Cheque OCR Classification and Prediction System

Cheque OCR Classification and Prediction System

Executive Summary
The goal of this project was to develop a system that reads scanned cheque images and extracts important fields using Optical Character Recognition (OCR) and Machine Learning (ML) techniques. The fields include: bank name, payor name, cheque amount, and cheque date. This initiative helps automate manual cheque processing, reduces data entry errors, and streamlines document verification.

1. Methodology
1.1 Data Source
The system is trained using data provided in .parquet format, where each row contains scanned cheque images and corresponding labels in a structured string format. Each image is associated with a set of messages including metadata generated by an assistant in the form of pipe-delimited (|) labels.

1.2 Label Parsing
The label string is parsed into a dictionary of key-value pairs. Each field, such as bank_name, payor_name, etc., is split and extracted to ensure the correct mapping of values to image bytes.

1.3 OCR Extraction
We used Tesseract OCR via the pytesseract Python library to extract text from cheque images. All images were first converted to grayscale (L mode in PIL) for better accuracy before applying OCR.

1.4 Feature Engineering
The raw text output from OCR was processed using TF-IDF Vectorization. This converts text into numerical features suitable for input to machine learning models.

1.5 Model Selection
Each target field (bank name, payor name, cheque amount, cheque date) was trained using a separate Random Forest Classifier. The models were trained and evaluated individually for better control and interpretability.

1.6 Model Evaluation
Each model was evaluated using:
- Train-test split (80/20)
- Accuracy, Precision, Recall, F1-Score from classification_report
- Filtering out samples with missing labels or OCR failures

1.7 Prediction and Reporting
The trained models were used to predict fields from OCR-extracted text for each cheque. Results were exported into a well-formatted Excel file using the openpyxl library. Auto-fit columns, centered alignment, and sheet naming were handled automatically.

2. Tools and Libraries Used
Tool / Library	Purpose
pandas	Data manipulation
numpy	Numerical operations
pytesseract	OCR text extraction from cheque images
PIL (Pillow)	Image processing
sklearn	Machine learning (TF-IDF, RandomForest, train/test split, metrics)
tqdm	Progress bars for processing large datasets
openpyxl	Excel export with formatting
io	Byte stream handling for image inputs
datetime	Timestamp generation for output filenames

3. File and Model Architecture
- Input:
  - .parquet files containing image_bytes and label messages.

- Processing Functions:
  - extract_text_from_image_bytes: Converts byte stream into text using OCR.
  - parse_label_text: Extracts metadata fields from assistant message.
  - prepare_training_data: Prepares structured DataFrame for model training.
  - train_field_model: Trains and evaluates a model per target field.
  - predict_and_export: Generates predictions and saves them to Excel.

- Output:
  - Excel file with columns: cheque_number, bank_name, payor_name, cheque_amount, cheque_date

4. Output Sample (Screenshot)
The screenshot displays the final predictions for each cheque image. The model successfully classified a majority of rows with high accuracy. Missing predictions (e.g., row 1 bank name) indicate potential OCR failure or insufficient training data for rare labels.

cheque_number	bank_name	payor_name	cheque_amount	cheque_date
1	nan	Synovus	10900.56	13-02-2024
2	First Citizens Bank	Geoffrey R. Mcdonald,Pc	8028.8	21-02-2024
3	PNC Bank	Booth Manufacturing Co	11178.48	06-03-2024
4	TD Bank	Boys & Girls Clubs of Philadelphia	3801.66	19-10-2023
5	M&T Bank	Hopkings & Wayson INC	15430.0	22-12-2023
6	United Bank	Columbia Country Club	9016.0	15-02-2024
7	Central State Bank	Helena High School	8885.95	08-02-2024
8	Servisfirst Bank	Lakeshore Station LLC	13135.51	02-10-2023
9	Synovus Bank	Dick Dyer & Associates INC	2317.68	15-12-2023
10	U.S. Bank	Formatech Inc	9288.05	04-01-2024
12	JP Morgan Chase Bank	Skycon Group LLC	4396.35	29-12-2023
13	The First Natonat Bank	Ottawa Tribe Of Oklahoma	40275.6	28-12-2023
14	Central State Bank	Contractors' Building Supply & Hardware,Inc.	4351.36	08-03-2024
15	South State Bank	Holman, Inc	13694.0	14-03-2024
16	Westbury Bank	Precision Plus, Inc	13153.53	13-03-2024
17	Wells Fargo Bank	Caravan Supply Chain	2500.0	15-02-2024
18	Enterprise Bank & Trust	Gourmet Specialty Foods, Llc	6672.0	09-04-2024
19	Sterling Bank	Saint Joseph'S Hospital Center	24205.55	14-02-2024
20	Wells Fargo Bank	Thresher Energy, Inc	2500.0	10-02-2024
21	Wells Fargo Bank	Westbrook Service Corporation	2500.0	08-03-2024
22	First Horizon Bank	Springs Creative Products Groups Llc	9017.12	09-02-2024
23	Bank Of America	The Boy'S Farmers Market Inc	11263.83	23-02-2024
24	Sturdy Savings Bank	Middle Township Board Of Education	8678.91	19-12-2023
25	First Republic Bank	Novo Construction	1986.3	20-10-2023
26	Truist	Contractors' Building Supply & Hardware,Inc.	4351.36	10-02-2024
27	Independent Financial	E&J Cabinets	14070.0	24-10-2023
28	Pathways Financial Credit Union	nan	1189.29	16-01-2024
29	Sterling National Bank	Ohel Children Home & Family Services	7920.98	22-02-2024
30	Wells Fargo Bank	H & N North America Llc	2242.9	21-03-2024
31	Wells Fargo Bank	Christopher T Metz Kerry C Metz	6145.92	19-10-2023
32	J.P. Morgan Chase Bank	Christopher T Metz Kerry C Metz	16302.12	13-03-2024
33	Bank Of America	J.Smith Lanier & Co	5360.0	13-03-2024
34	Truist	Susan N.Weiner	1285.63	15-02-2024
35	Wells Fargo Bank	Metro National Corp	3594.27	10-10-2023
36	Centennial Bank	Tom Jenkin's Bar-B-Q	1883.5	06-02-2024

5. Challenges and Solutions
Challenge	Solution
Noisy OCR text	Used grayscale preprocessing to improve quality
Mismatched label formats	Added robust parsing logic with validation checks
Sparse class distribution	Handled missing fields and excluded null samples before training
Excel formatting	Automated column sizing, text alignment, and sheet management

6. Recommendations
- Data Augmentation: Enrich the training set with more diverse examples per label.
- Deep Learning Models: Explore LayoutLM, TrOCR, or BERT for document understanding.
- Cloud OCR APIs: For better OCR accuracy, consider Google Vision, AWS Textract.
- Web Interface: Build a simple UI for uploading cheques and downloading predictions.

7. Conclusion

This system demonstrates a functional pipeline that automates cheque field extraction using OCR and ML. It reduces manual effort and sets the foundation for scalable document intelligence solutions. With further improvements, this pipeline can be extended to support additional document types and multilingual OCR.

Appendix

- Script file includes end-to-end functionality with clear modular design.
- Timestamped output file names prevent accidental overwrite.
- The system is robust to missing data and logs skipped rows.