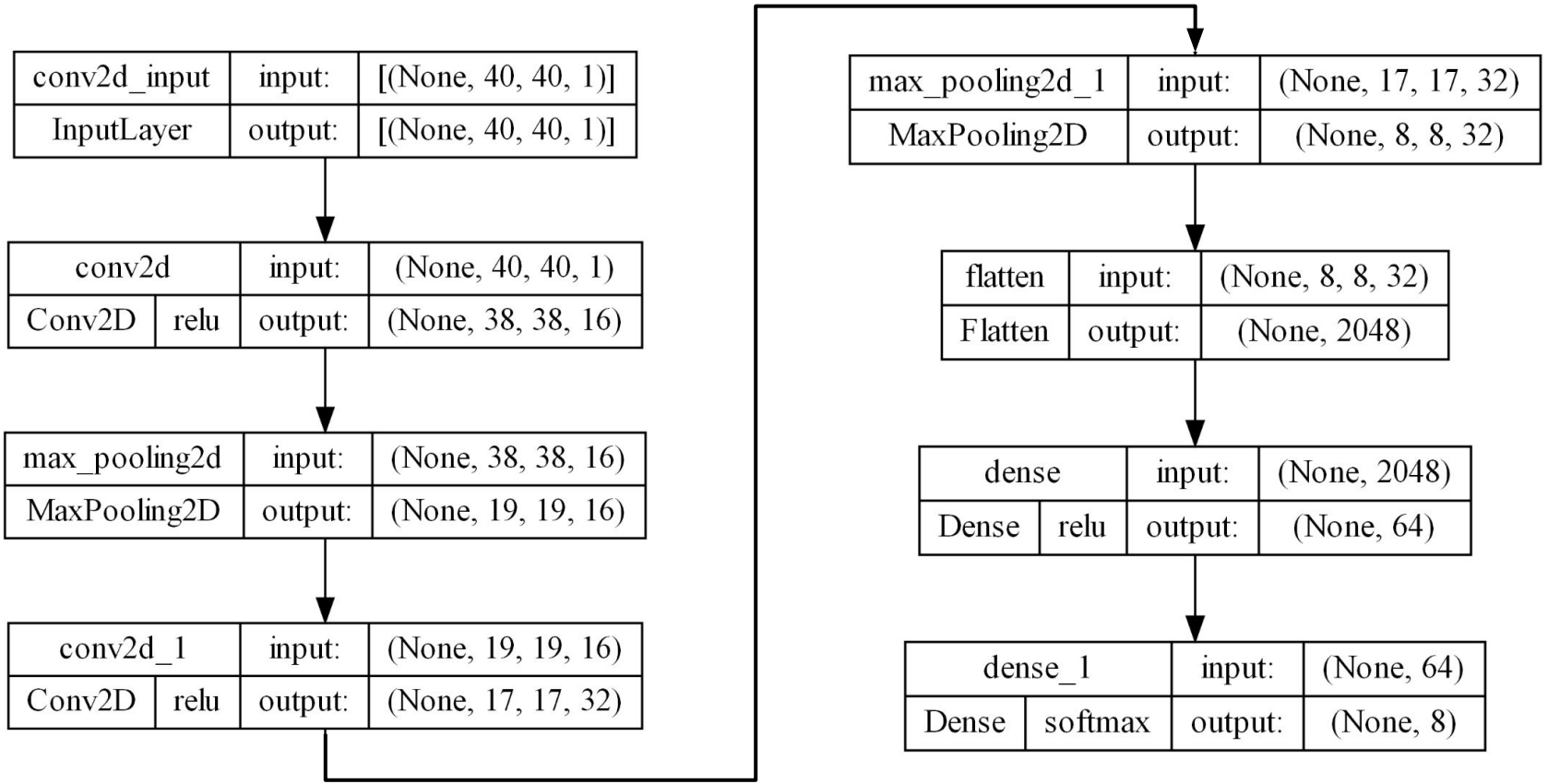
- Optical character recognition,classification,languages,feature extraction,deep learning.
- Implementation of MLP, use of datasets like CEDAR,MNIST,CHARS74K.
- Gap - Publicly available datasets also include stimuli that are aligned well with each other and fail to incorporate examples that correspond well with real-life scenarios, i.e. writing styles, distorted strokes, variable character,thickness and illumination.
- Future scope - implementation of deep learning architectures like CNN,RNN and LSTM will steadily increase.

Block Diagram

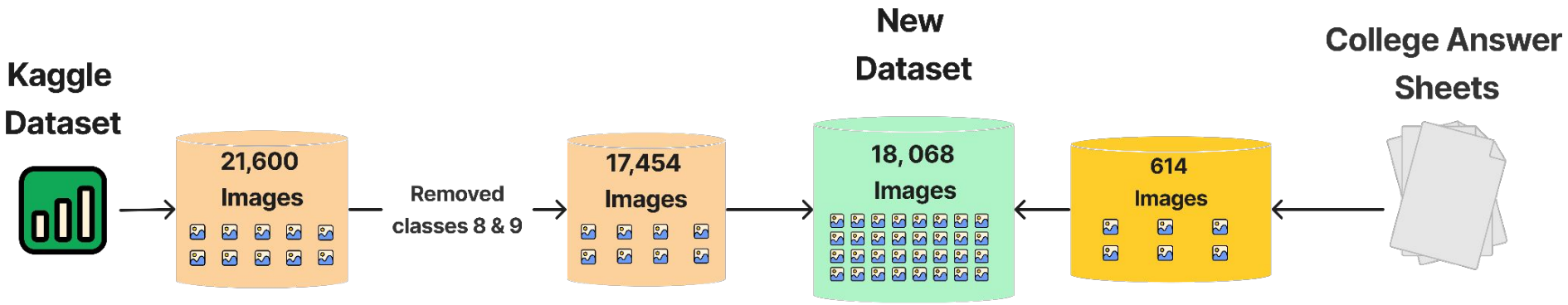


Block Diagram

Neural Network Architecture

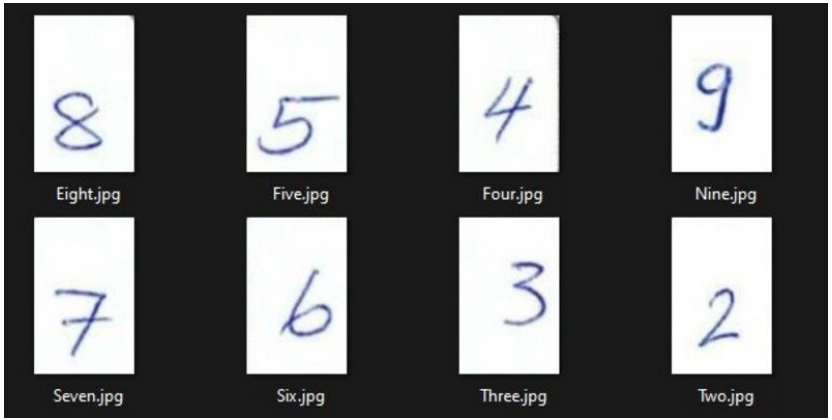


Data Collection

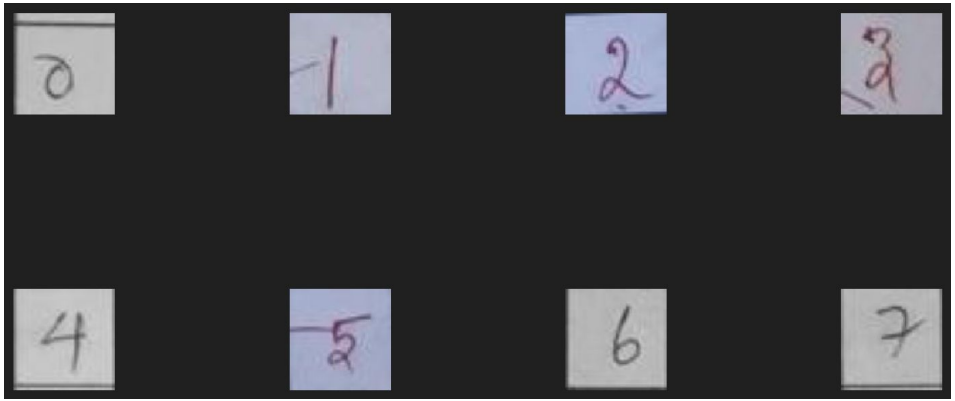


Work Done So Far

- Used a library '**img2table**' for detecting and extracting cells.
- Created dataset using an **online handwritten dataset** & **images from answer sheets**.



Online Dataset Header Image



Answer Sheet Dataset Sample Images

- Mistakes in **Paddle OCR** predictions.
- Created **neural network for number detecting** (Custom OCR tool).

1a	1b	2a	3a	3c	4a	5a	6a	7a	7c	8a	9a	9b	10a	11c	12a
			2												
		.			dr					2	44		51		
22	22			25					65	65	7	7		6X	6X
		13				3	53	4			6		6		6
					2	23					3				
			M					6					7		
					3	3							63		

Output of Paddle OCR tool

Performance Evaluation

```
Testing accuracy: 95.81%
```

Neural Network Accuracy

```
Total completion time: 14.81 minutes
```

Time taken for Prediction

```
Accuracy of 0 : 93.97%
Accuracy of 1 : 99.29%
Accuracy of 2 : 96.22%
Accuracy of 3 : 95.10%
Accuracy of 4 : 97.59%
Accuracy of 5 : 93.33%
Accuracy of 6 : 97.33%
Accuracy of 7 : 97.57%
```

Prediction Accuracy

Work To Be Done

- Collect more datasets
- Analyze model performance using additional methods
- Integrate custom OCR tool
- Preprocess DataFrame
- Physical equipment for holding the camera

Conclusion

- Dataset consist of **18,068 Images**
- Testing **accuracy** of CNN is **95.81%.**
- **Custom OCR tool** has achieved **greater accuracy than previous OCR tool**(Paddle OCR).

References

[1] A.Raj, S.Sharma, J.Singh, A.Singh, “Revolutionizing Data Entry: An In-Depth Study of Optical Character Recognition Technology and Its Future Potential”, International Journal for Research in Applied Science & Engineering Technology, Vol. 11 No.2, pp: 645-653, Feb 2023.

[2] Ömer Aydin, “Classification of Documents Extracted from Images with Optical Character Recognition Methods”, Anatolian Journal of Computer Sciences, Vol.6 No.2 pp:46-55, 01 Jun, 2021.

[3] Raajkumar G., Indumathi D., “Optical Character Recognition using Deep Neural Network”, International Journal of Computer Applications, Vol. 176 No. 41 pp:61-65, July 2020.

[4] J.Memon, R.Sami, Rizwan A.Khan, M.Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)”, IEEE Access, Vol. 8, pp:142642-142668, 2020.

[5] Colin G.White-Dzuro, Jacob D.Schultz, C.Ye,Joseph R. Coco, Janet M. Myers, C.Shackelford, S.T.Rosenbloom,D.Fabbri, “Extracting Medical Information from Paper COVID-19 Assessment Forms”, Applied Clinical Informatics Vol. 12 No. 1, pp:170–178, 2021.

Questions?

Thank You

Presentation Setting

Introduction – Ajay

Problem statement – Ajay

Application – Emil

Literature Survey – Justin

Block Diagram – Justin

Data Collection – Vishnu

work done so far – Ajay

performance evaluation – justin

Work To Be Done– Emil

Conclusion – Vishnu

References – Vishnu

Don’t Present this

Don’t Present this

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_4 (Conv2D)           (None, 88, 138, 16)       160

 max_pooling2d_4 (MaxPooling  (None, 44, 69, 16)       0
 2D)

 conv2d_5 (Conv2D)           (None, 42, 67, 32)        4640

 max_pooling2d_5 (MaxPooling  (None, 21, 33, 32)       0
 2D)

 flatten_2 (Flatten)         (None, 22176)             0

 dense_4 (Dense)             (None, 64)                1419328

 dense_5 (Dense)             (None, 8)                 520

=================================================================
Total params: 1,424,648
Trainable params: 1,424,648
Non-trainable params: 0
```

Custom OCR Model Summary

Conclusion

- Ensure that the predictions are accurate need to improve the accuracy further.
- need to integrate the table detection algorithm & a custom OCR model.
- aim to reduce the time taken by teachers to do the mark data entry procedure.
- Preliminary results has yielded great success but can still be improved.
- Validation accuracy of CNN is 95.81%.

Data Description

Excluded
May be useful in future

- Primary data are ‘Answer sheets of our college’, which are image data of A4 size
- The data comes under ‘college documents’ domain
- Initial data can be collected from our department, for more data variability, we could request it from other departments of our college itself.
- Data can be compressed to file format like PDF for easy portability
- Data size: Input data can have 5-10MB (PDF), and output data can have 4KB - 2MB (CSV)
- Data may be unstructured and papers may be damaged or faded.