Marks2CSV

A simple solution to convert tabular mark fields to CSV file

Mini Project Presentation

Guided by:
Dr. Deepa V.

Presented by:

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

Outline

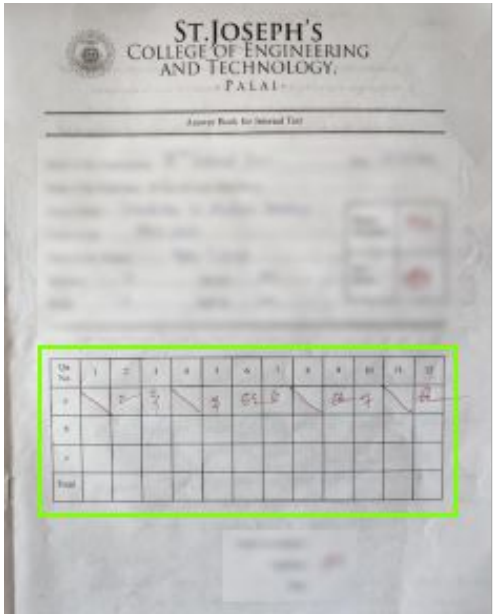
- Introduction
- Literature Survey
- Problem Statement
- Research Scope and Objectives
- Application
- Methodology
  - Data Collection
  - Block Diagram
  - Techniques
- Results and Discussion
- Conclusion and Future Scope
- References

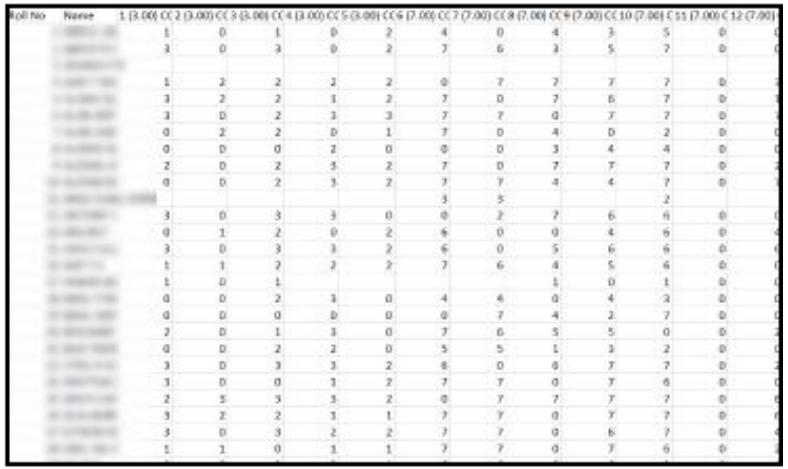
Introduction

"Technology will never replace great teachers, but technology in the hands of great teachers can be transformational." - The Innovator's Mindset, George Couros.

- **Digitizing handwritten data is a necessity in modern age**, to analyse and get insights.
- **Educational Institutions** also has **huge requirement** of data digitization.
- Introducing our **time-saving idea of digitizing handwritten documents to automate various data entry processes**.

- The key technology is **Optical Character Recognition (OCR)**.
- **Optical Character Recognition**, is a technology that **enables the conversion of scanned or photographed images of text into digital text** that can be edited, and searched by a computer. OCR is commonly **used to digitize handwritten or paper-based documents.**
- Instant conversion of **answer sheet data to CSV files.**
- **Saves time** with **fast performance**.





Conversion of mark table in answer sheets to CSV File

Literature Survey

[1] C.ShanWei, S.LiWang, Ng T.Foo, Dzati A.Ramli, “A CNN based Handwritten Numeral Recognition Model for Four Arithmetic Operations 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems”, Procedia Computer Science, Vol.192, pp:4416-4424, 2021.

- AI, CNN, SVM, Softmax Classifier, ReLU, ADAM.
- Skew image correction, image segmentation, training data acquisition, algorithm improvement.
- Gap - use of MNIST dataset instead of sample images from user’s input, thus only works on handwritten numerals.
- Future scope - Extend the potential of CNN in recognizing handwritten English letters and Chinese characters.

[2] 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.

[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 if not set at a particular angle.
- Future scope - implies more usage of PyTesseract over SVM.

[4] J.Yuan, H.Li, M.Wang, R.Liu, C.Li, B.Wang, “An OpenCV-based Framework for Table Information Extraction”, 2020 IEEE International Conference on Knowledge Graph (ICKG), IEEE Xplore, pp: 621-628, Sep 2020.

- PDF, Information Extraction, OpenCV, TesseractOCR.
- PDF Preprocessing, Table Detection, Table Cell Value Extraction.
- Gap - Fails to extract table data if it is not built according to the present standard, Small kernel size - more noisy vertical lines, Large kernel size - Loss of table lines
- Future scope - introduce machine learning methods to build several kinds of target detection model that can be used to perform area detection and content recognition on complex table types.

[5] 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.

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
- A teacher needs to spend **4 to 6 hours for manually entering mark data**.
- This project aims to develop an **automated solution** that streamlines the aforementioned problems.

Research Scope and Objectives

- Scopes:
  - **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**.
- Objectives:
  - Create a tool to **help teachers in data entry process**.
  - Implementation of **OCR** using the **CNN Architecture**.
  - Create **modular and customizable program** for the system.

Application

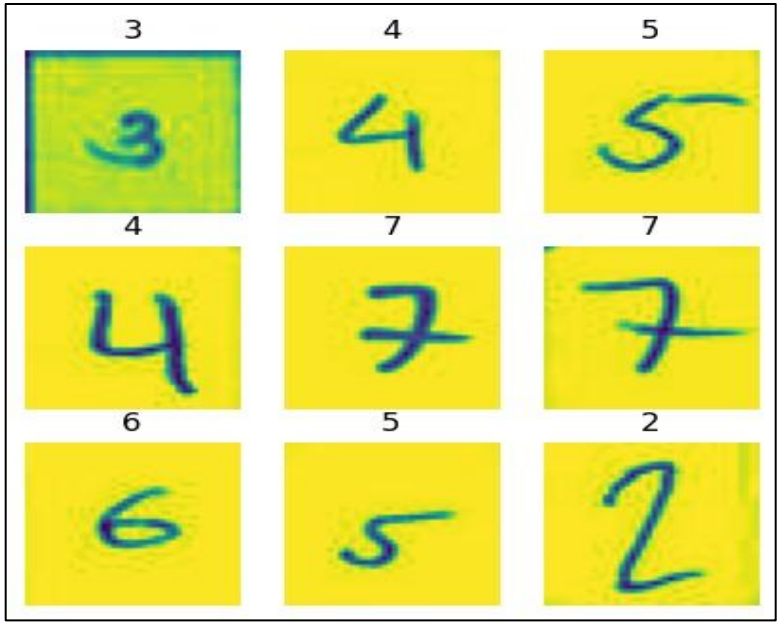
- **Detection of handwritten marks** from images and their **conversion to CSV** format.
- **Simplify the mark entry process** by **avoiding manual input** on each cell.
- Final **output obtained as CSV file** comprising all numbers extracted from the input images.
- Designed specifically for **SJCET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

Methodology - Data Collection

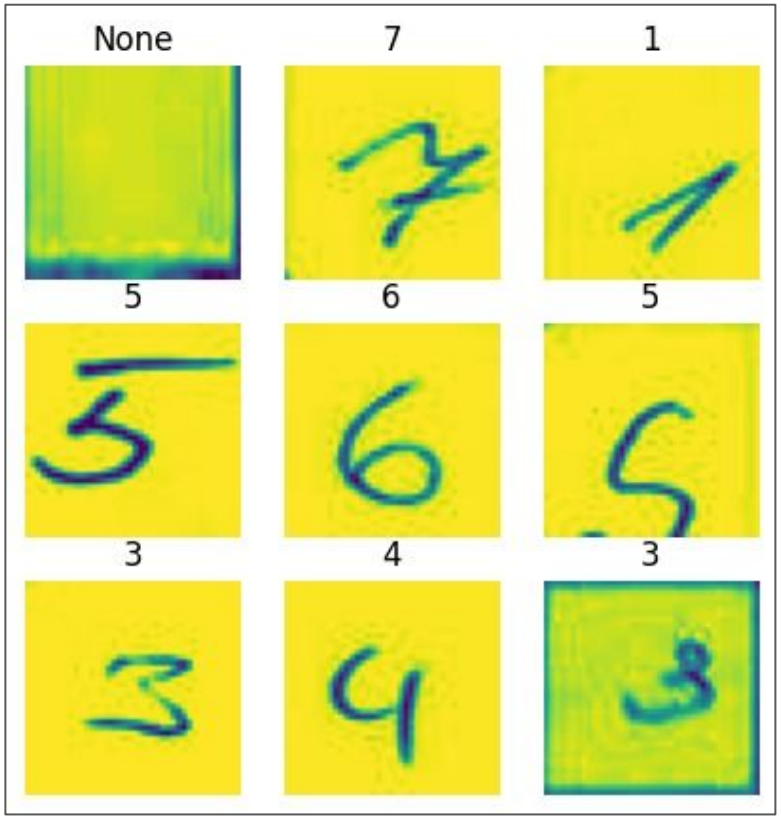
- Main data sources -

  College answer sheets (7,420) and

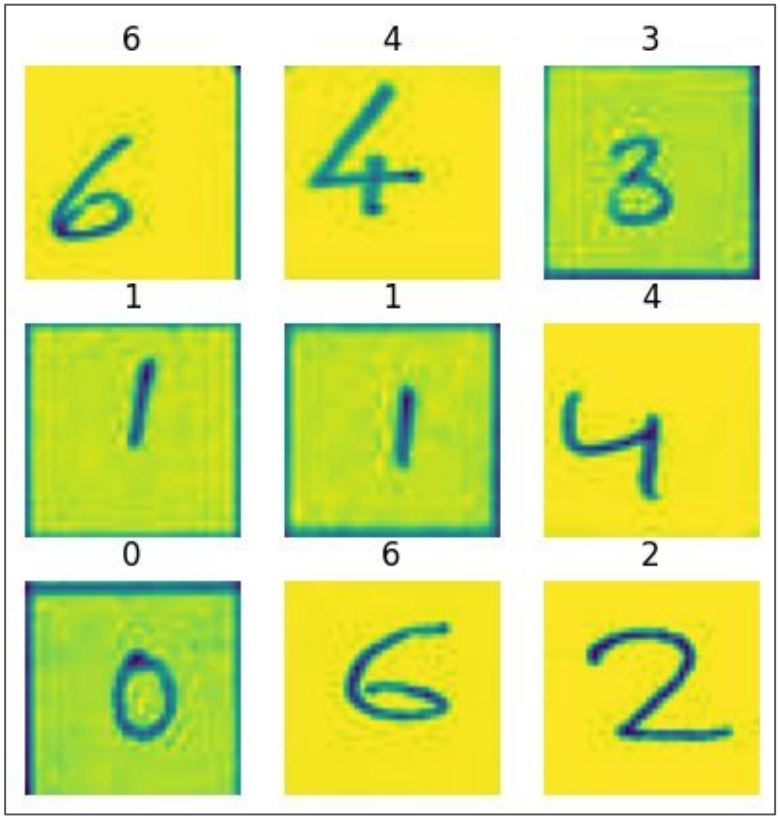
  Kaggle number dataset (13, 855)
- Data Type - Image data of JPG format.
- Data Size - 40px x 40px x 3 channel
- Data Storage - Zip Format in Google Cloud.



Training Set



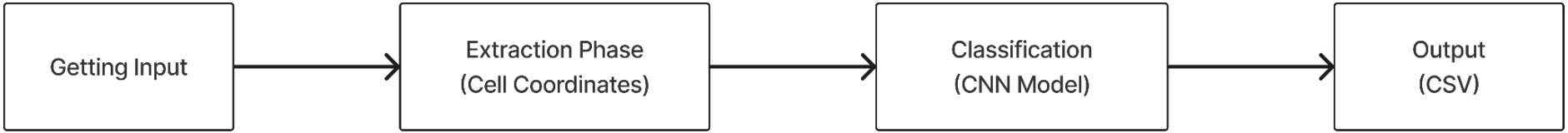
Validation Set



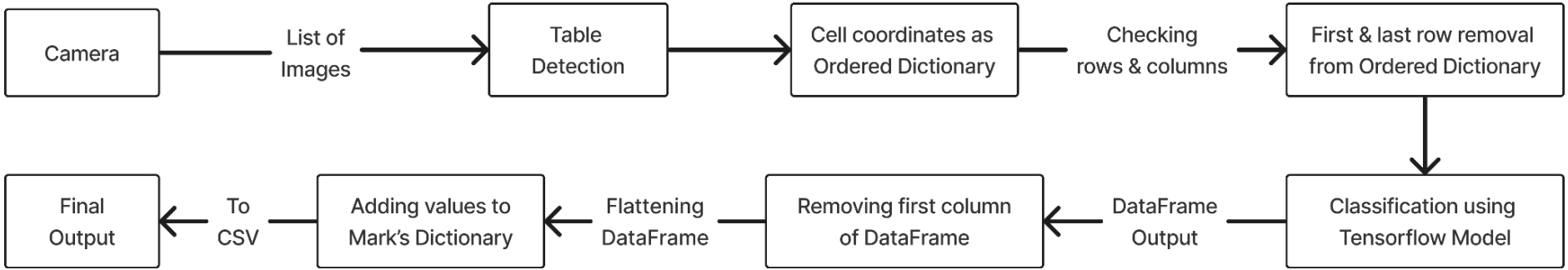
Testing Set

Methodology - Block Diagram

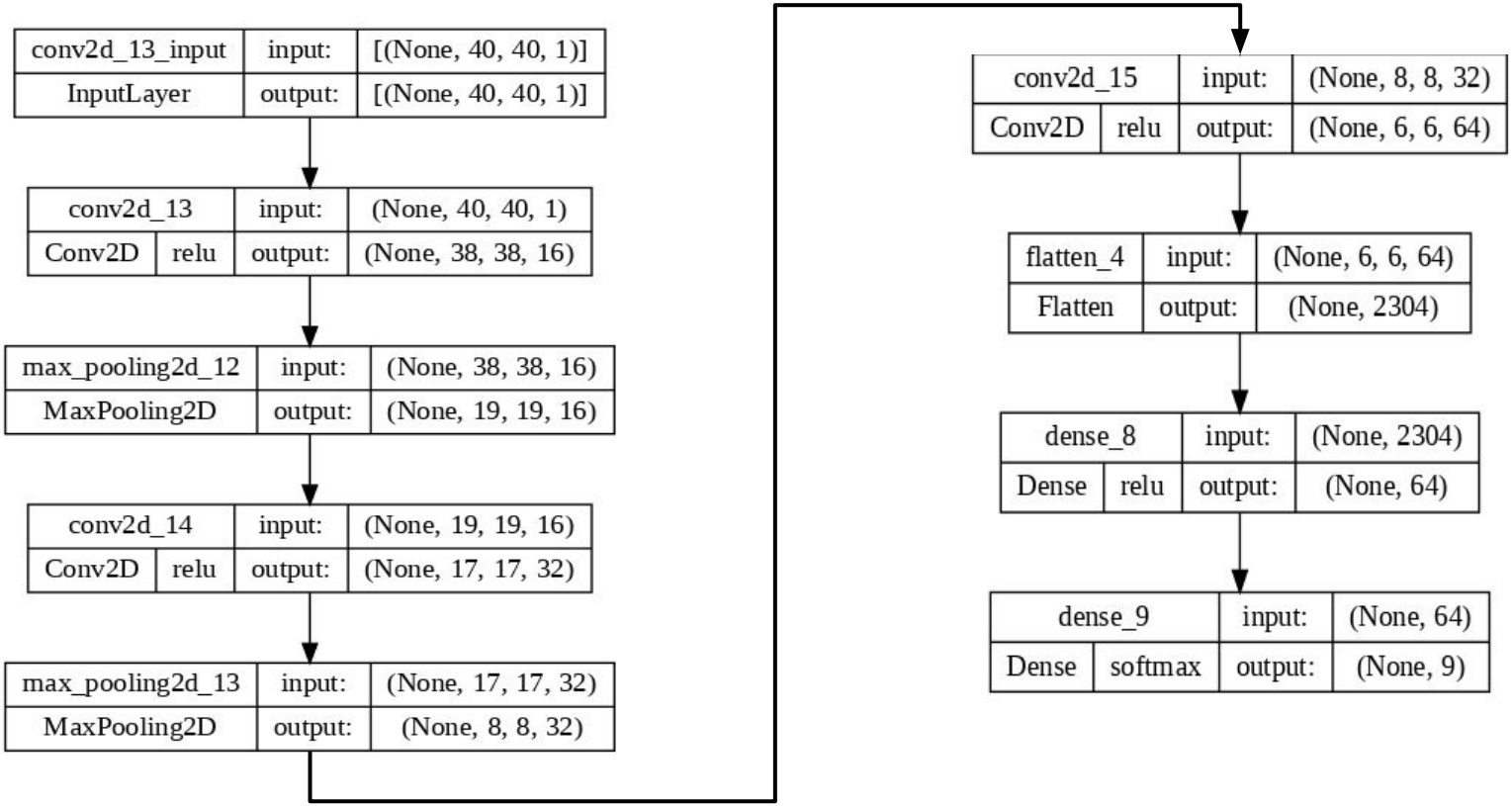
Our Work in 4 Steps



Working of System



Neural Network Architecture

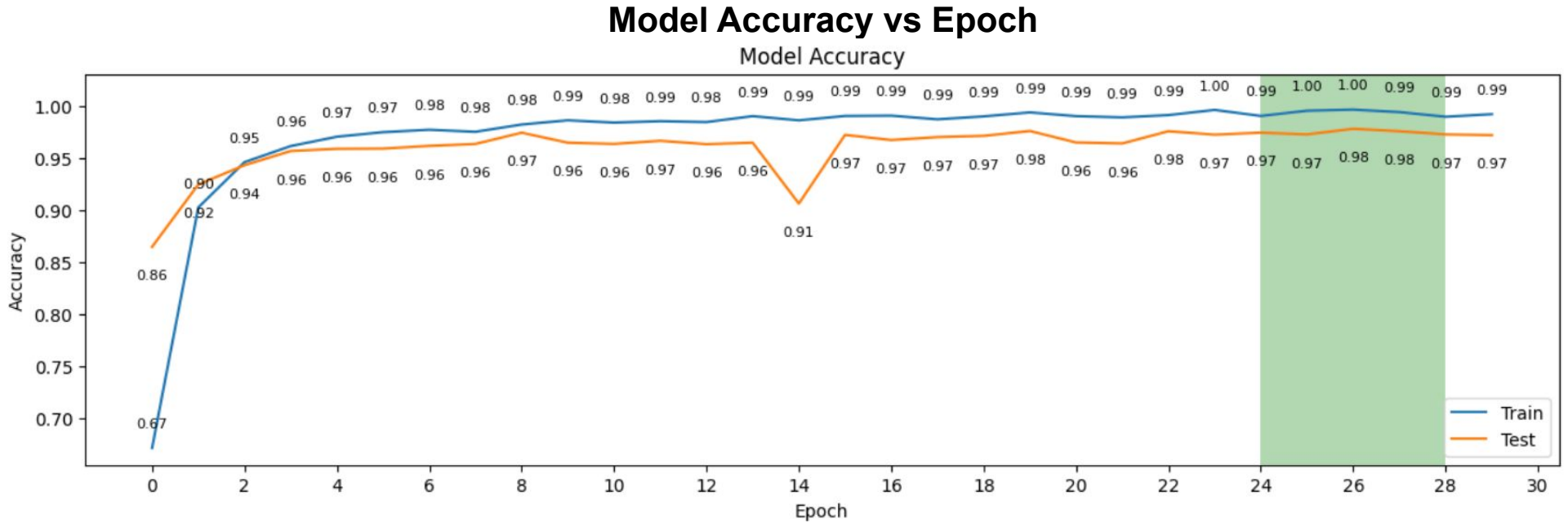


Methodology - Techniques

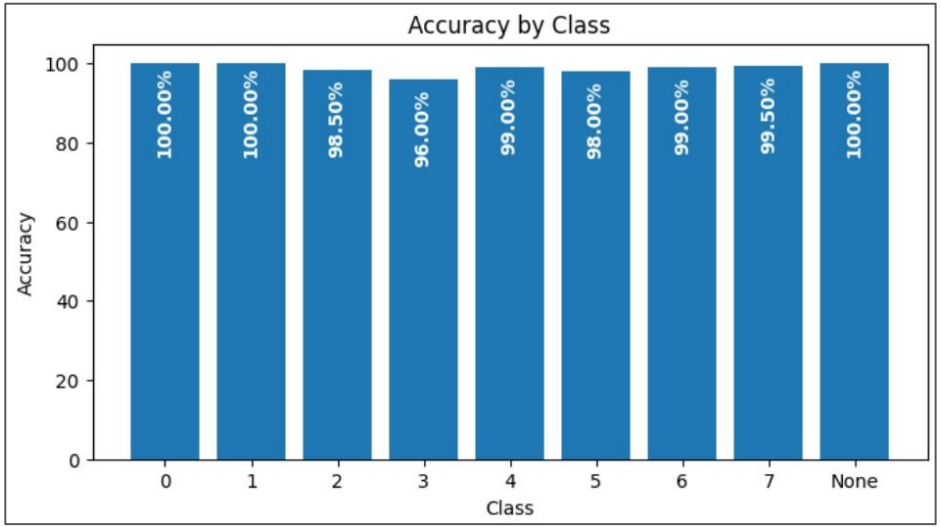
- Increased the speed of processing from **10 minutes to 3 minutes**
  - **PaddleOCR takes 10 minutes** to complete whole process.
  - New processing is done on **built-in data structure** (ordered dictionary).
  - First & last row are removed by **deleting the keys** of ordered dictionary.
  - Then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

Results and Discussion

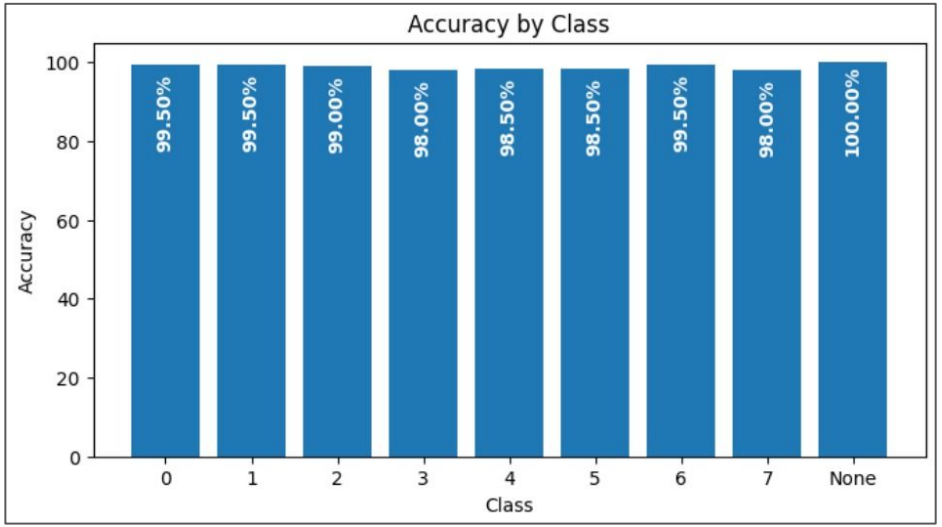
- As epoch value increases, we see training and testing accuracies gradually increasing.
- Model performs best when the epoch value is within the shaded region.



Accuracy by Class



CNN_Model_0

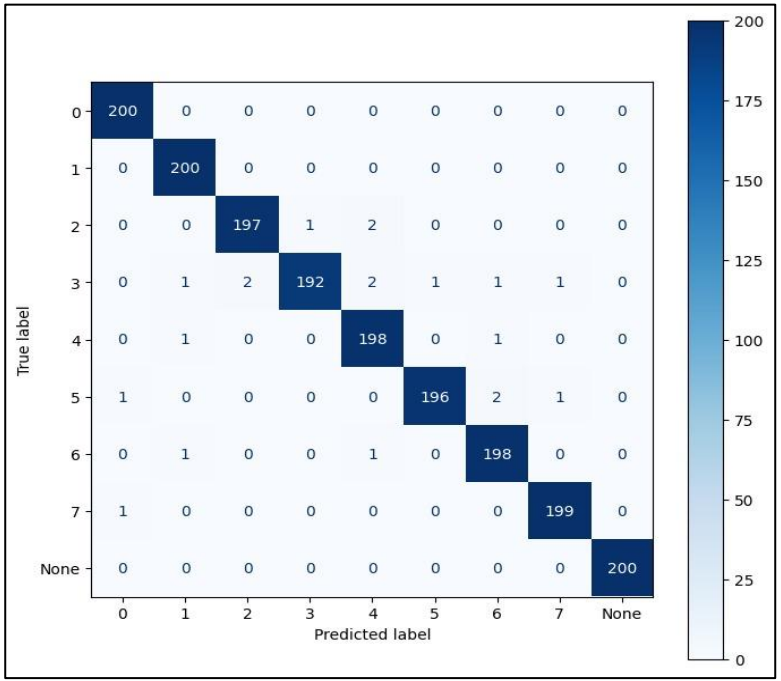


CNN_Model_1

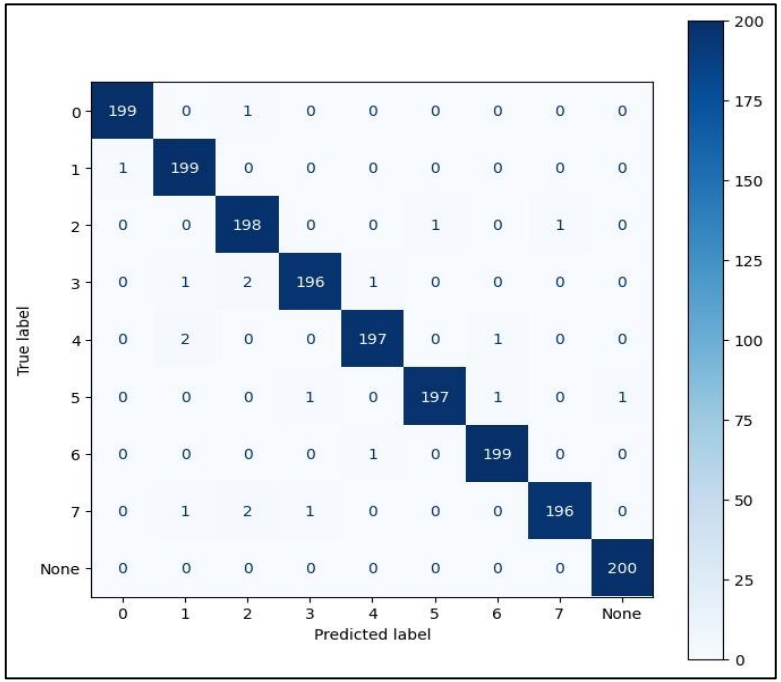
- Comparison of detection accuracy of each class for CNN_Model_0(99%), and CNN_Model_1(99.2%).
- Accuracy values improve in CNN_Model_1 over values of CNN_Model_0.

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	Overall
	CNN_Model_0									
Precision	0.99	0.99	0.99	0.99	0.98	0.99	0.98	0.99	1.00	0.988
Recall	1.00	1.00	0.98	0.96	0.99	0.98	0.99	0.99	1.00	0.987
F1-Score	1.00	0.99	0.99	0.98	0.98	0.99	0.99	0.99	1.00	0.999
	CNN_Model_1									
Precision	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.99	1.00	0.988
Recall	0.99	0.99	0.99	0.98	0.98	0.98	0.99	0.98	1.00	0.986
F1-Score	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99	1.00	0.988

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

Figure 1. Cells with mark written correctly

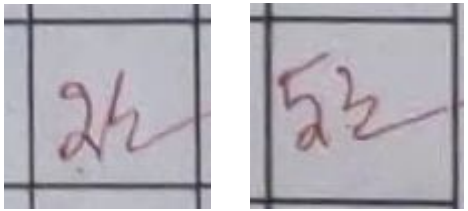


Figure 2. Cells with half marks (unable to detect)

Figure 3. Cells with cuts & corrections (unable to detect)

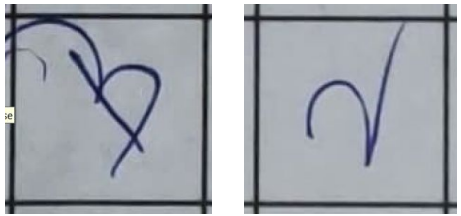


Figure 4. Cells with hard-to-recognize marks (may give false result)

Conclusion and Future Scope

- The project can significantly **accelerate data processing workflows**.
- The system can **completely process 60 papers in just 2.5 minutes**, which shows **speed of the application**.
- The **system performed well for our objective** of designing an OCR tool for detecting handwritten marks that helps for data entry process.

- **Limitations:**
  - Difficulty in **detecting decimal part** (.5) of numbers like 3.5.
  - OCR tool's **efficiency is affected** by **stray marks, corrections, and unsteady writings**.
  - **Proper lighting** is required for **best camera performance**.
- **Future scope:**
  - **Improve model architecture** or **use of another neural network architecture** to the current limitations.
  - Developing a **mobile application** for the project.
  - **Integration with LMS** or other Student Information Systems.

References

[1] C.ShanWei, S.LiWang, Ng T.Foo, Dzati A.Ramli, “A CNN based Handwritten Numeral Recognition Model for Four Arithmetic Operations” 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science, Vol.192, pp:4416-4424, 2021.

[2] 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.

[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.Yuan, H.Li, M.Wang, R.Liu, C.Li, B.Wang, “An OpenCV-based Framework for Table Information Extraction”, 2020 IEEE International Conference on Knowledge Graph (ICKG), IEEE Xplore, pp: 621-628, Sep 2020.

[5] 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.

[6] Md. Ajij, S.Pratihar, Diptendu S.Roy, T.Hanne, “Robust Detection of Tables in Documents Using Scores from Table Cell Cores”, SpringerNature Computer Science Journal, Vol.3, No.161, pp: 1-19, Feb 2022.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] B.Gatos, D.Danatsas, I.Pratikakis, S.J.Perantonis, “Automatic Table Detection in Document Images”, International Conference on Pattern Recognition and Image Analysis (ICAPR), Pattern Recognition and Data Mining, pp: 609–618, 2005.

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

Classification Report

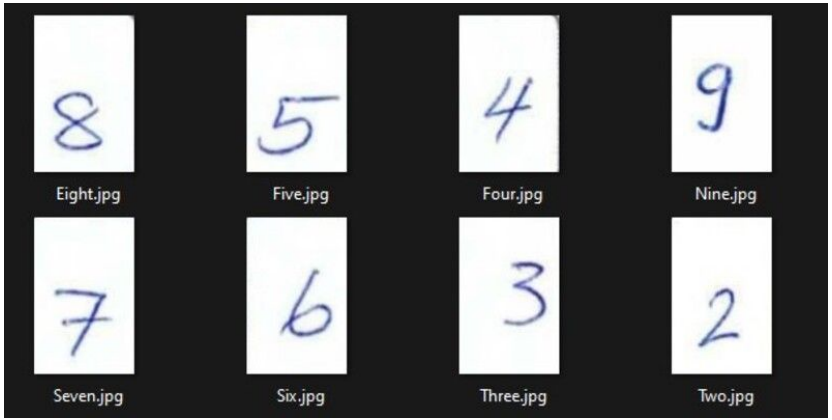
Precision: How accurate the models are in predicting positive samples.

Recall: How effectively a model can identify positive samples.

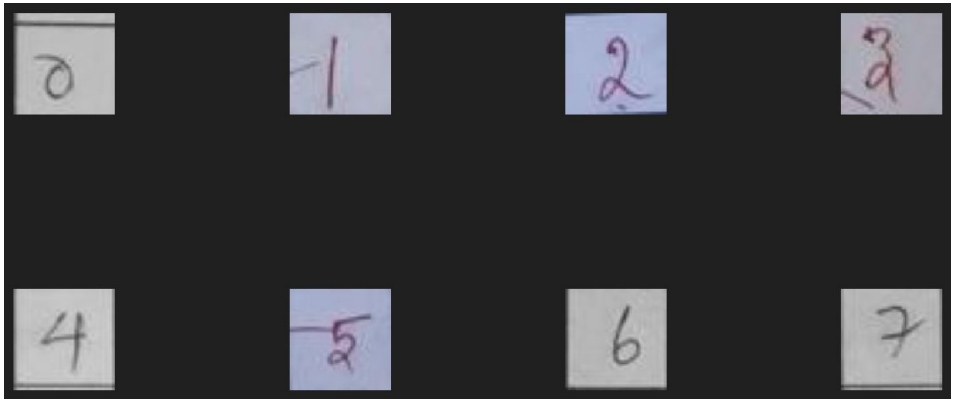
F1-Score: The average of precision and recall, providing a balanced measure of a model's performance.

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  (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.

- 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

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.

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.

[4] 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.

[11] 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.

[13] 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.