

Marks2CSV

A simple solution to convert tabular mark fields to CSV file

Mini Project Presentation: Second Review

Guided by:
Dr. Deepa V.

Presented by:

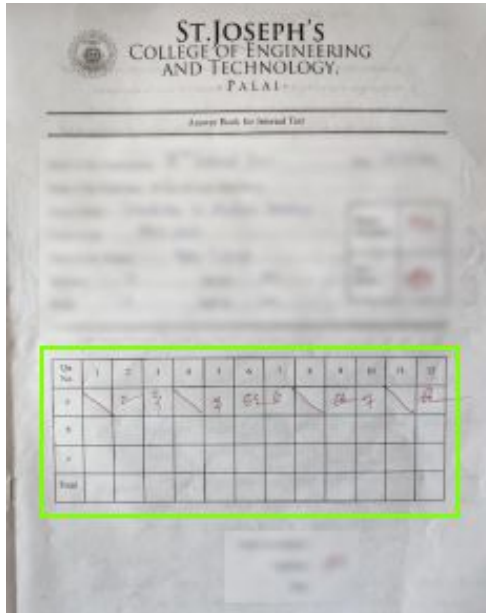
Ajay T Shaju,	SJC20AD004
Emil Saj Abraham,	SJC20AD028
Justin Thomas Jo,	SJC20AD046
Vishnuprasad K G,	SJC20AD063

Outline

- Introduction
- Literature Survey
- Problem Statement
- Research Scope and Objectives
- Application
- Methodology
 - Data Collection
 - Block Diagram
 - Techniques
- Results and Discussion
- Conclusion and Future Scope
- References

Introduction

- Idea: Instant conversion of answer sheet data to CSV files.
- Saves time with its fast performance.
- Easy to use.



Roll No.	Name	1 (3.00)	CC 2 (3.00)	CC 3 (3.00)	CC 4 (3.00)	CC 5 (3.00)	CC 6 (7.00)	CC 7 (7.00)	CC 8 (7.00)	CC 9 (7.00)	CC 10 (7.00)	CC 11 (7.00)	CC 12 (7.00)
1		0	1	0	2	4	0	4	3	5	0		
2		0	3	0	2	7	6	3	5	7	0		
3		2	2	2	2	0	7	7	7	7	0		
4		2	2	1	2	7	0	7	6	7	0		
5		0	2	3	3	7	7	0	7	7	0		
6		0	2	0	1	7	0	4	0	2	0		
7		0	0	2	0	0	0	3	4	4	0		
8		2	0	2	3	2	7	0	7	7	0		
9		0	0	2	3	2	7	7	4	4	7	0	
10		3	0	3	3	0	0	2	7	6	0		
11		0	1	2	0	2	6	0	0	4	6	0	
12		3	0	3	3	2	6	0	5	6	6	0	
13		1	1	2	2	2	7	6	4	5	4	0	
14		1	0	1					1	0	1	0	
15		0	0	2	3	0	4	4	0	4	3	0	
16		0	0	0	0	0	0	7	4	2	7	0	
17		2	0	1	3	0	7	6	5	5	0	0	
18		0	0	2	3	0	5	5	1	3	2	0	
19		3	0	3	3	2	8	0	6	7	7	0	
20		3	0	0	1	2	7	7	0	7	6	0	
21		2	3	3	3	2	0	7	7	7	7	0	
22		3	2	2	1	1	7	7	0	7	7	0	
23		3	0	3	2	2	7	7	0	6	7	0	
24		1	1	0	1	1	7	7	0	7	6	0	

Literature Survey

[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.

Problem Statement

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

Research Scope and Objectives

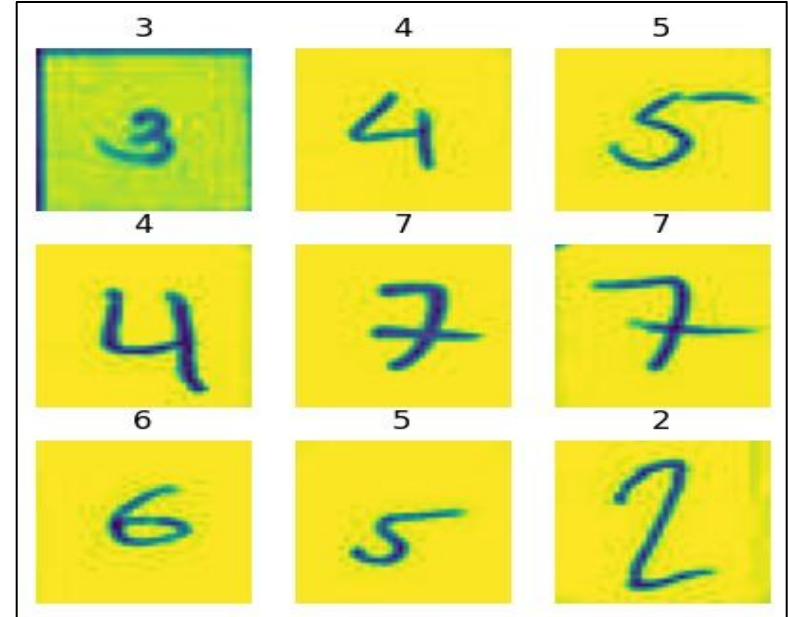
- Scope:
 - **Automated system** to make CSV out of images of answer sheets.
 - Use **neural networks to clearly classify the digits** to attain optimal results.
 - To **detect decimal marks** during the OCR processing.
 - Train the model to **detect marks as per customer** requirements.
 - Create a **smaller device to encase the tool**.
- Objective:
 - Create a tool to **help teachers in data entry process**.
 - Implementation of **OCR** using the **CNN Architecture**.
 - A **customizable program** to extract table marks with **programmer's control**.

Application

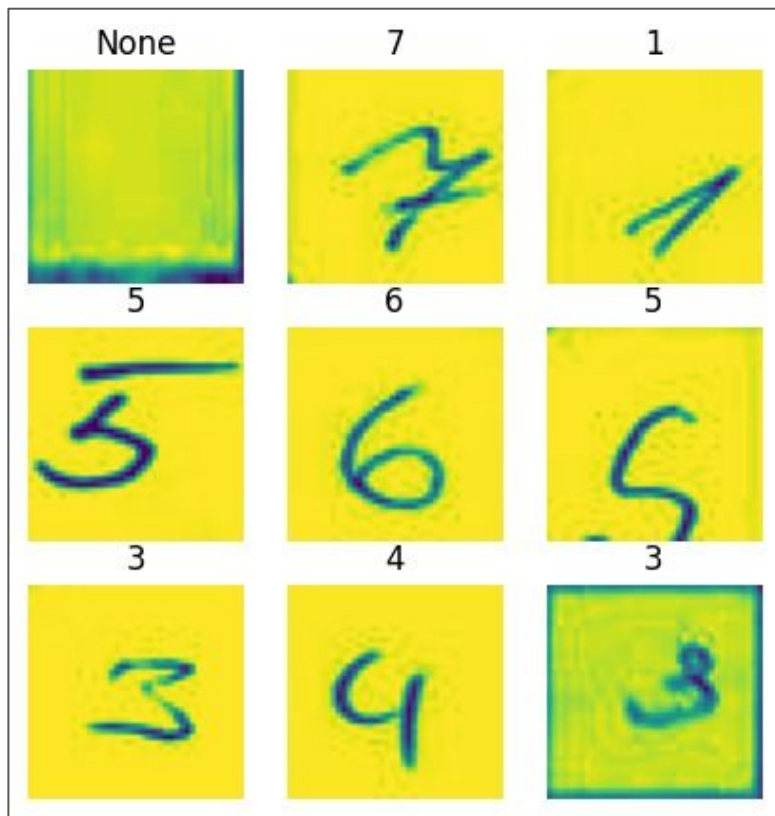
- **Detection of handwritten marks** from images and their **conversion to CSV** format.
- **Simplify the mark entry process** by **avoiding manual input** on each cell.
- Final **output obtained as CSV file** comprising all numbers extracted from the input images.
- Designed specifically for **SJCET teachers**.

Methodology - Data Collection

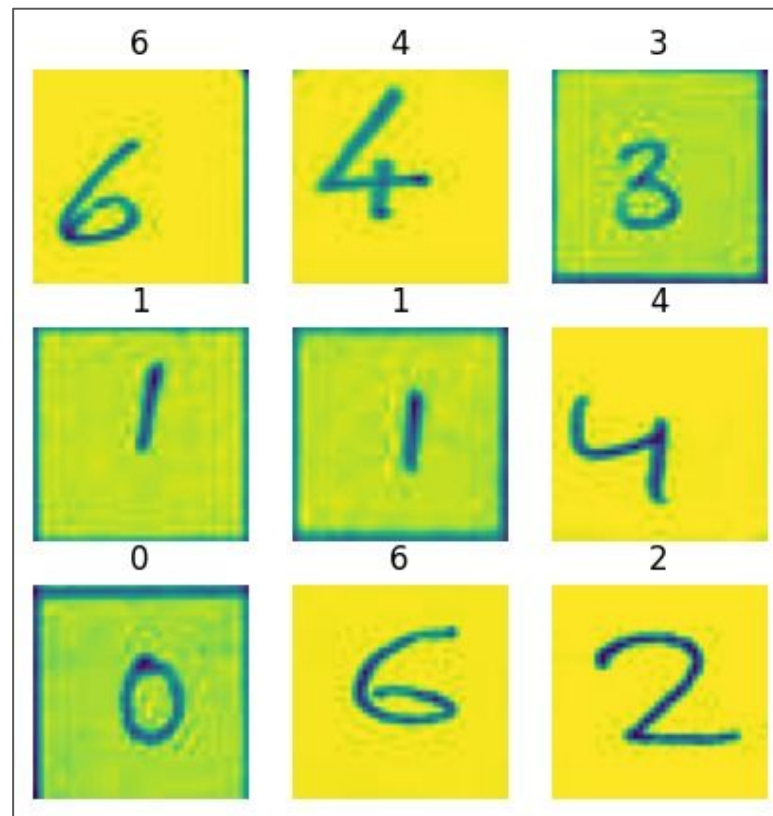
- Main data sources -
College answer sheets (7,420) and
Kaggle number dataset (13, 855)
- Data Type - Image data of JPG format
- Data Size - 40px x 40px x 3 channel
- Data Storage -
Windows File System & Zip Formats



Training Set



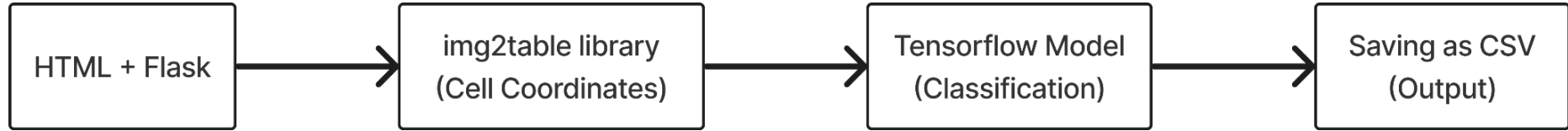
Validation Set



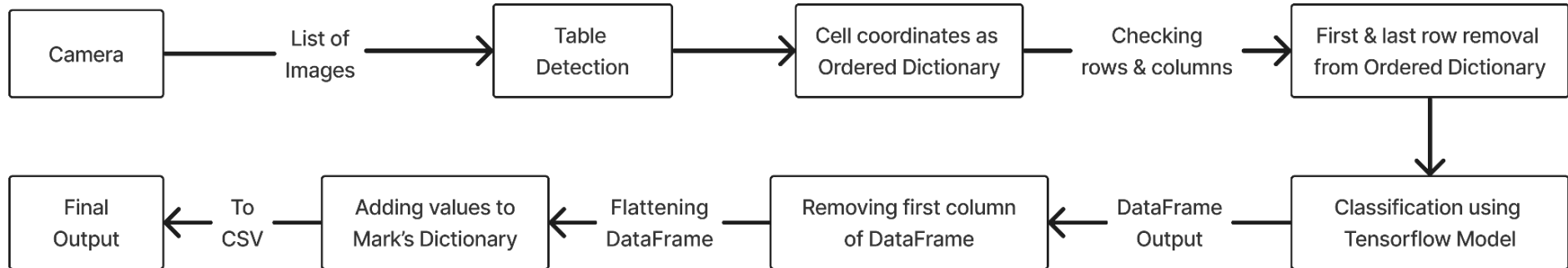
Testing Set

Methodology - Block Diagram

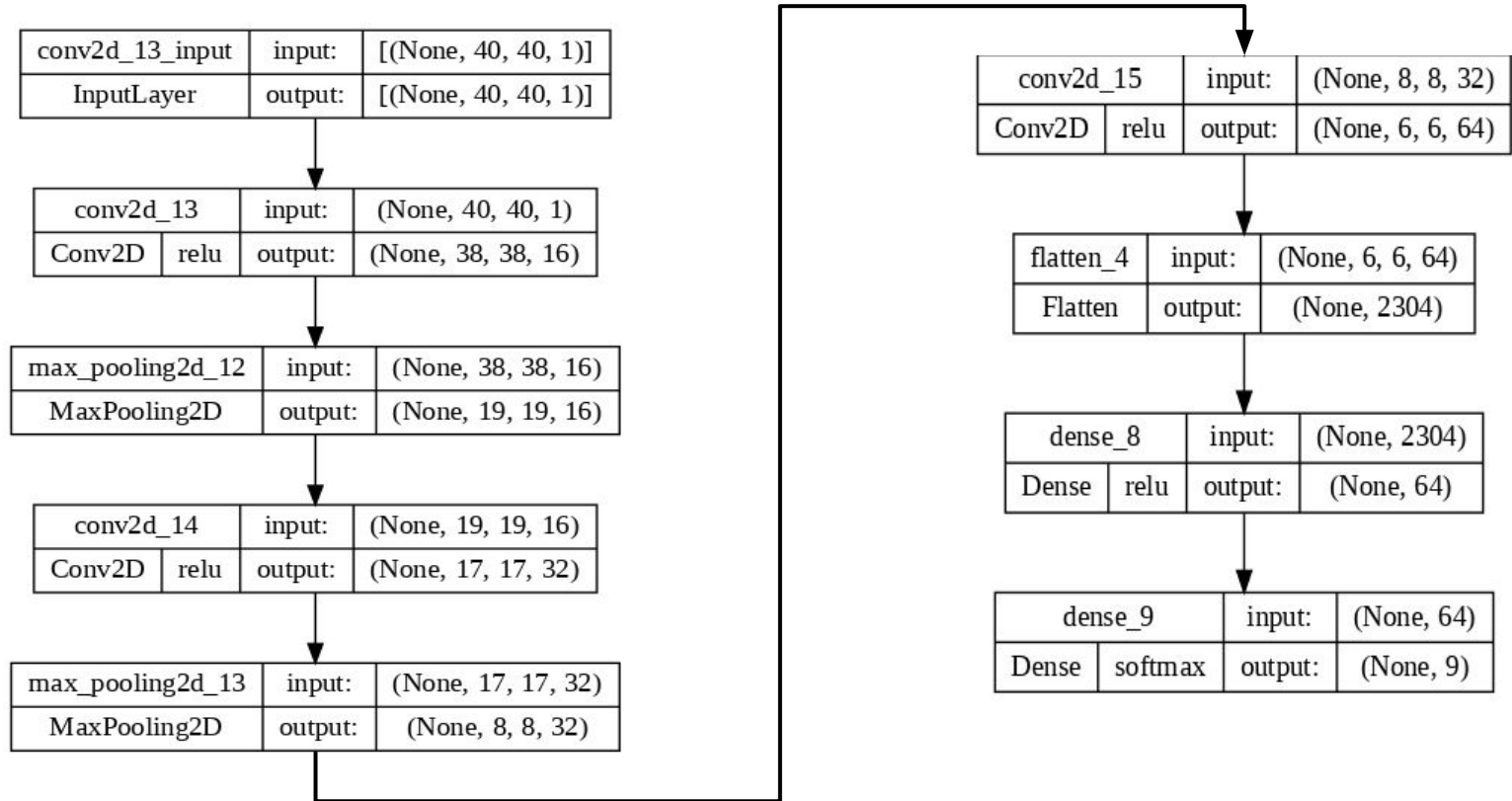
Our Work in 4 Steps



Working of System



Neural Network Architecture

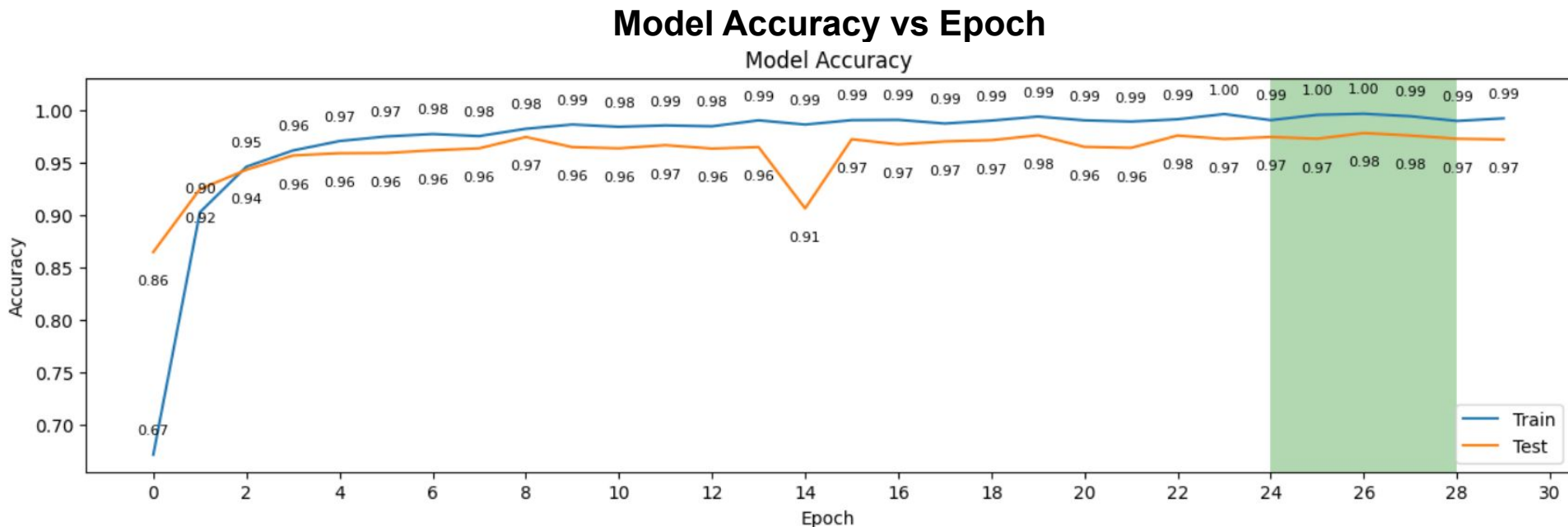


Methodology - Techniques

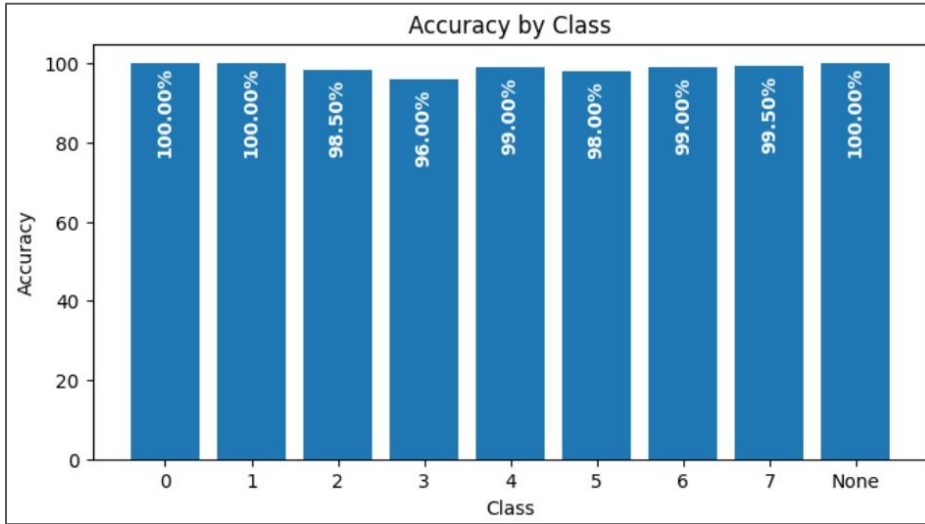
- Increased the speed of processing from **10 minutes to 3 minutes**
 - PaddleOCR takes 10 minutes to complete whole process.
 - New processing is done on **built-in data structure** (ordered dictionary).
 - First & last row are removed by **deleting the keys** of ordered dictionary.
 - We **run OCR** only on remaining **39 cells**.
- Using **custom trained OCR** tool supported by a **CNN architecture** for classification.

Results and Discussion

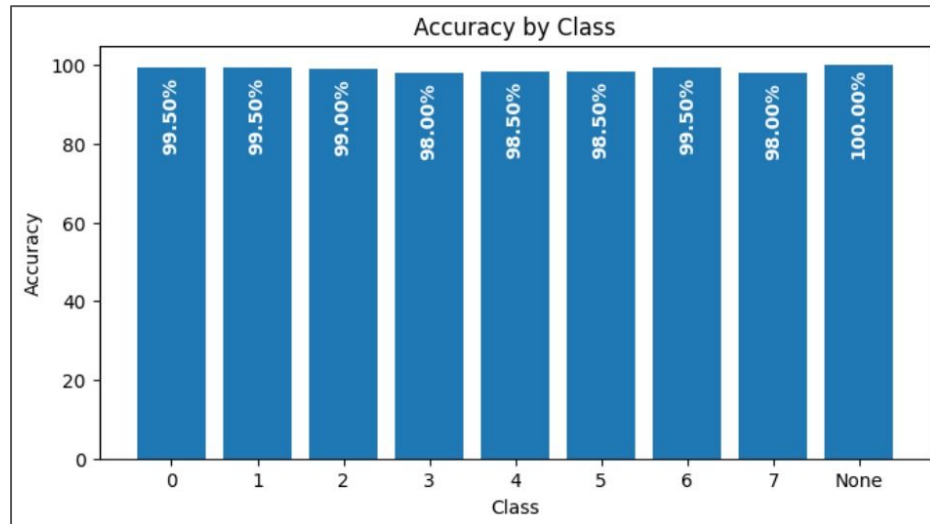
- As epoch value increases, we see training and testing accuracies gradually increasing.
- Model performs best when the epoch value is within the shaded region.



Accuracy by Class



CNN_Model_0

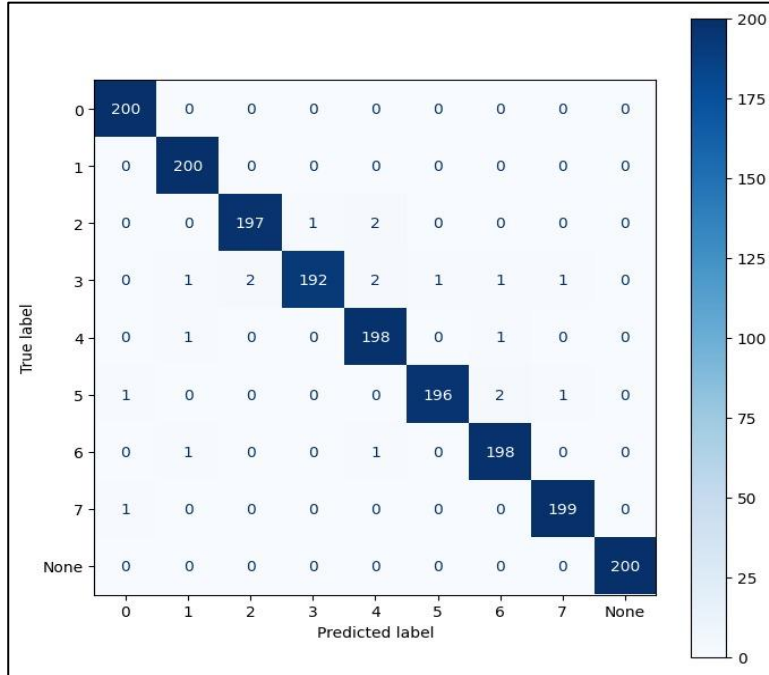


CNN_Model_1

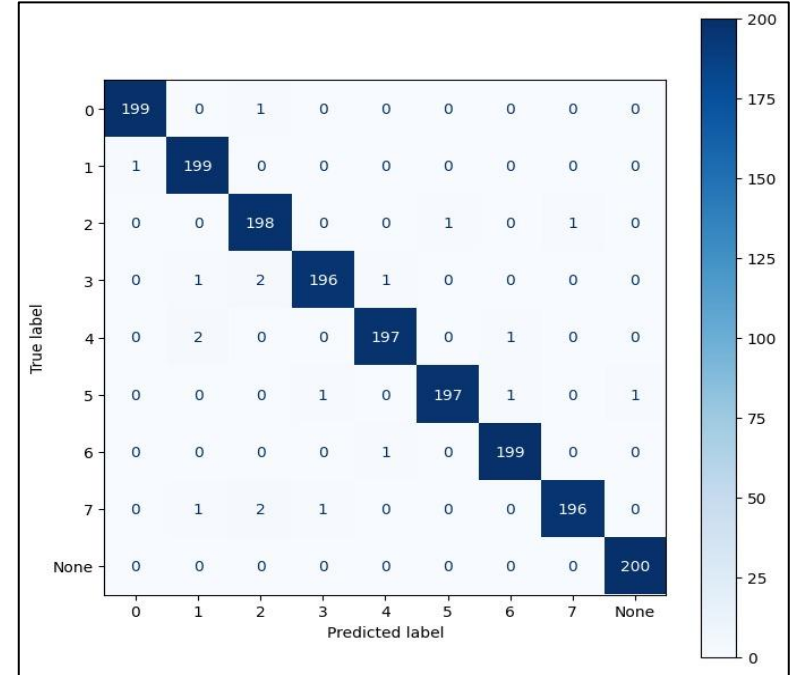
- Comparison of detection accuracy of each class for CNN_Model_0 and CNN_Model_1.
- Accuracy values improve in CNN_Model_1 over values of CNN_Model_0.

Comparison of Confusion Matrices

- Notice the TP value changing for certain numbers in both CNN_Model_0 and CNN_Model_1.



CNN_Model_0



CNN_Model_1

Performance Evaluation

	0	1	2	3	4	5	6	7	None
	CNN_Model_0								
Precision	0.99	0.99	0.99	0.99	0.98	0.99	0.98	0.99	1.00
Recall	1.00	1.00	0.98	0.96	0.99	0.98	0.99	0.99	1.00
F1-Score	1.00	0.99	0.99	0.98	0.98	0.99	0.99	0.99	1.00
	CNN_Model_1								
Precision	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.99	1.00
Recall	0.99	0.99	0.99	0.98	0.98	0.98	0.99	0.98	1.00
F1-Score	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99	1.00

Overall the **CNN_Model_1** has equal performance for all classes

Conclusion and Future Scope

- We have achieved our objective of designing an OCR tool for detecting handwritten marks that helps for data entry process.
- We have successfully implemented a program for **real-time processing** to make the tool more user friendly and simpler.
- CNN_Model_1 has **training, validation & testing accuracy of 99.37%, 97.11% & 99.2%** respectively compared to accuracy values of **99.79%, 96.80% & 99%** for CNN_Model_0.

- **Limitations:**

- Difficulty in **detecting decimal part** of numbers like 0.5.
- OCR tool's efficiency is affected by **stray marks, corrections, and unsteady cell lines.**

- **Future scope:**

- Improve model architecture to work on **decimal marks, different styles of writing, and to detect stray marks accurately.**
- **Getting user inputs** prior to starting recognition.

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.

[6] S.Shrivatsava, Sanjeev K.Singh, K.Shrivatsava, V.Sharma, “CNN based Automated Vehicle Registration Number Plate Recognition System”, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE Xplore, pp: 795-802, 01 March 2021.

[7] A.Das, Gyana R.Patra, Mihir N.Mohanty, “LSTM based Odia Handwritten Numeral Recognition”, 2020 International Conference on Communication and Signal Processing (ICCSP), IEEE Xplore, pp: 538-541, 01 September 2020.

[8] B.Barz, J.Denzler, “Deep Learning on Small Datasets without Pre-Training using Cosine Loss”, 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE Xplore, pp: 1360-1369, 14 May 2020.

[9] A.Yaganteeswarudu, “Multi Disease Prediction Model by using Machine Learning and Flask API”, 2020 5th International Conference on Communication and Electronics Systems (ICCES), IEEE Xplore, pp: 1242-1246, 10 July 2020.

[10] A. Arivoli, D.Golwala, R.Reddy, “CoviExpert: COVID-19 detection from chest X-ray using CNN”, Journal of Measurement: Sensors 23, pp: 1-8, 2022.

Questions?

Thank You

Presentation Setting

Introduction - Justin

Literature Survey - Justin

Problem Statement -Ajay

Research Scope & Objectives - Vishnu

Application - Emil

Methodology

--> Data Collection - Ajay

--> Block Diagram - Emil (BD), Justin (NN)

--> Techniques - Vishnu

Results - Justin (Epoch, Acc), Vishnu (CM,Report)

Conclusion - Ajay

References - Ajay

Don't Present this

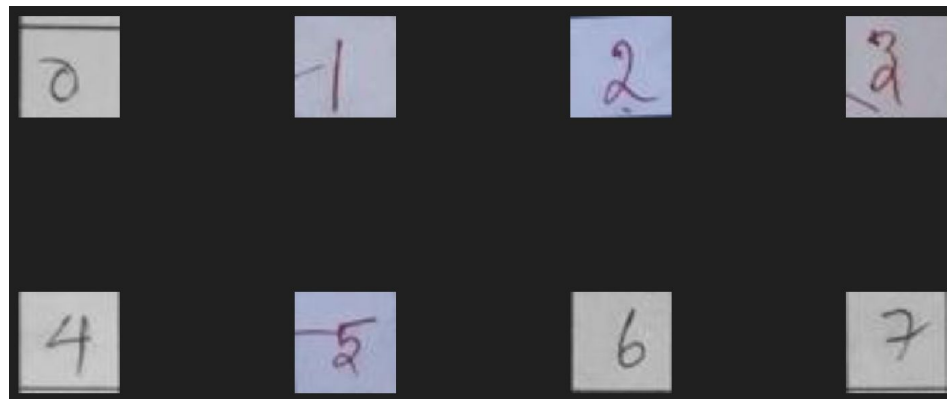
Don't Present this

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

Model: "sequential_2"

Layer (type)	Output Shape	Param #
=====		
conv2d_4 (Conv2D)	(None, 88, 138, 16)	160
max_pooling2d_4 (MaxPooling 2D)	(None, 44, 69, 16)	0
conv2d_5 (Conv2D)	(None, 42, 67, 32)	4640
max_pooling2d_5 (MaxPooling 2D)	(None, 21, 33, 32)	0
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.

Outline

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

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

- 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

Research Scope and Objective

- To detect decimal point marks during the OCR processing.
- To clearly identify the digits properly to attain optimal results.
- To train the model more to detect marks greater than 7 if the customer's input is so.
- Create a smaller device to encase the tool.
- Objective: Create a tool to help ease the teacher's efforts in data entry process, by implementing OCR techniques with the help of CNN.

Techniques

- User interface created in **Flask & HTML**.
- **Real time capturing** of the answer sheets.
- **img2table** library for **coordinate extraction**.
- **Processing on built-in data structure** (ordered dictionary) for speed.
- Using **custom trained OCR** tool supported by a **CNN architecture** for classification.

Classification Report

Precision: How accurate the models are in predicting positive samples.

Recall: How effectively a model can identify positive samples.

F1-Score: The average of precision and recall, providing a balanced measure of a model's performance.