Expense Receipt OCR and AI-Driven Insight Generation for Sales CRM Systems

Omkar Sanjay Kamble[1] Mr. Shripad S. Bhide[2]
[1]Student [2]Assistant Professor
[1,2]Master of Computer Applications
[1,2]P.E.S. Modern College of Engineering, Pune, Maharashtra, India

Abstract — Expense management in modern organizations is increasingly dependent on digital workflows and automation systems. However, many organizations still rely on manual expense receipt processing and fragmented operational analysis methods. Traditional OCR systems frequently fail when handling low-quality receipts, payment screenshots, inconsistent layouts, and noisy image conditions. Additionally, organizations require intelligent business insights derived from CRM operational data for effective decision-making. This paper presents a hybrid Expense Receipt OCR and AI- Driven Insight Generation framework integrated into a Sales CRM platform. The proposed system combines image preprocessing, hybrid OCR extraction using Tesseract.js and Python OCR ensemble methods, machine learning based field extraction, expense category prediction, confidence-aware review logic, and AI-powered CRM insight generation. The framework supports both physical receipts and digital payment screenshots from applications such as Google Pay, PhonePe, and Paytm. The OCR module extracts structured expense information including vendor name, date, GST amount, category, total amount, and confidence metrics. The AI insight module analyzes CRM operational data including leads, deals, quotations, meetings, expenses, and follow-ups to generate business recommendations using Gemini AI models with deterministic fallback support. The proposed architecture improves operational efficiency, reduces manual expense processing effort, enhances structured data extraction, and supports intelligent business decision-making in CRM environments.

Keywords: OCR, Artificial Intelligence, CRM Analytics, Expense Automation, Document Intelligence, Tesseract, Easy- OCR, Machine Learning

I. Introduction

Organizations that rely heavily on sales operations process a large number of expense receipts every day. These receipts are generated through travel, accommodation, food, client meetings, office purchases, and operational activities. Manual entry of these receipts into CRM or accounting systems is time-consuming and often results in inconsistent data storage. Expense receipts exist in multiple formats such as printed bills, photographed invoices, scanned receipts, and payment screenshots captured from digital payment applications. The diversity in document quality, font styles, layouts, lighting conditions, and image orientation creates major challenges for traditional OCR systems.

Conventional OCR systems focus primarily on extracting raw text from images. However, business applications require structured information extraction, intelligent validation, category prediction, review workflows, and operational analytics rather than raw OCR output alone.

At the same time, modern CRM systems generate large amounts of operational data related to leads, deals, quotations, meetings, clients, and follow-ups. Organizations increasingly require intelligent insight generation systems that can analyze CRM activities and provide recommendations for business growth and operational improvement.

This research proposes a hybrid OCR and AI-driven insight generation framework integrated into a Sales CRM system. The proposed system combines preprocessing techniques, OCR ensemble extraction, machine learning field inference, confidence scoring, review workflows, and AI-generated operational recommendations.

II. Problem Statement

Many organizations still depend on manual expense processing workflows. Employees upload receipts manually and finance teams often spend significant time validating vendor names, dates, amounts, and categories.

The major challenges associated with traditional systems include:

- Low OCR accuracy on noisy or blurred receipts
- Inconsistent extraction due to varying receipt layouts
- Difficulty handling payment screenshots
- Absence of structured field extraction
- Lack of confidence-aware review systems
- No integration with CRM operational analytics
- Dependence on manual operational monitoring

Additionally, existing CRM systems often fail to provide actionable business recommendations derived from operational data. Businesses require automated analytics systems capable of identifying stale leads, overdue follow-ups, pipeline risks, and revenue opportunities.

Therefore, there is a need for a practical hybrid OCR and AI insight framework capable of structured expense extraction and operational intelligence generation.

III. Objectives

The primary objectives of the proposed system are:

1) Develop a hybrid OCR framework for receipt understanding.
2) Support printed receipts and digital payment screenshots.
3) Improve OCR accuracy through preprocessing and OCR ensembles.
4) Extract structured expense fields automatically.
5) Predict expense categories intelligently.
6) Compute confidence scores for validation support.
7) Provide review workflows for low-confidence records.
8) Integrate expense automation into CRM workflows.
9) Generate AI-driven CRM operational insights.
10) Provide deterministic fallback analytics support.

IV. Literature Review

OCR technology has evolved significantly over the last decade. Tesseract OCR remains one of the most widely used

open-source OCR engines because of its flexibility and practical deployment capabilities. Tesseract performs effectively on structured printed text but often struggles under noisy image conditions.

EasyOCR introduced deep learning-based recognition techniques that improved OCR performance on scene text and multilingual content. Ensemble OCR methods that combine multiple OCR engines are increasingly used to improve ex- traction robustness.

Research in document intelligence extends beyond OCR ex- traction toward semantic understanding, structured field extraction, and layout-aware analysis. Modern systems frequently combine OCR outputs with machine learning classifiers and heuristic validation techniques.

Machine learning based field extraction systems analyze textual and spatial patterns to identify vendor names, dates, totals, and semantic labels. Hybrid approaches combining heuristics and ML inference often outperform purely OCR- based methods.

Large language models such as Gemini and GPT-based systems are increasingly used for operational analytics and business summarization. However, production systems require fallback deterministic logic due to network failures, quota limitations, and unpredictable model responses.

V. PROPOSED SYSTEM

The proposed framework consists of two major modules:

1) Expense Receipt OCR Module
2) AI Insight Generation Module

The OCR module processes uploaded receipt images and payment screenshots to generate structured expense records. The AI insight engine analyzes CRM operational data and generates business recommendations, priorities, summaries, and analytics.

The complete workflow includes:

- Receipt upload and preprocessing
- Hybrid OCR extraction
- OCR post-processing
- Document style detection
- ML-based field extraction
- Expense category prediction
- Confidence scoring and review
- CRM storage integration
- AI insight generation
- Dashboard analytics visualization

VI. SYSTEM ARCHITECTURE

The proposed architecture contains the following components:

- User Upload Interface
- OCR Service API
- Image Preprocessing Engine
- Tesseract.js OCR Engine
- Python OCR Ensemble
- ML Field Extractor
- Category Prediction Engine
- Confidence Scoring Module
- Expense Database
- CRM Analytics Layer
- Gemini AI Engine
- Deterministic Fallback Engine
- Dashboard and Reporting Layer

A. Expense OCR Architecture

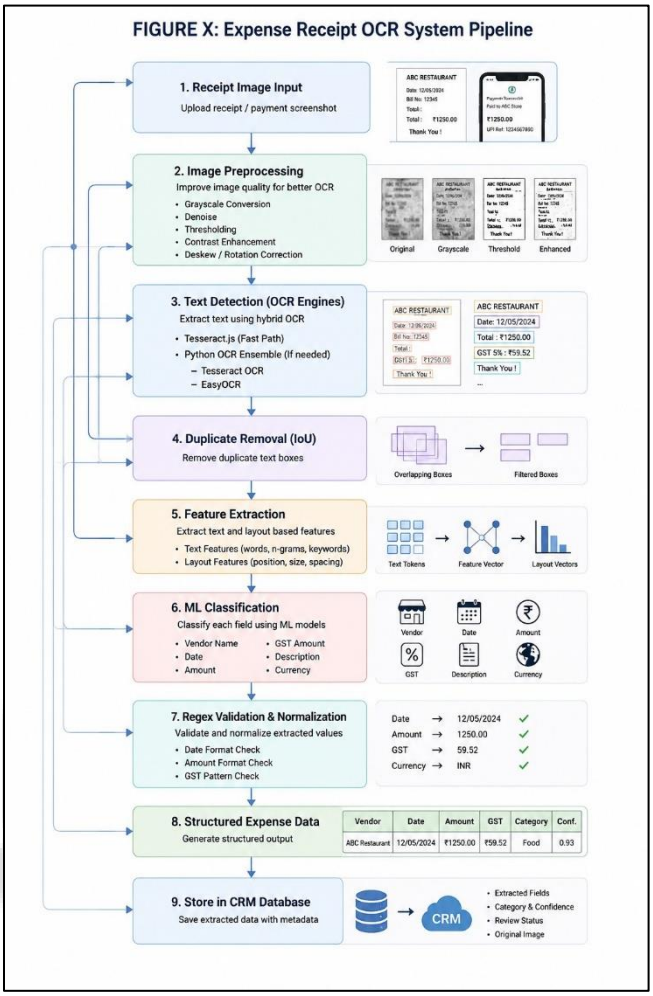

Fig. 1: Expense Receipt OCR System Pipeline

B. AI Insight Architecture

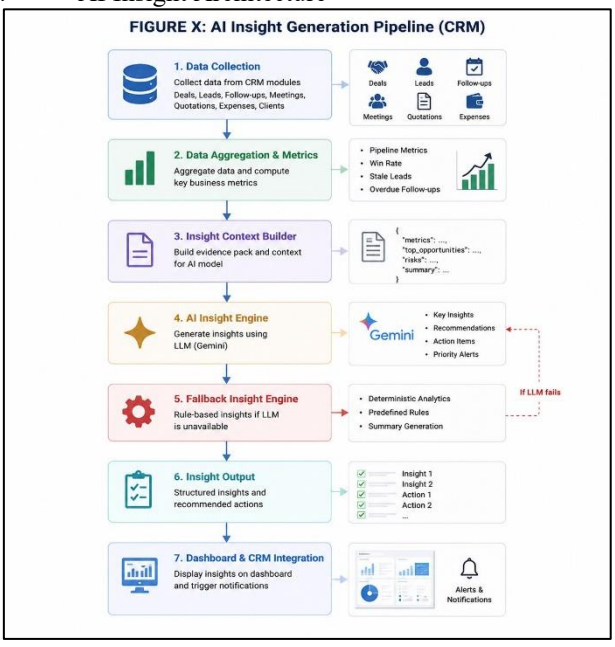

Fig. 2: AI Insight Generation Pipeline

VII. METHODOLOGY

A. Image Acquisition

The system accepts the following input types:

- Printed receipts
- Mobile captured receipts
- Scanned invoices
- Payment screenshots
- Cropped receipt images

The upload interface validates file type and prepares the image for OCR processing.

B. Expense OCR Workspace

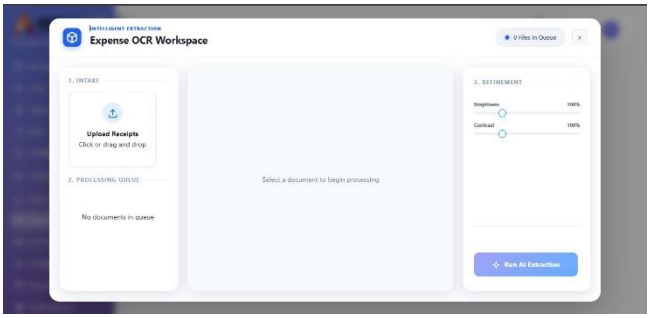

Fig. 3: Expense OCR Workspace Interface

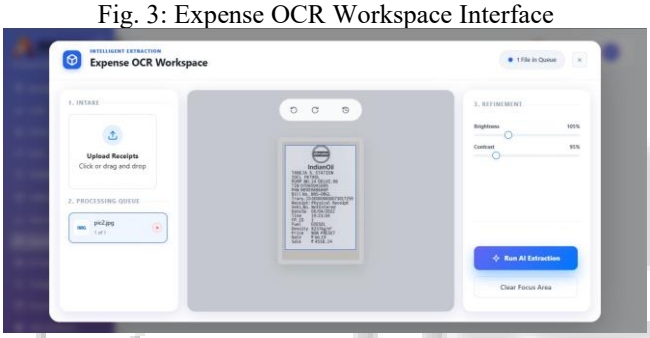

Fig. 4: Expense OCR Processing Queue and Receipt Preview

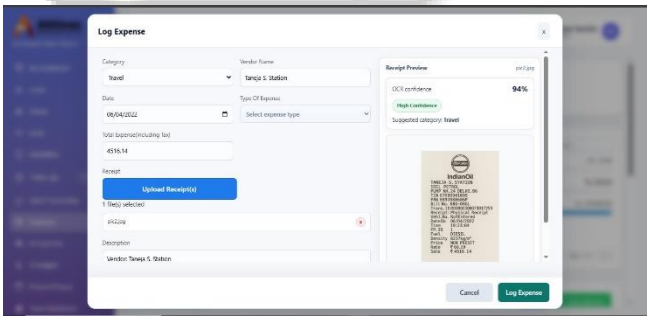

Fig. 5: Expense Logging and OCR Confidence Extraction

C. Image Preprocessing

The preprocessing stage improves OCR readability and extraction quality. The following preprocessing operations are supported:

- Cropping
- Rotation correction
- Brightness adjustment
- Contrast enhancement
- Denoising
- Thresholding
- Grayscale conversion

OpenCV and Python preprocessing scripts are used to improve image quality before OCR extraction.

D. Hybrid OCR Extraction

The system uses a hybrid OCR strategy consisting of:

- Tesseract.js OCR
- Python Tesseract OCR
- EasyOCR Ensemble

The workflow first attempts fast extraction using Tesser- act.js. If extraction confidence is low or fields are incomplete, Python OCR ensemble extraction is triggered.

OCR outputs are merged using line grouping, duplicate removal, and normalization logic.

E. Document Style Detection

The system identifies whether the uploaded image is:

- Standard receipt
- Payment screenshot

Keyword detection is used for style classification. Examples include:

- UPI
- Transaction
- Paid To
- Credited
- Debited
- Google Pay
- PhonePe
- Paytm

This classification improves field extraction accuracy be- cause receipts and payment screenshots follow different layout patterns.

F. Field Extraction and Classification

The system extracts:

- Vendor Name
- Expense Date
- Total Amount
- GST Amount
- Description
- Currency
- Expense Category

Field extraction combines:

- Heuristic rules
- ML field inference
- Regex validation
- Confidence-aware validation

Machine learning models analyze text patterns and layout information to classify fields.

G. Category Prediction

Expense category prediction uses:

- Vendor text
- Description keywords
- OCR lines
- Rule-based scoring
- ML category classification Supported categories include:
  - Travel
  - Hotel
  - Food

- Stationery
- Marketing
- Client Meeting
- Event
- Other

H. Confidence Scoring

The system computes:

– OCR engine confidence
– Field-level confidence
– Overall extraction confidence

Low-confidence records are automatically flagged for manual review.

I. AI Insight Generation

The AI insight module analyzes CRM operational data including:

– Leads
– Deals
– Quotations
– Meetings
– Expenses
– Clients
– Follow-ups

The analytics layer computes:

– Open pipeline
– Weighted forecast
– Win rate
– Stale leads
– Overdue follow-ups
– Revenue opportunities
– Team performance

Gemini AI models generate summaries, recommendations, priorities, and outlook indicators.

J. AI Insight Dashboard

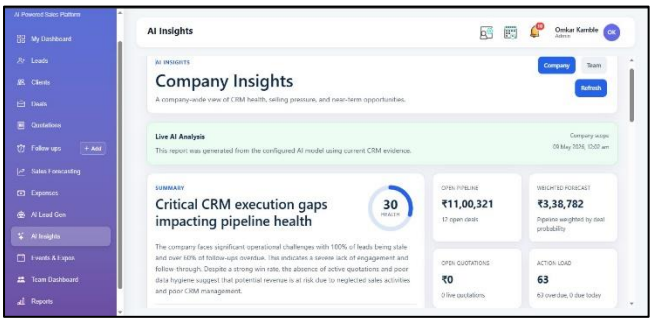

Fig. 6: AI Insight Dashboard Overview

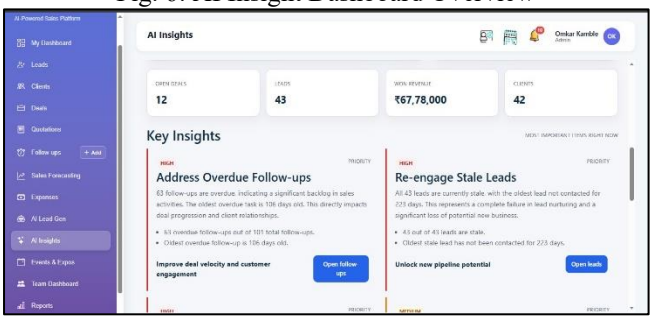

Fig. 7: AI Recommendation and Action Analysis

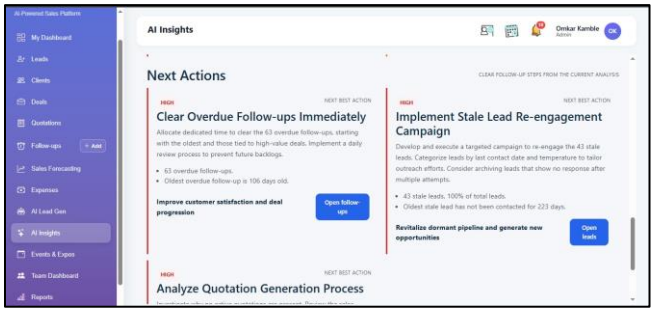

Fig. 8: Company Insight Analytics Dashboard

VIII. ALGORITHMS USED

The proposed framework uses the following algorithms and technologies:

– Tesseract OCR
– EasyOCR
– OpenCV preprocessing
– OCR ensemble extraction
– ML-based field inference
– Regex validation
– Confidence scoring
– Gemini AI insight generation

IX. EXPERIMENTAL RESULTS

Experimental evaluation is to be conducted on real-world receipt datasets containing both printed receipts and payment screenshots.

Suggested evaluation metrics include:

– OCR extraction accuracy
– Field extraction precision
– Category prediction accuracy
– Average processing time
– Manual review reduction

Metric	Value
Vendor Extraction Accuracy	TBD
Date Extraction Accuracy	TBD
Amount Extraction Accuracy	TBD
Category Prediction Accuracy	TBD
Average Processing Time	TBD

Table I: Suggested Evaluation Metrics

X. ADVANTAGES

The proposed framework provides several advantages:

– Supports multiple receipt formats
– Handles payment screenshots effectively
– Hybrid OCR improves extraction robustness
– ML-based field inference improves structured extraction
– Confidence-aware review workflow improves reliability
– CRM integration enables operational automation
– AI-driven insights improve business decision-making
– Deterministic fallback ensures operational continuity

XI. LIMITATIONS

Despite the advantages, certain limitations remain:

– OCR accuracy depends on image quality
– Highly blurred receipts reduce extraction quality

- Complex layouts may affect field detection
- ML models require labeled training data
- AI insight quality depends on CRM data quality

XII. FUTURE SCOPE

Future improvements may include:

- Multilingual OCR support
- Transformer-based document understanding
- Fraud detection systems
- Mobile OCR applications
- Voice-assisted expense reporting
- Advanced analytics dashboards
- Real-time operational forecasting

XIII. CONCLUSION

This paper presented a hybrid Expense Receipt OCR and AI-Driven Insight Generation framework for Sales CRM systems. The proposed architecture combines preprocessing, OCR ensemble extraction, machine learning based field inference, confidence-aware validation, and AI-powered CRM analytics. The framework improves expense automation, reduces manual effort, enhances structured data extraction, and supports intelligent operational decision-making in CRM environments. The proposed system demonstrates how OCR automation and AI-driven analytics can be combined into a unified enterprise workflow platform.

REFERENCES

[1] R. Smith, "An Overview of the Tesseract OCR Engine," ICDAR, 2007.
[2] EasyOCR Documentation, Jaided AI, 2024.
[3] OpenCV Documentation, Open-Source Computer Vision Library, 2024.
[4] Google Gemini AI Models Documentation, 2025.
[5] S. Mori, "Historical Review of OCR Research and Development," Proceedings of the IEEE, 1999.
[6] A. Graves et al., "Deep Learning for Scene Text Recognition," Springer, 2013.
[7] Ian Goodfellow et al., "Deep Learning," MIT Press, 2016.
[8] Christopher Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

