

**IMPROVING DOCUMENT UNDERSTANDING OF MEDICAL INVOICES
USING MULTIMODAL APPROACH: A COMPARATIVE STUDY**

By

Abyakta Nadhif Mahendratama
12102003

BACHELOR'S DEGREE
in

INFORMATION TECHNOLOGY - AI & DATA SCIENCE
FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

Revised after the Thesis Defence on 3 July 2025

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

Abyakta Nadhif Mahendratama

Student



Date

Approved by:

Dr. Ir. Heru Purnomo (Pung), M.Eng

Thesis Advisor

Date

Andi Danna Taqwim

Thesis Co-Advisor

Date

Dr. Maulahikmah Galinium, S.Kom, M.Sc

Dean

Date

Abyakta Nadhif Mahendratama

ABSTRACT

IMPROVING DOCUMENT UNDERSTANDING OF MEDICAL INVOICES USING MULTIMODAL APPROACH: A COMPARATIVE STUDY

By

Abyakta Nadhif Mahendratama
Dr. Ir. Heru Purnomo Ipung, M.Eng, Advisor
Andi Darma Taqwim, Co-Advisor

SWISS GERMAN UNIVERSITY

This thesis presents a comprehensive comparative study on document understanding of medical invoices using both OCR and advanced multimodal models. A stratified dataset of 186 real-world Indonesian medical invoices, encompassing diverse forms of visual degradation, was manually annotated at the field level to ensure robust and representative evaluation. The research investigates the effectiveness of OCR systems, advanced multimodal models, and the combination of both systems in extracting structured information from preprocessed invoice images. Standard image preprocessing techniques were applied to all samples prior to evaluation. Performance was quantitatively assessed using established metrics, including precision, recall, and F1-score. The results demonstrate that multimodal models consistently outperform OCR systems on visually degraded invoices. This work offers practical insights for deploying robust automated document understanding solutions in claims processing, highlighting the advantages of integrating preprocessing with modern multimodal models for real-world, domain-specific applications.

Keywords: Document understanding, Optical Character Recognition, multimodal models, medical invoices, Gemini, structured data extraction, preprocessing, Mistral OCR, Pixtral, Qwen 2.5-VL, PaddleOCR,



© Copyright 2025
by Abyakta Nadhif Mahendratama
All rights reserved

DEDICATION

I dedicate this work to my future self and to the journey that lies ahead. May the lessons learned and the challenges overcome here serve as a foundation for growth, perseverance, discovery, and meaningful impact.



ACKNOWLEDGEMENTS

First and foremost, I express sincere gratitude to God Almighty for the guidance, strength, and blessings throughout my studies and this research.

I would like to sincerely thank my thesis advisor, Dr. Ir. Heru Purnomo Ipung, M.Eng, for his invaluable guidance, patience, and insightful feedback, without which this work would not have been possible. His encouragement, expertise, and motivation have profoundly shaped my academic growth.

A special thank-you to my thesis co-advisor, Mr. Andi Darma Taqwim, for generously providing access to the dataset and valuable resources that made this research feasible.

Special thanks to my colleagues and friends for their continuous support, valuable discussions, and cherished camaraderie during challenging times.

Lastly, my appreciation goes to my family and my partner, whose continuous love, support, and prayers have been my greatest source of motivation throughout this academic journey.

Thank you all for your invaluable contributions to this milestone. May blessings be upon us all.

TABLE OF CONTENTS

	Page
STATEMENT BY THE AUTHOR.....	2
ABSTRACT.....	3
DEDICATION.....	5
ACKNOWLEDGEMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES.....	9
LIST OF TABLES.....	10
CHAPTER 1 - INTRODUCTION.....	11
1.1 Background.....	11
1.2 Research Problems.....	12
1.3 Research Objectives.....	12
1.4 Research Limitations.....	13
1.5 Significance of Study.....	13
1.6 Research Question.....	13
1.7 Hypotheses.....	14
CHAPTER 2 - LITERATURE REVIEW.....	15
2.1 General Theory.....	15
2.1.1 Optical Character Recognition.....	15
2.1.2 Multimodal Models.....	15
2.1.3 Document Understanding.....	15
2.1.4 Image Degradation Factors.....	16
2.1.5 Preprocessing Techniques.....	16
2.2 Related Works.....	16
2.2.1 Recent Studies on OCR Models.....	16
2.2.2 Recent Studies on Multimodal Models.....	19
2.3 Table Comparison.....	22
2.4 The Proliferation of AI and the Need for Systematic Validation.....	25
CHAPTER 3 – RESEARCH METHODS.....	27
3.1 Research Overview.....	27
3.2 Dataset Sampling and Annotation.....	28
3.3 Preprocessing Variants.....	29
3.4 Ground Truth Preparation.....	30
3.5 Model Implementation.....	31
3.6 Evaluation Metrics.....	32
3.7 Hardware and Software Specifications.....	34
CHAPTER 4 – RESULTS.....	35
4.1 Quantitative Results.....	35
4.2 Impact of Preprocessing.....	38

4.3 Statistical Significance of Model Differences.....	39
4.4 Error Analysis.....	39
4.5 Robustness on Degraded Inputs.....	41
4.6 Discussion.....	41
4.7 Summary of Findings: A Model Validator's Perspective.....	42
CHAPTER 5 – CONCLUSIONS AND RECOMMENDATIONS.....	43
5.1 Conclusions.....	43
5.2 Recommendations.....	43
GLOSSARY.....	45
REFERENCES.....	46
APPENDICES.....	48
Appendix A: Sample Stimuli.....	48
A.1 Sample Degraded Invoice.....	49
A.2 Multimodal Model Prompt.....	49
Appendix B: Detailed Performance Results.....	50
B.1 Gemini 2.5 Flash (Aggressive Preprocessing).....	50
B.2 Gemini 2.5 Flash (Raw Images).....	50
B.3 PP-OCRv3 (Raw Images).....	51
B.4 PP-OCRv3 (Moderate Preprocessing).....	51
B.5 Qwen 2.5-VL (Aggressive Preprocessing).....	51
B.6 Mistral OCR + Pixtral (Aggressive Preprocessing).....	52
CURRICULUM VITAE.....	53

LIST OF FIGURES

Figure	Page
3.1 Research Overview	27
3.2 Analysis using the Elbow method and Silhouette scores	28
3.6 Classification Metrics Formulae	33
4.4a Date-bleed and blur on date-field extraction	39
4.4b Occlusion on line-item description extraction	40
4.5a Exhibit A on the robustness on degraded input	41
A.1 Sample Invoice from The Dataset with Some Extracted Data	49
A.2 Prompt to Instruct Multimodal Models on JSON Structure and Field Extraction Requirements	49



LIST OF TABLES

Table	Page
2.3a Comparison Table of OCR Models	22
2.3b Comparison Table of Multimodal Models	23
2.3c Benchmark by Models	24
2.4 AI Growing Trends	26
3.3 Preprocessing Variants	29
3.5 Model Configuration	31
3.7 Experimental Platforms and Software Stack	34
4.1a Header Field F-Scores	35
4.1b Line Items F1-Score	36
4.1c Micro-Averaged Performance	37
4.2 Result from OCR Baseline and Multimodal Approach	38
B.1 Gemini's Per-Field Scores with Aggressive Preprocessing	50
B.2 Gemini's Per-Field Scores with Raw Images	50
B.3 PP-OCRv3's Per-Field Scores with Raw Images	50
B.4 PP-OCRv3's Per-Field Scores with Moderate Preprocessing	50
B.5 Qwen 2.5-VL's Per-Field Scores with Aggressive	50
B.6 Mistral OCR and Pixtral's Per-Field Scores with Aggressive Preprocessing	51