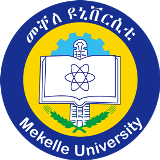
** Comprehensive OCR Error Analysis **

**For Tigrigna  
Using Google Vision, Tesseract, DeepSeek, and Meta-Based OCR Models**

Mekelle University

Mekelle Institute of Technology

Department of Computer Science and Engineering

Name: 1. Leake Abrha gebremedhin

2. Zeru Gebrekidan Grmay

Supervisors: 1. Asmelash Teka Hadgu (Ph.D)

2. Yaecob Girmay (Researcher)

3. Mussie Abrehaley (Lecturer)

Duration: October 2025–January 2026

***Sponsored By Lesan AI***

**Contents**

[1. Executive Summary 1](#_heading=h.ewuvet1dmqsj)

[2. Problem Definition 2](#_heading=h.7sn8uhnks6to)

[2.1 What are you trying to do? 2](#_heading=h.bi0yg1i0iiwi)

[2.2 Why is this problem important? 2](#_heading=h.e5bmuvaydt5h)

[3. Background and Current Practice 3](#_heading=h.eynjjc56sr16)

[3.1 How is it done today? 3](#_heading=h.ksrigo5or88e)

[3.2 Limits of Current Practice 3](#_heading=h.2msut23e6ebf)

[4.](#_heading=h.waiojvmmfsh8) Proposed Approach 3

[4.1 What is new in this approach? 4](#_heading=h.dycg3n3f8k6u)

[4.2 Why Will It Be Successful? 4](#_heading=h.3ak4aqszgsbj)

[5.](#_heading=h.kukcqq3z57v0) Research Plan (3-Month Plan) 4

[5.1 Methodology 4](#_heading=h.v2zrrlnljp8e)

[5.2 Error Analysis Framework 5](#_heading=h.qfmpkge6nlg1)

[6. Risks and Mitigation 5](#_heading=h.nxmc151ewdpp)

[7. Expected Outcomes 5](#_heading=h.vcm8sisikwmm)

[7.1. Scientific or Technical Contributions 5](#_heading=h.hgllgwt12vyg)

[7.2. Impact 6](#_heading=h.pqzq0iane75k)

[8. Evaluation Plan 6](#_heading=h.i2vdf57wq677)

[9. Resources Needed 6](#_heading=h.z5l8qsxmvx29)

[10. Budget 6](#_heading=h.789zfgs4x6ig)

[11. References 7](#_heading=h.13wmeexvbn4n)

# 

# 

# 

# 1. Executive Summary

This research project presents a rigorous and comprehensive Optical Character Recognition (OCR) error analysis focused exclusively on Tigrinya language. Despite major developments in OCR technologies, languages written in the Ethiopic script remain underrepresented in modern machine learning research [[1]](#bookmark=id.dof6ppnlzw49), [[2]](#bookmark=id.cm3ird3qif73), [[3]**,**](#bookmark=id.evhm9xtc31of)[[4]](#bookmark=id.8x330cwdwlad). Consequently, existing OCR systems frequently exhibit high error rates when applied to Tigrigna text, especially on manuscripts, historical documents, low-quality scans, or texts with complex diacritics [[2],](#bookmark=id.cm3ird3qif73) [[5]](#bookmark=id.oi9f1jgwxapi), [[6]](#bookmark=id.vktfc1vnpfw0).

This study evaluates two major OCR systems - Google Cloud Vision, and Meta open-source OCR - across a diverse manually collected dataset including printed books, scanned PDFs, historical manuscripts, exam papers, biographies, newspapers, health books, and 104.4 FM Radio Mekelle. The project will produce Tigrigna-specific OCR benchmark, a detailed error taxonomy, and empirically grounded recommendations for improving OCR accuracy in low-resource African languages. The outcome will support future research efforts, assist in digitizing cultural heritage, and contribute to the development of more inclusive OCR technologies.

# 2. Problem Definition

Current OCR systems struggle to deliver reliable performance on Tigrigna texts due to the unique properties of the Ethiopic script, such as complex diacritics, visually similar glyphs, and rich morphological variations [[2]](#bookmark=id.cm3ird3qif73), [[3]](#bookmark=id.evhm9xtc31of), [[5]](#bookmark=id.oi9f1jgwxapi). Existing OCR tools are not optimized for Tigrinya or other Ethiopic-derived languages [[7],](#bookmark=id.2652f0t9x52x) resulting in high character error rates, misclassification of glyphs, incorrect segmentation, and difficulty handling degraded or historical documents [[5]](#bookmark=id.oi9f1jgwxapi), [[6].](#bookmark=id.vktfc1vnpfw0) The core objective of this project is to systematically identify, analyze, and categorize errors produced by the two selected OCR systems across different Tigrigna text types - to reveal performance limitations and generate a clear, data-driven understanding of Tigrigna OCR challenges.

**Research Questions:**

1. Which error types occur most frequently when processing different categories of Tigrigna text using the selected OCR systems?
2. How do Google Cloud Vision, and Meta open-source OCR models differ in accuracy and robustness across diverse Tigrigna datasets?
3. Which error categories most significantly contribute to OCR performance failures?

## 2.1 What are you trying to do?

The core problem is that current commercial and open-source OCR tools, designed primarily for Latin or high-resource scripts, perform poorly on the Ethiopic script like Tigrigna [[3]](#bookmark=id.evhm9xtc31of), [[7]](#bookmark=id.2652f0t9x52x). Our goal is to analyze how well the two different OCR systems perform on Tigrigna text, determine where they fail, and lay the foundation for future improvements. This project therefore conducts quantitative evaluations, classifies error types, analyzes misrecognition patterns, and comparing system performance using standard OCR metrics. The analysis aims to uncover underlying causes of recognition failures and inform future improvements to Tigrigna OCR technologies.

## 2.2 Why is this problem important?

Tigrigna, spoken by millions, remains severely underrepresented in Natural Language Processing (NLP) research [[8](#bookmark=id.qtfn39anclci)] and lacks robust digital support despite its cultural and historical importance. The absence of well-developed OCR systems for Ethiopic languages limits the digitization and accessibility of vast collections of Tigrigna texts, including historical and cultural documents [[5]](#bookmark=id.oi9f1jgwxapi). Improving OCR for Tigrigna plays a vital role in digitizing books, enhancing accessibility, enabling searchable archives, supporting linguistic research, preserving cultural heritage, integrating the language into modern digital systems, automating document processing and also is essential to prevent loss from physical damage like fire, flooding, or dust [[2]](#bookmark=id.cm3ird3qif73), [[9]](#bookmark=id.nnm26ytkz4fa).

# 3. Background and Current Practice

## 3.1 How is it done today?

Current OCR systems rely on deep learning architectures such as CNNs, RNNs, LSTMs, and Transformers [[9]](#bookmark=id.nnm26ytkz4fa), [[10]](#bookmark=id.7wt6wactpy94), [[11].](#bookmark=id.vtau95ng80tj) While OCR for high-resource languages has matured substantially, low-resource languages - including Tigrigna - lack optimized solutions. Google Vision provides general OCR capabilities, and TrOCR represents a state-of-the-art transformer-based OCR model [[5]](#bookmark=id.oi9f1jgwxapi), [[12]](#bookmark=id.e006iflvzund). Although prior work exists for Amharic, Tigrigna receives comparatively limited attention and lacks specialized evaluation resources [[7].](#bookmark=id.2652f0t9x52x)

In Ethiopia, the integration of OCR into institutional workflows remains extremely limited. Hospitals, municipalities, courts, universities, and government bureaus still depend on manual data entry and paper-based documentation. Common practices involve manually transcribing printed or handwritten records, scanning pages into image or PDF formats without applying text extraction, and storing large collections of physical documents [[2]](#bookmark=id.cm3ird3qif73) with minimal digitization. Existing scanning tools generally produce non-searchable image-based files, and there is little to no adoption of OCR technologies for Ethiopian languages. As a result, digitization processes across sectors remain slow, labor-intensive, error-prone, and difficult to scale.

## 3.2 Limits of Current Practice

Effective OCR systems have primarily been developed for high-resource, Latin-based languages, while the technology for African indigenous scripts, including Tigrigna, is far behind, often lacking functional systems [**[6]**](#bookmark=id.vktfc1vnpfw0), [**[10]**](#bookmark=id.7wt6wactpy94).

Current OCR systems are limited by:

* Scarcity of curated datasets and systematic error analysis
* Frequent confusion between diacritic variations within the same consonant family [[2]](#bookmark=id.cm3ird3qif73), [[5]](#bookmark=id.oi9f1jgwxapi).
* Poor performance on degraded or historical documents
* Incorrect word segmentation and spacing.
* Lack of standardized benchmark datasets [**[1]**](#bookmark=id.dof6ppnlzw49)
* Limited attention to script-specific linguistic features

# Proposed Approach

This project evaluates OCR outputs across printed, scanned, and manuscript texts. It introduces a structured error taxonomy that captures the unique error types found in Tigrigna OCR. By systematically comparing the performance of Google Vision, and Meta-based models, this research maps the technological strengths and weaknesses. Quantitative evaluation using metrics such as Character Error Rate and Word Error Rate, along with linguistic analysis of error categories [[11].](#bookmark=id.vtau95ng80tj)

## 4.1 What is new in this approach?

To the best of our knowledge, no prior work has conducted a systematic, comparative error analysis exclusively focused on Tigrigna. This study introduces a structured, language-specific error taxonomy and evaluates two OCR paradigms. It integrates OCR error analysis techniques, including confusion matrix analysis, character-level error classification (substitutions, insertions, and deletions), and the use of standardized OCR evaluation frameworks [[11]](#bookmark=id.vtau95ng80tj) such as ocrevalUAtion and Dingle-hopper for comprehensive accuracy reporting. Additionally, an optional exploratory fine-tuning step will be conducted to assess whether targeted improvements can address the weaknesses identified through the error analysis.

## 4.2 Why Will It Be Successful?

This study leverages a diverse dataset representing real-world Tigrigna text, ensuring that the analysis captures practical OCR challenges. By integrating quantitative evaluation metrics with linguistically informed error categorization, the study offers both depth and methodological rigor. This combination, along with the exploratory model improvement stage, ensure meaningful and impactful results.

# Research Plan (3-Month Plan)

**Setup & Data Collection: 5 weeks**

Key Tasks: API access setup, Collect and organize dataset, Conduct ground-truth transcription, perform image preprocessing: noise removal, contrast enhancement, deskewing, segmentation, Conduct literature review, and evaluation metrics selection.

**Baseline Testing & Initial Error Analysis: 5 weeks**

Key Tasks: Evaluate the selected OCR systems on the dataset, Compute CER, WER, and edit-distance metrics, Begin manual error annotation, Develop initial error taxonomy.

**Deep Analysis & Reporting: 3 weeks**

Key Tasks: Full error classification, statistical analysis—cross-system performance comparison, visualization, final report, recommendations, and presentation.

## 5.1 Methodology

Data sources**:** include printed books, images and scanned PDFs such as መፅሓፍ ቅዱስ - ሓዲስ ኪዳን & ብሉይ ኪዳን፣ ትፅቢት ባህጉ፥ ትግራይ ትሕቲ ሰማይ፣ ገድሊ ኣባ፣ ሓራ ዉፃእ፣ ህያው ደብሪ፣ መንፈሳዊ ህይወት ኣብ ግብሪ፣ ምኽሪ ንፈታዊ፥ ብሄራዊ ፈተና ስነ ሂወት 8ይ ክፍሊ፣ መፅሃፍ ተምሃራይ ሒሳብ 7ይን 8ይን ክፍሊ፣ ወይን ጋዜጣ፣FM 104.4፣ Tigrigna MEDICAL HEP-C, and others.

Preprocessing**:** noise reduction, binarization, contrast adjustment, cropping, de-skewing, resolution normalization, segmentation, and conversion from PDF to images.

Annotation Process**:** Multiple annotators will verify ground-truth transcriptions, with periodic agreement checks to reduce labeling inconsistencies.

OCR outputscome from Google Vision, and Meta open-source models.

Evaluation Metrics**:** Character Error Rate (CER), Word Error Rate (WER), and edit distance.

Errors are categorized**:** into diacritic errors, glyph confusion/Similar glyph misclassification, word/phrase segmentation errors, layout/column detection errors, and Hyphenation/punctuation errors.

## 5.2 Error Analysis Framework

A structured error taxonomy will be developed based on sequence alignment (Levenshtein distance) and linguistic analysis, and structure of Ethiopic characters. Major categories include:

* Base consonant confusion between distinct characters (e.g., ቨ → በ).
* Vowel-order confusions within the same base consonant (e.g., ሰ → ሱ).
* Errors in punctuation and diacritics
* Word and character segmentation failures
* Layout and formatting errors
* Insertions, deletions, substitutions in recognized output

NB: All data used in this study will be processed exclusively for academic research purposes, following fair-use principles and respecting the intellectual property rights of the original content creators.

# 6. Risks and Mitigation

The following risks are manageable and expected for OCR research involving low-resource scripts.

|  |  |  |
| --- | --- | --- |
| **S/N** | **Risks** | **Mitigations** |
| 1 | Degraded scans reducing OCR accuracy | Apply stronger image preprocessing techniques to enhance input quality. |
| 2 | Time-intensive annotation process | Use selective sampling to ensure representative and diverse data. |
| 3 | API cost and rate limitations | Utilize GPU resources (Colab) for efficient experimentation |
| 4 | Potentially small improvements from fine-tuning | Incorporate multiple annotators to reduce bias |
| 5 | Ground-truth transcription inconsistencies | Maintain verification checkpoints during annotation |

# 7. Expected Outcomes

## 7.1. Scientific or Technical Contributions

* A curated and annotated Tigrigna OCR benchmark dataset
* A structured and reproducible Tigrigna-specific OCR error taxonomy
* A comprehensive comparison of Google Vision, and Meta open-source
* Detailed recommendations for improving Tigrigna OCR
* New insights into OCR challenges for low-resource African languages

## 7.2. Impact

* Baseline for future academic OCR research on Ethiopic-script languages
* Foundation for future post-processing and fine-tuning efforts
* Local Language Communities: Enables the efficient and accurate digitization of Tigrigna cultural heritage, historical documents, and official records, increasing information accessibility.
* Contribute valuable resources to open-source African-language technology ecosystems.
* Guide for future dataset collection efforts for low-resource languages.
* Provide actionable guidance for OCR API developers.
* Enhances automation and accessibility in document processing.

# 8. Evaluation Plan

**Midterm**:   
- Dataset prepared (dataset with ground-truth transcriptions)

- Metadata ready  
- Baseline OCR results for systems  
- Initial analyses (Initial CER/WER calculations)  
**Final**:   
- Complete error analysis with classified error types

- Visualizations and statistical comparisons  
- Finalized error taxonomy  
- Submission of full report, presentation.

# 9. Resources Needed

* Data Sources: Digitized images, scanned PDFs, manuscripts, printed materials
* Compute**:** Access to high-performance GPUs for TrOCR fine-tuning, as large transformer models require significant compute [[12]](#bookmark=id.e006iflvzund).
* Storage: Sufficient cloud or local storage for the image dataset
* Software Tools: Python, PyTorch, TensorFlow, Tesseract, OpenCV, Pillow, HuggingFace Transformers, pdf2image, pymupdf, ocrevalUAtion toolkit, jiwer, pandas, matplotlib/seaborn
* APIs: Google Cloud Vision APIs.
* Annotation Tools: Image-to-text annotation interface and PDF processing tools

# 10. Budget

* Human effort for data annotation
* Access to a high-quality scanner
* API usage for commercial OCR services.

# 11. References

[1] Henok Biadlign Ademtew and Mikiyas Girma Birbo*, “*AGE: Amharic, Ge'ez and English Parallel Dataset*,*” *5th Workshop on African Natural Language processing,* 2024.

[2] Kibrom Tsegay Fitsum, Yask Patel, "Optical character recogniation for Tigrigna Printed Documents using HOG and SVM," *in Proc. 2nd Int. Conf. Inventive Communication and Computational Technologies (ICICCT)*, 2018.

[3] Y. Assabie, & J. Bigun, “Ethiopic character recognition using direction field tensor,” *In Proc.18th International Conference on Pattern Recognition,* pp. 284-287, 2006.

[4] H. Aaron, H. Abiel, G. Danait, H. Robel, G. Tesfana, and T. Yemane, “Tigrinya OCR: Applying CRNN for Text Recognition,” 2023.

[5] Kasaye Akanie Guangul, Dagne Walle Girmaw, Million Meshesha, "Character recognition of ancient ethiopic Ge’ez manuscripts using deep convolutional neural networks," *Discover Imaging, vol. 2, no. 5*, 2025.

[6] M. Meshesha and C. V. Jawahar, “Optical Character Recognition of Amharic Documents,” *IIIT Hyderabad, India*, 2007.

[7] B. H. Belay, “Deep learning for Amharic text-image recognition: Algorithm, dataset and application,” *Ph.D. dissertation, Addis Ababa University*, 2021.

[8] Fitsum Gaim, Jong C. Park, "Natural Language Processing for Tigrinya: Current State and Future Directions," *arXiv:2507.17974v2*, 2025.

[9] H. T. Weldegebriel, J. Chen, and D. Zhang, “Deep Learning for Ethiopian Ge’ez Script Optical Character Recognision,” in *Proc. 2018 Tenth Int. Conf. on Advanced Computational Intelligence (ICACI)*, 2018.

[10] A. Afkari-Fahandari, E. Shabaninia, F. Asadi-Zeydabadi, & H. Nezamabadi-Pour, “A Comprehensive Survey of Transformers in Text Recognition: Techniques, Challenges, and Future Directions,” *Journal of King Saud University - Computer and Information Sciences*, 2024.

[11] R. Jana, A. R. Chowdhury, and M. Islam, “Optical character recognition from text image,” *International Journal of Computer Applications Technology and Research*, 2014.

[12] M. Li et al., “TrOCR: Transformer-based optical character recognition with pre-trained models,” *arXiv preprint*, 2022.