PROJECT REPORT

On

TextEvolve

Preserving History, Empowering Research

Submitted by

Yuva Nandhini M – 711621243059

Praveenkumar S – 711621243036

Sibi Siddharth S – 711621243046

Uma Maheswari P – 711621243053

Under the Guidance of

Kavitha M,

Assistant Professor / AI&DS

Kathir College of Engineering

1. Executive Summary

TextEvolve is an advanced AI-driven platform developed by team Dynamic Dreamers from Kathir College of Engineering. Its mission is to digitize historical handwritten documents into accessible and editable digital formats. By leveraging a dual OCR solution—integrating a custom ML model with commercial OCR APIs such as Google Vision, Azure Vision, and Amazon Textract—the platform ensures high-accuracy text extraction. Enhanced by features like text polishing via Google Gemini API and translation through Google Translate API, TextEvolve not only preserves cultural heritage but also facilitates robust academic research. In the future, plans to establish a "TextEvolve Community" will enable users to share their digitized documents and research insights.

2. Project Overview

The primary goal of TextEvolve is to address the challenges posed by manual transcription of historical documents. Aging physical documents, handwriting variability, and regional language intricacies have long hindered accessibility. TextEvolve overcomes these obstacles by using artificial intelligence and optical character recognition (OCR) technologies that ensure high accuracy, ease of access, and multiple format outputs (Word, PDF, TXT). Designed for researchers, students, and cultural preservationists, this platform is positioned as a complete, end-to-end solution for historical document digitization.

3. Key Features and Innovations

3.1 Dual OCR Approach

Custom OCR Model:

Engineered using machine learning frameworks such as TensorFlow, PyTorch, and Keras.

Specially trained for recognizing handwritten historical documents, particularly those in regional languages and with degradation issues.

Commercial OCR APIs:

Google Vision API: Reliable text extraction from diverse document layouts.

Azure Vision OCR API: Provides scalable OCR services with robust error handling.

Amazon Textract: Capable of extracting detailed information even from complex document formats.

This dual approach is designed to cross-verify and maximize text extraction accuracy, ensuring that even the most challenging handwritten texts are processed effectively.

3.2 Enhanced Text Processing

Google Gemini API Integration: Improves the quality of extracted text through automatic grammar, spelling, and formatting corrections.

Delivers refined, publication-ready text with minimal manual intervention.

3.3 Translation Capabilities

Google Translate API: Translates the extracted content into various regional languages.

Enhances accessibility and allows users to view historical documents in their native language for deeper understanding.

3.4 Multi-Format Export

Output Flexibility: Generates digitized documents in multiple formats such as Word, PDF, and TXT.

All outputs are editable, making it easier for further annotation and research.

3.5 Interactive Research and Query Features

Query and Insights Interface: Users can pose questions on scanned documents to get summaries and in-depth insights.

Designed to serve educational and research communities, this feature accelerates data discovery and scholarly analysis.

3.6 Future Expansion: TextEvolve Community

Community Platform: A forthcoming feature that will let users publish their scanned documents along with personal insights and research findings.

Fosters a collaborative environment where researchers, scholars, and students can share knowledge and discuss interpretations of historical texts.

4. Technology Stack

4.1 Programming Languages

Python: Backbone for developing machine learning models and backend services.

JavaScript:

Used for building dynamic, responsive frontend components as well as server-side scripts.

4.2 Databases

MongoDB: Employed for scalable document and data storage.

Prisma: Utilized as an ORM to manage and interface between the application and databases efficiently.

4.3 Deployment

Hostinger KVM 2 - VPS: Provides reliable virtualized hosting to support the application’s scalability and real-time processing requirements.

4.4 Frameworks and Libraries

Backend and APIs:

Flask (Python): Lightweight framework for building the backend services.

Node.js: Serves as the runtime for executing JavaScript on the server-side.

Frontend:

React.js: For creating dynamic user interfaces.

Vite: A modern build tool that enhances frontend development.

Tailwind CSS: For fast and efficient UI styling.

Machine Learning:

TensorFlow, PyTorch, Keras: Fundamental for developing and deploying the custom OCR and text processing models.

4.5 Development Tools

Visual Studio Code (VSCode): Integrated development environment (IDE) for efficient coding and debugging.

Figma: Used for UI/UX design.

Git & GitHub: Version control and collaboration tools.

Postman: For testing and validating API endpoints.

npm: Package manager supporting JavaScript development.

Adobe Illustrator & Canva: Tools for designing graphics and other visual assets.

5. System Architecture & Workflow

5.1 Workflow Overview

Document Ingestion: Scanned physical documents are uploaded to the system.

Pre-Processing: Image enhancement techniques are applied to improve quality before OCR.

OCR Conversion: Initial extraction is performed via leading OCR APIs.

Custom OCR model refines the text extraction for higher accuracy.

Text Enhancement & Translation: Google Gemini API polishes the extracted text. Google Translate API converts text into multiple languages.

Output Generation: Digitized documents are provided in editable formats (Word, PDF, TXT).

Query and Insights Interface: Users can ask for summaries or deeper insights into the text.

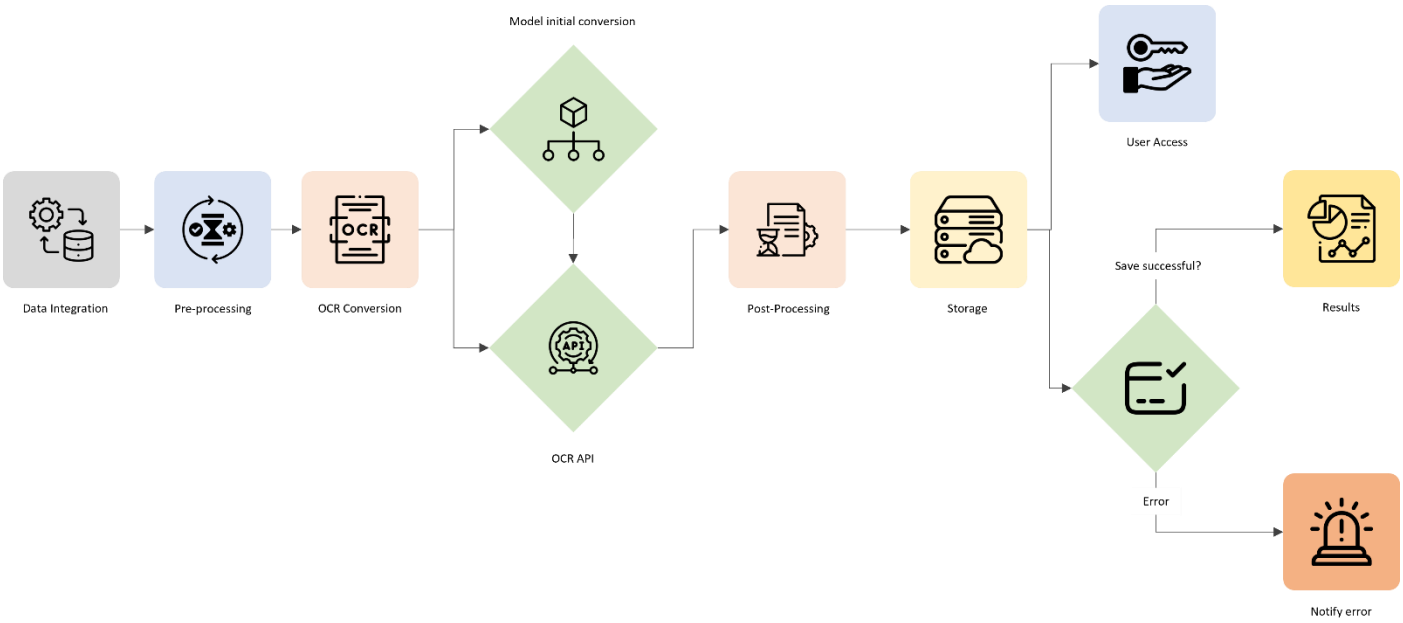
Future Community Platform: Plans to include a feature where users can publish and collaborate on digitized historical documents.

5.2 Infrastructure Overview

Cloud Deployment: The system is hosted on Hostinger KVM 2 - VPS, ensuring high availability and scalability.

API Integrations and Custom Models: Combines commercial OCR services with custom ML models for robust text extraction.

Real-Time Processing: The workflow enables progress tracking and timely notifications, ensuring a smooth user experience.



6. Impact and Benefits

6.1 Cultural Preservation

Preserving Historical Documents: Digitizes fragile, historical, and handwritten documents to save them from physical decay.

Maintaining Regional Heritage: Supports extraction and translation of documents in regional languages, preserving cultural nuances.

6.2 Academic Research and Education

Enhanced Research Capabilities: Provides editable, searchable texts that facilitate detailed study and academic research.

Interactive Learning: The query interface aids students and researchers in quickly summarizing and extracting insights from historical documents.

6.3 Collaborative and Community Initiatives

TextEvolve Community (Future): A dedicated platform for users to share, discuss, and expand on digitized content.

Collaborative Research: Encourages cross-disciplinary studies by offering a common platform for scholarly exchange.

7. Future Directions

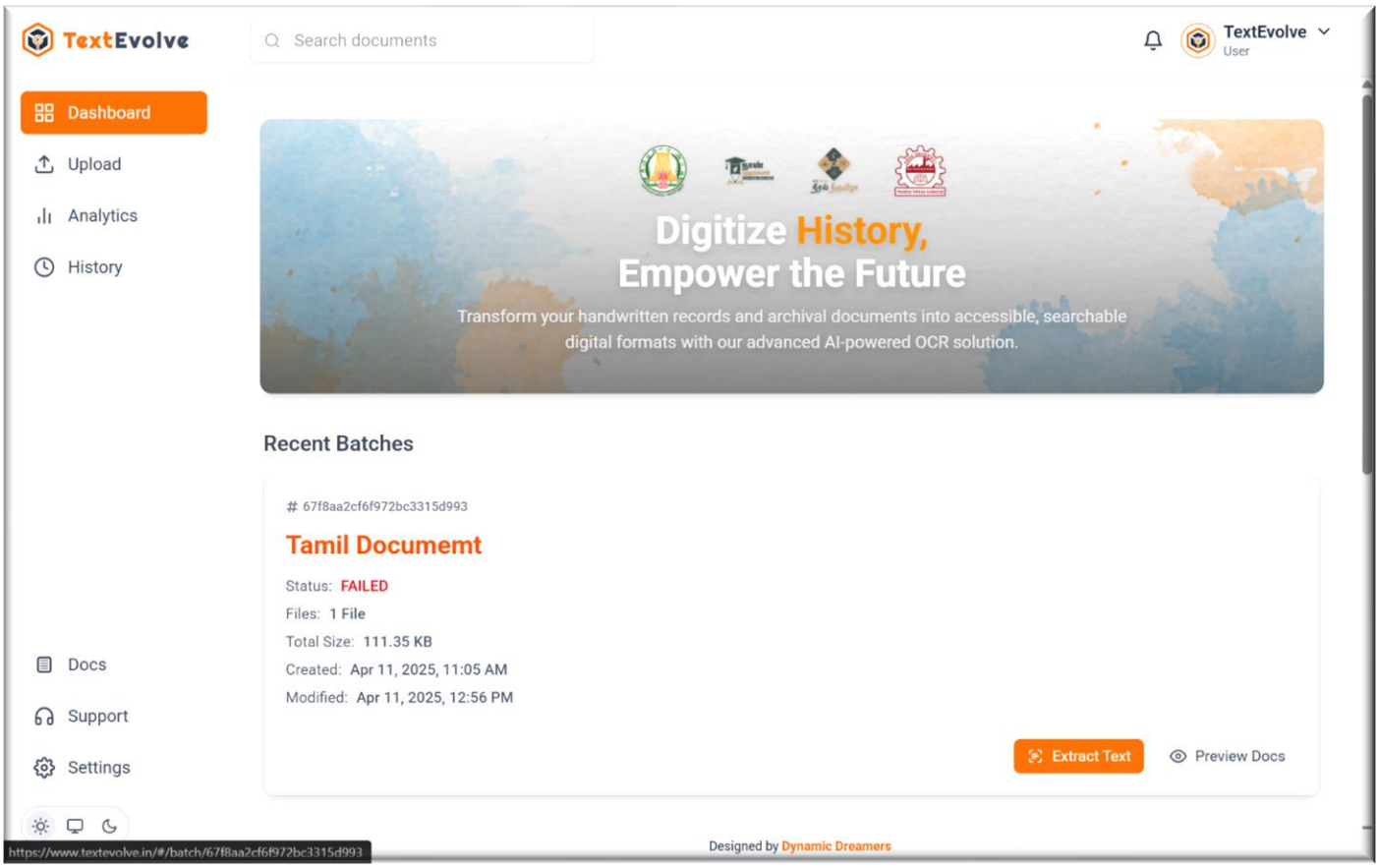
Model Enhancement and Expansion: Ongoing improvements in machine learning models to support a broader range of handwriting and additional regional languages.

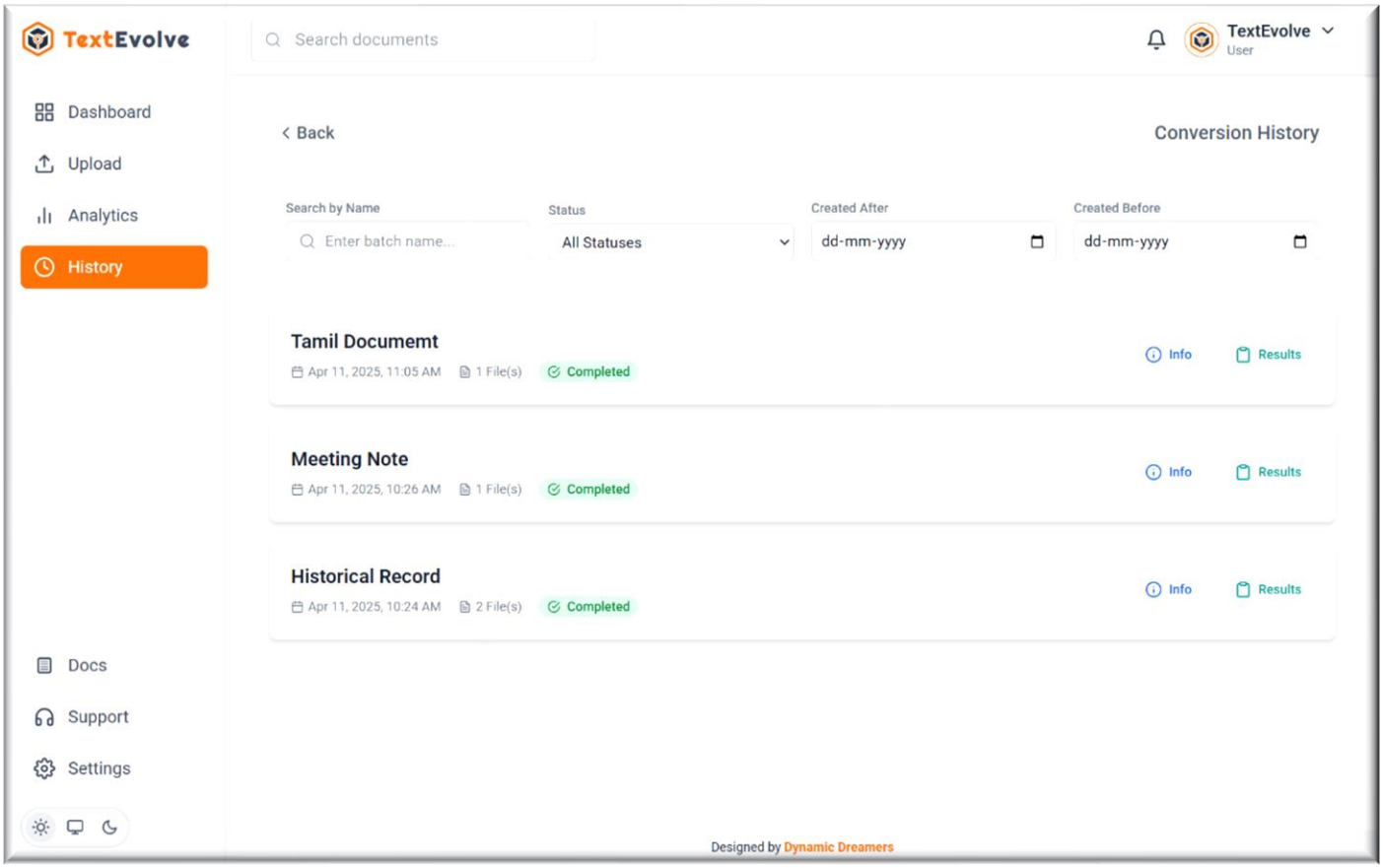
User Experience Improvements: Refinement of the user interface for even greater usability and accessibility.

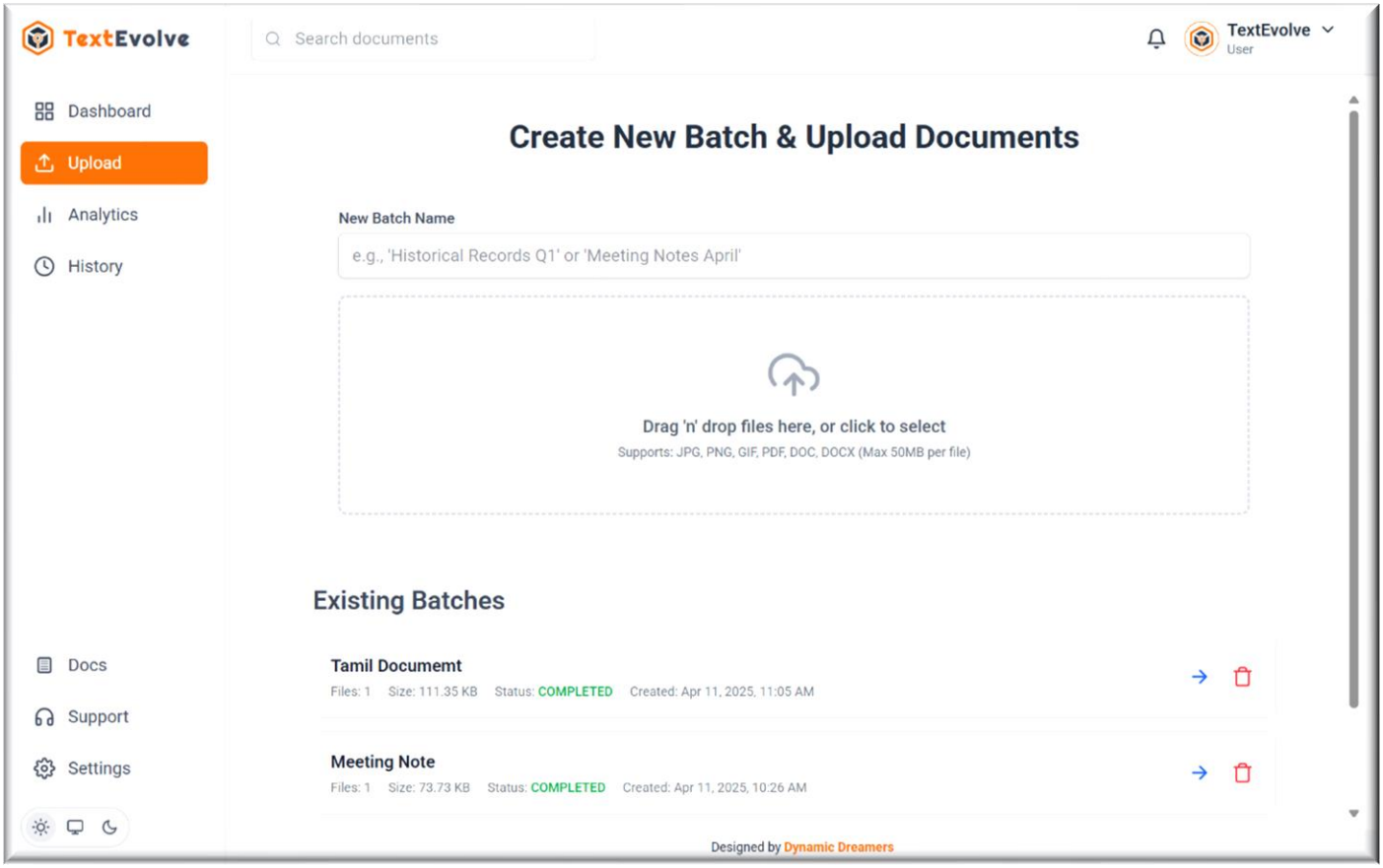
Community-Driven Features: Development of the TextEvolve Community to promote collaboration among researchers, students, and cultural custodians.

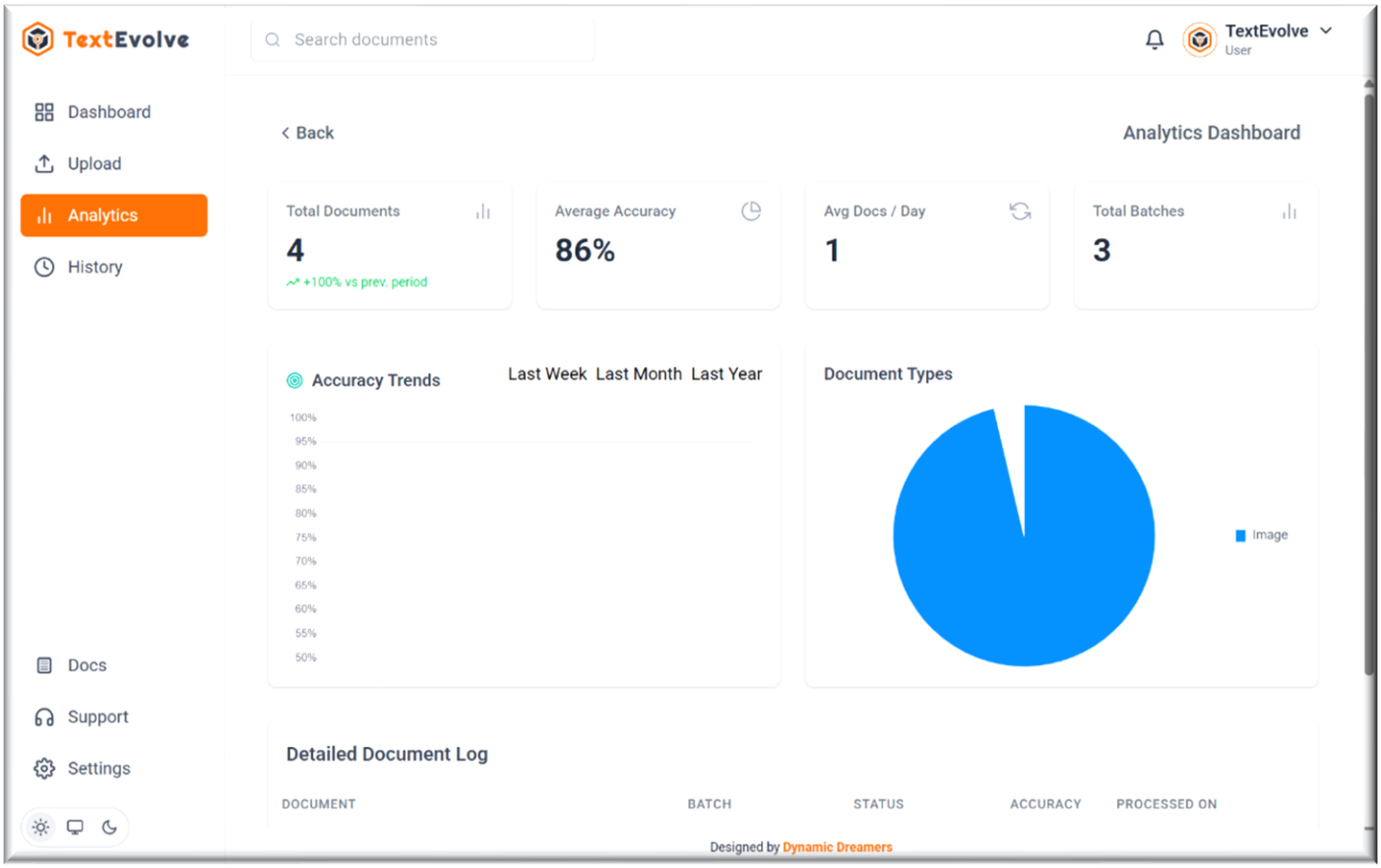
Advanced Query Capabilities:Further enhancements to the interactive query system to provide deeper, more granular insights.

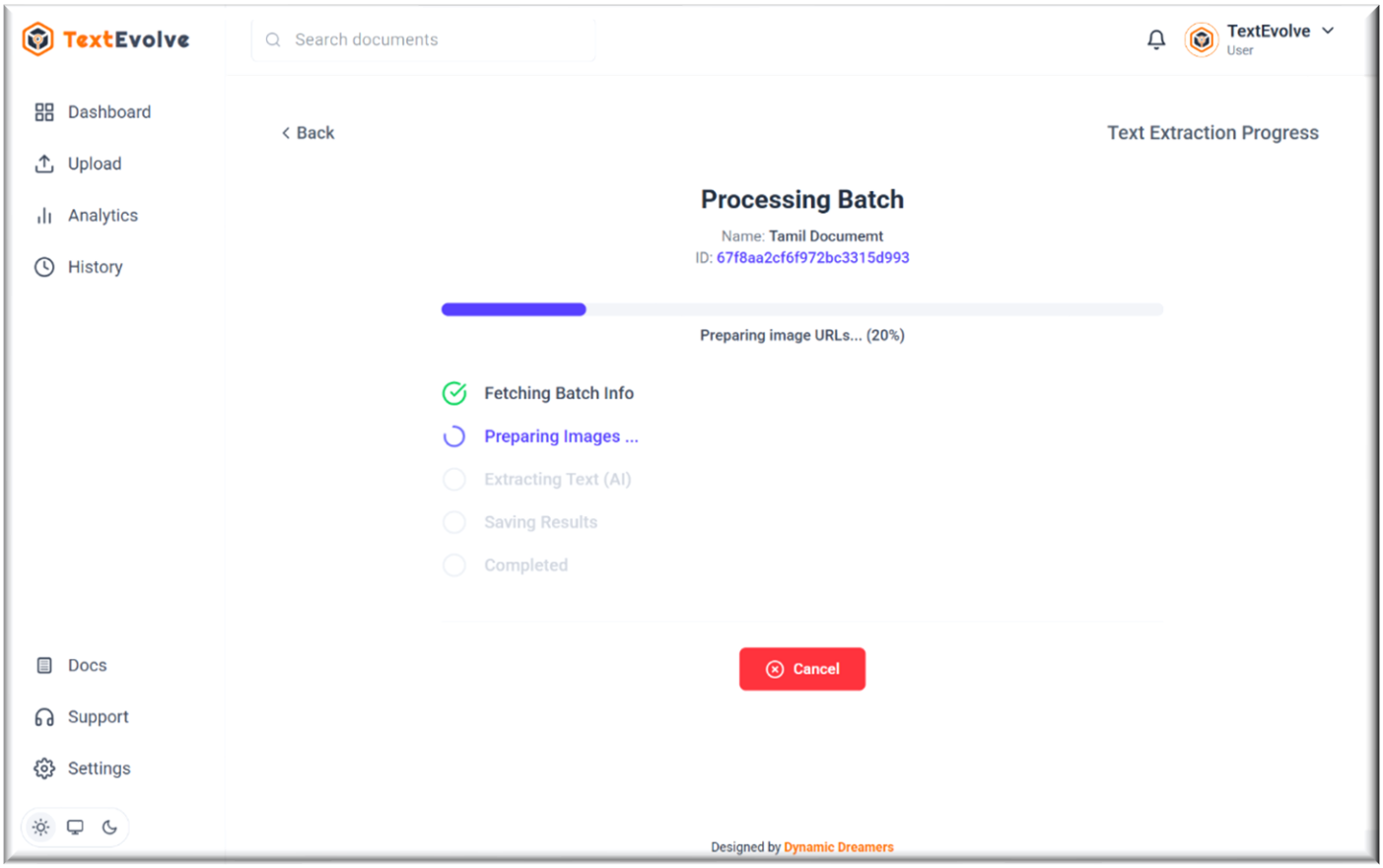
8. Appendix

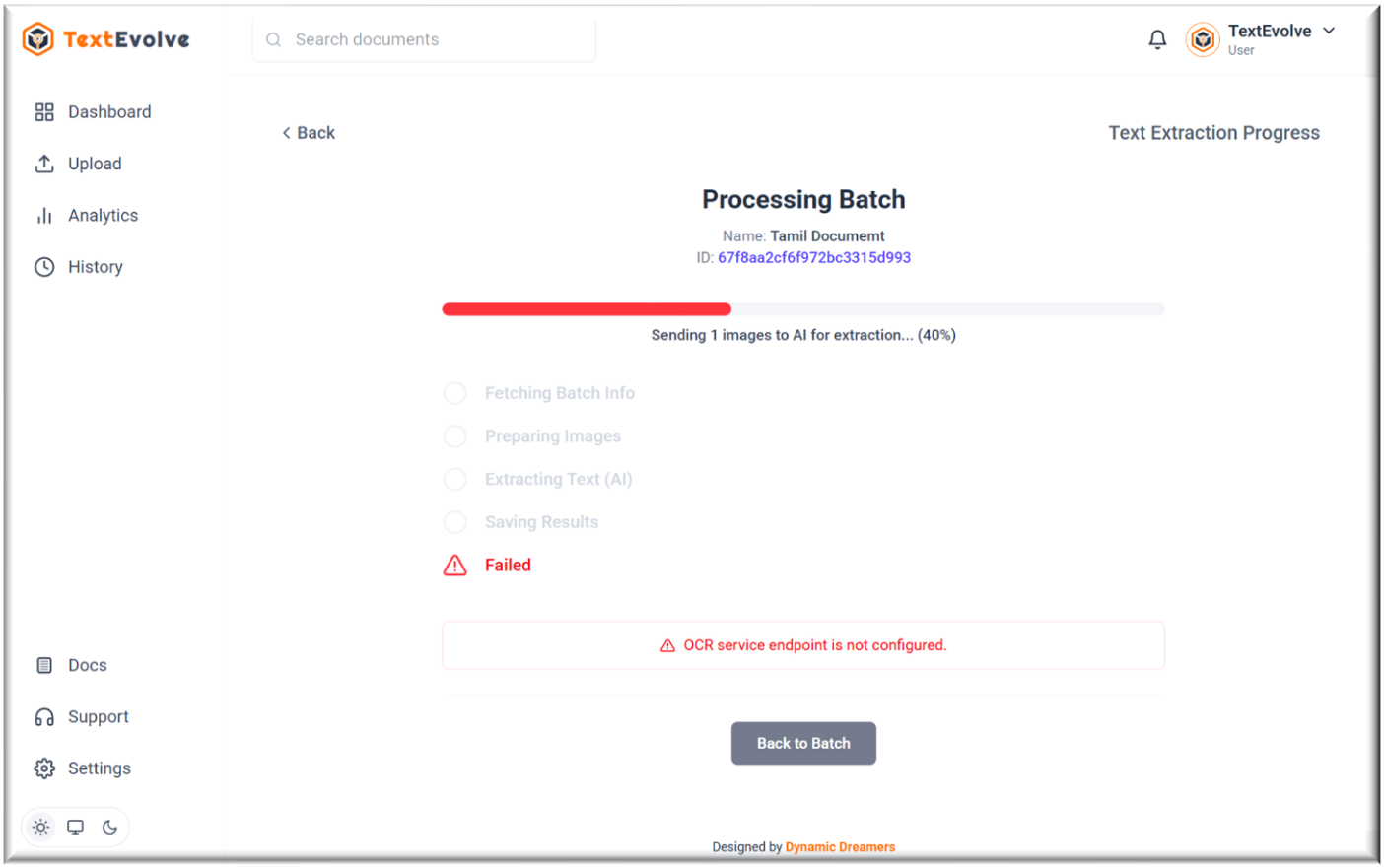


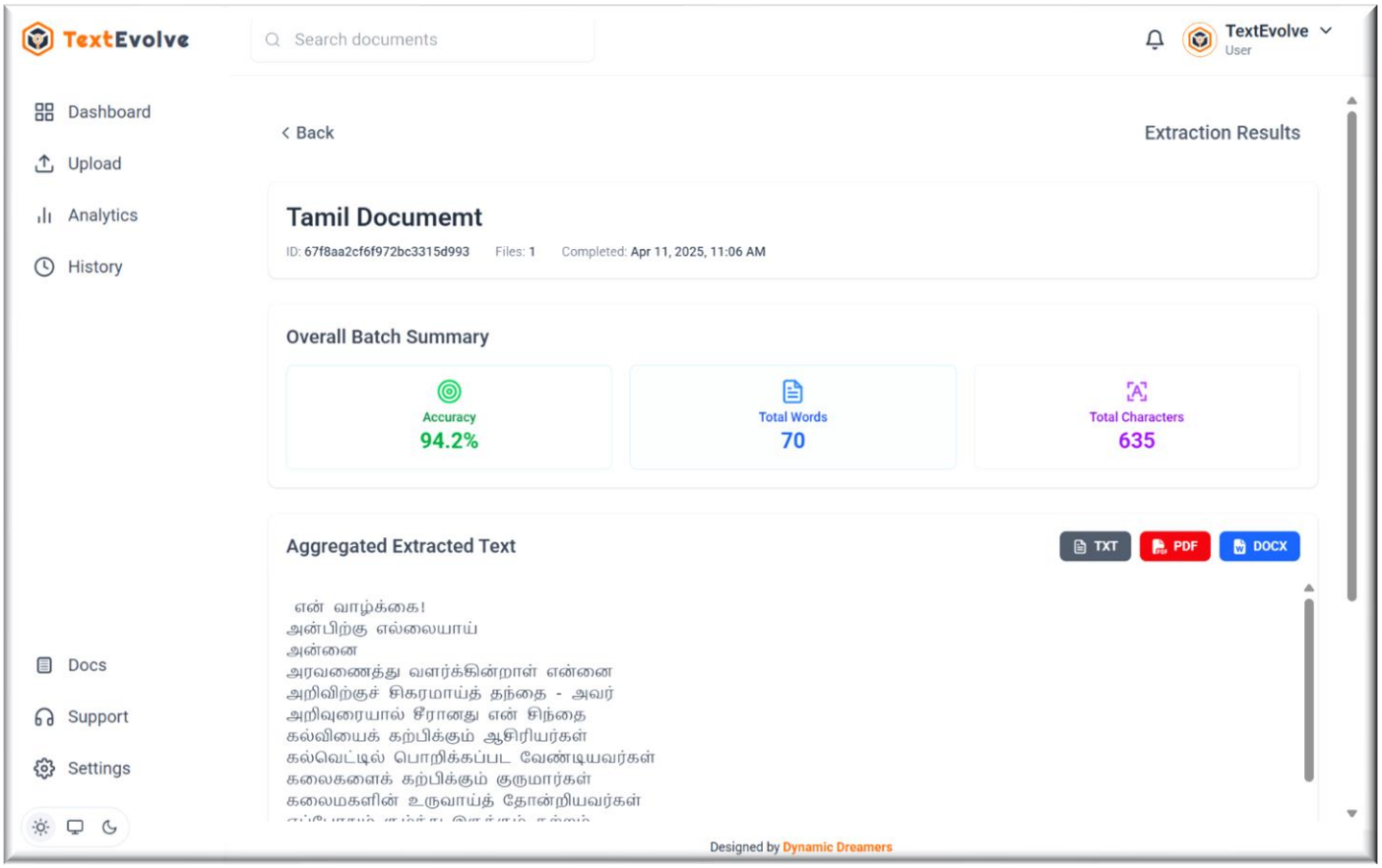

TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
Conversion History
Search by Name
Enter batch name...
Status
All Statuses
Created After
dd-mm-yyyy
Created Before
dd-mm-yyyy
Tamil Documemt
Apr 11, 2025, 11:05 AM
1 File(s)
Completed
Info
Results
Meeting Note
Apr 11, 2025, 10:26 AM
1 File(s)
Completed
Info
Results
Historical Record
Apr 11, 2025, 10:24 AM
2 File(s)
Completed
Info
Results
Docs
Support
Settings
Designed by Dynamic Dreamers

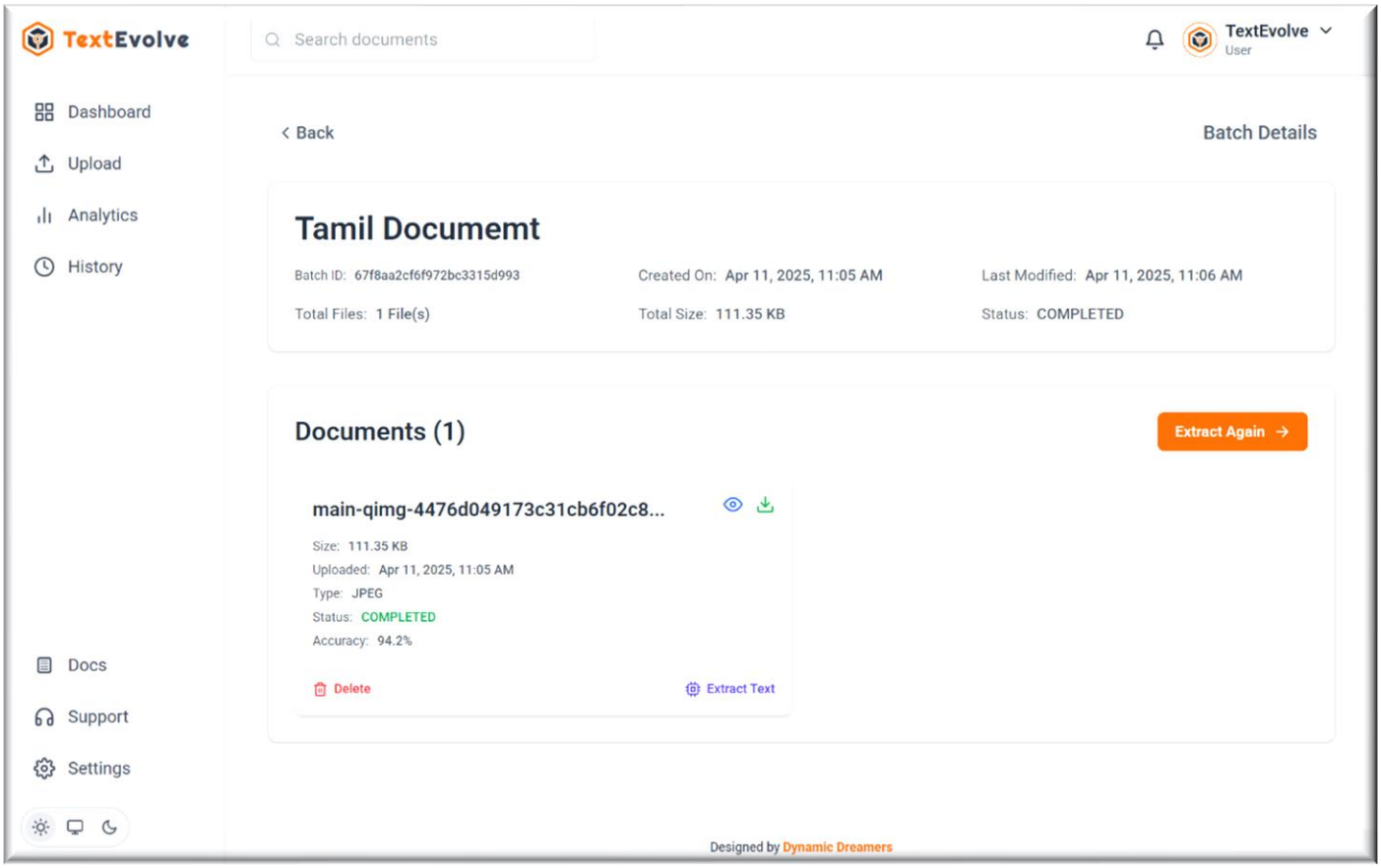

TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Create New Batch & Upload Documents
New Batch Name
e.g., 'Historical Records Q1' or 'Meeting Notes April'
Drag 'n' drop files here, or click to select
Supports: JPG, PNG, GIF, PDF, DOC, DOCX (Max 50MB per file)
Existing Batches
Tamil Documemt
Files: 1 Size: 111.35 KB Status: COMPLETED Created: Apr 11, 2025, 11:05 AM
Meeting Note
Files: 1 Size: 73.73 KB Status: COMPLETED Created: Apr 11, 2025, 10:26 AM
Docs
Support
Settings
Designed by Dynamic Dreamers

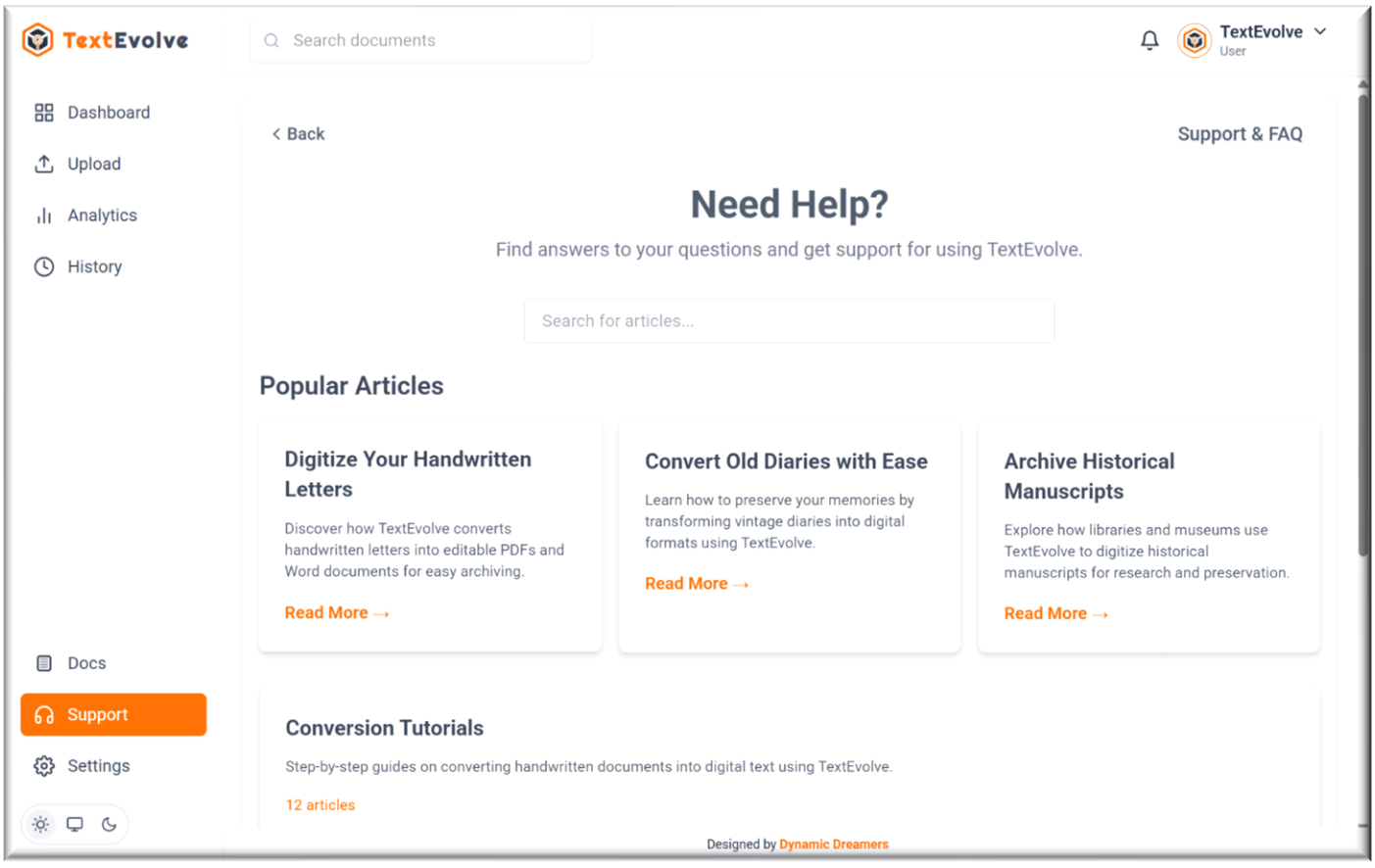

TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
Analytics Dashboard
Total Documents
4
+100% vs prev. period
Average Accuracy
86%
Avg Docs / Day
1
Total Batches
3
Accuracy Trends
Last Week Last Month Last Year
100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
Document Types
Image
Docs
Support
Settings
Detailed Document Log
DOCUMENT
BATCH
STATUS
ACCURACY
PROCESSED ON
Designed by Dynamic Dreamers

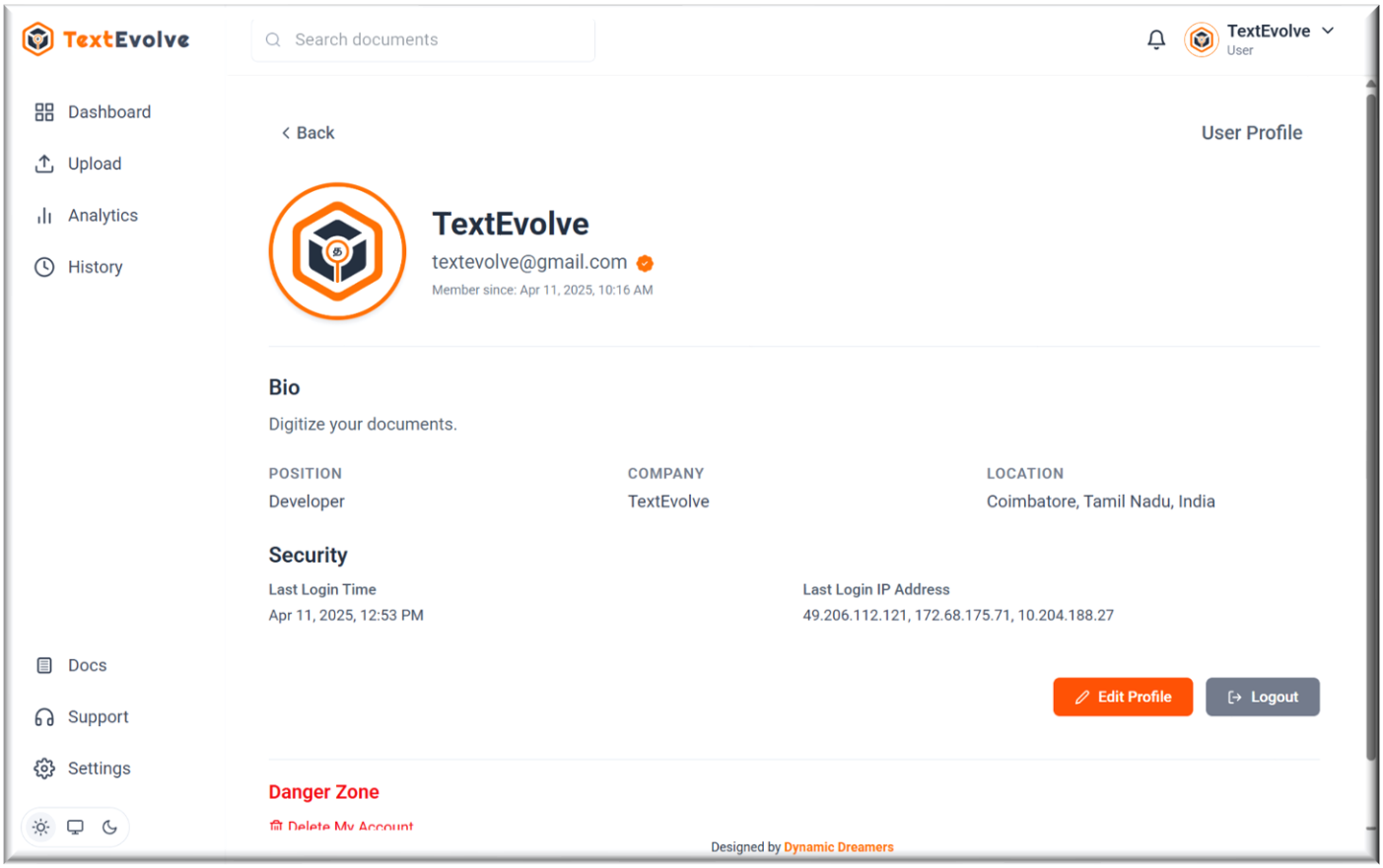

TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
Text Extraction Progress
Processing Batch
Name: Tamil Documemt
ID: 67f8aa2cf6f972bc3315d993
Preparing image URLs... (20%)
Fetching Batch Info
Preparing Images ...
Extracting Text (AI)
Saving Results
Completed
Cancel
Docs
Support
Settings
Designed by Dynamic Dreamers

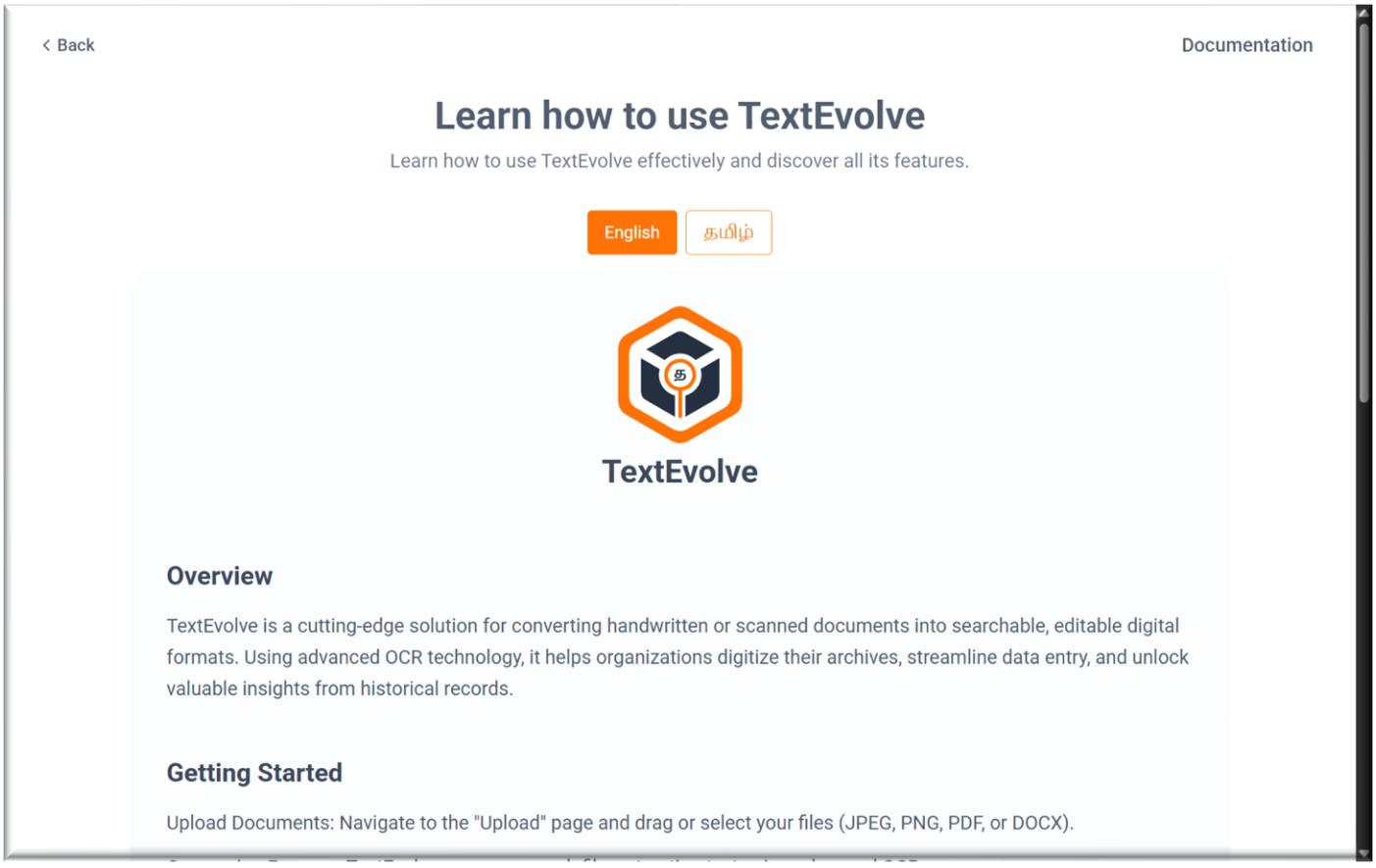

TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
Text Extraction Progress
Processing Batch
Name: Tamil Documemt
ID: 67f8aa2cf6f972bc3315d993
Sending 1 images to AI for extraction... (40%)
Fetching Batch Info
Preparing Images
Extracting Text (AI)
Saving Results
Failed
OCR service endpoint is not configured.
Back to Batch
Docs
Support
Settings
Designed by Dynamic Dreamers

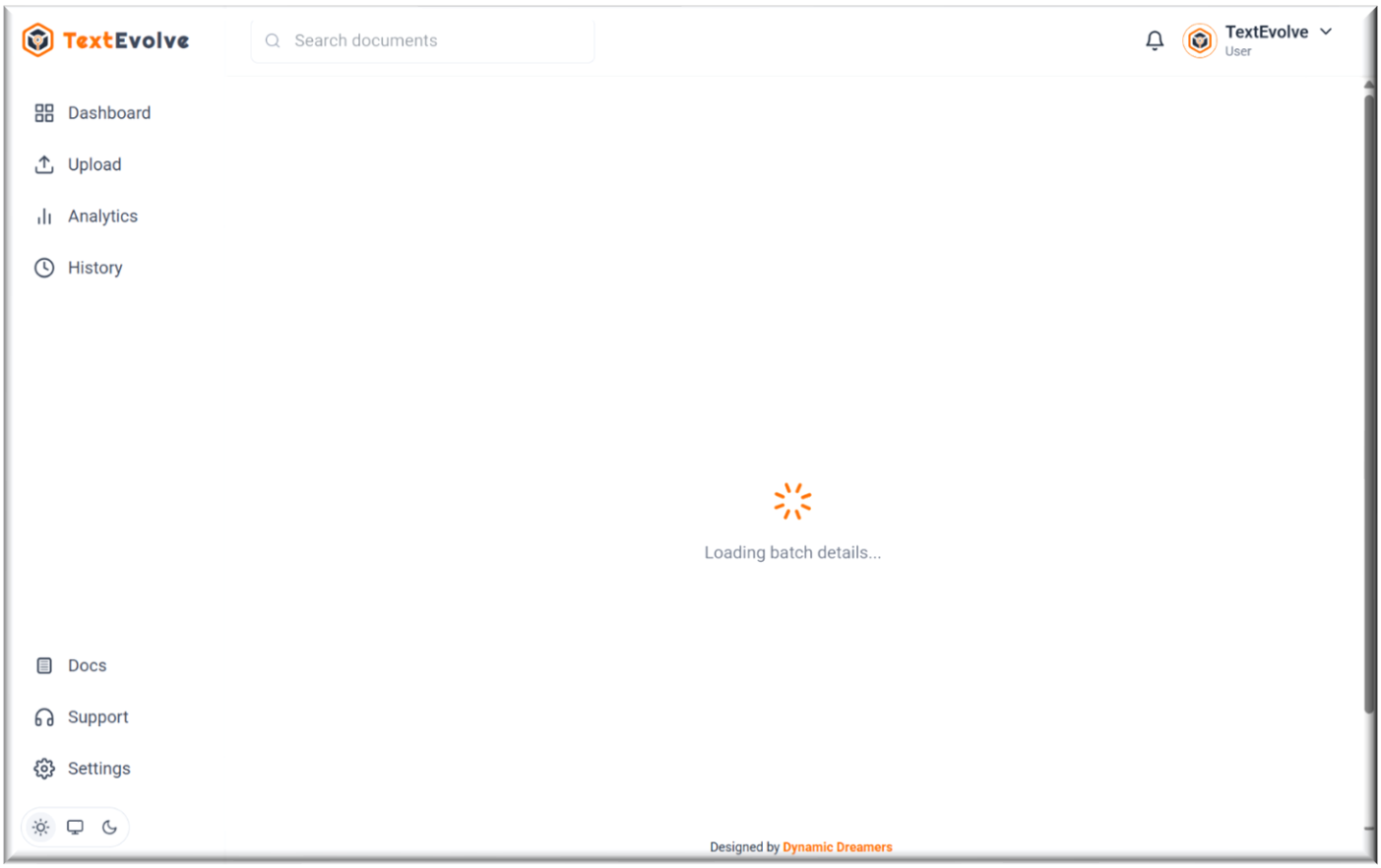

TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
Extraction Results
Tamil Documemt
ID: 67f8aa2cf6f972bc3315d993 Files: 1 Completed: Apr 11, 2025, 11:06 AM
Overall Batch Summary
Accuracy
94.2%
Total Words
70
Total Characters
635
Aggregated Extracted Text
TXT
PDF
DOCX
என் வாழ்க்கை!
அன்பிற்கு எல்லையாய்
அன்னை
அரவணைத்து வளர்க்கின்றாள் என்னை
அறிவிற்குச் சிகரமாய்த் தந்தை - அவர்
அறிவுரையால் சீரானது என் சிந்தை
கல்வியைக் கற்பிக்கும் ஆசிரியர்கள்
கல்வெட்டில் பொறிக்கப்பட வேண்டியவர்கள்
கலைகளைக் கற்பிக்கும் குருமார்கள்
கலைமகளின் உருவாய்த் தோன்றியவர்கள்
Docs
Support
Settings
Designed by Dynamic Dreamers


TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
Batch Details
Tamil Documemt
Batch ID: 67f8aa2cf6f972bc3315d993
Created On: Apr 11, 2025, 11:05 AM
Last Modified: Apr 11, 2025, 11:06 AM
Total Files: 1 File(s)
Total Size: 111.35 KB
Status: COMPLETED
Documents (1)
Extract Again
main-qimg-4476d049173c31cb6f02c8...
Size: 111.35 KB
Uploaded: Apr 11, 2025, 11:05 AM
Type: JPEG
Status: COMPLETED
Accuracy: 94.2%
Delete
Extract Text
Docs
Support
Settings
Designed by Dynamic Dreamers


TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
Support & FAQ
Need Help?
Find answers to your questions and get support for using TextEvolve.
Search for articles...
Popular Articles
Digitize Your Handwritten Letters
Discover how TextEvolve converts handwritten letters into editable PDFs and Word documents for easy archiving.
Read More →
Convert Old Diaries with Ease
Learn how to preserve your memories by transforming vintage diaries into digital formats using TextEvolve.
Read More →
Archive Historical Manuscripts
Explore how libraries and museums use TextEvolve to digitize historical manuscripts for research and preservation.
Read More →
Docs
Support
Settings
Conversion Tutorials
Step-by-step guides on converting handwritten documents into digital text using TextEvolve.
12 articles
Designed by Dynamic Dreamers


TextEvolve
Search documents
TextEvolve
User
Dashboard
Upload
Analytics
History
Back
User Profile
TextEvolve
textevolve@gmail.com
Member since: Apr 11, 2025, 10:16 AM
Bio
Digitize your documents.
POSITION
Developer
COMPANY
TextEvolve
LOCATION
Coimbatore, Tamil Nadu, India
Security
Last Login Time
Apr 11, 2025, 12:53 PM
Last Login IP Address
49.206.112.121, 172.68.175.71, 10.204.188.27
Docs
Support
Settings
Edit Profile
Logout
Danger Zone
Delete My Account
Designed by Dynamic Dreamers


Back
Documentation
Learn how to use TextEvolve
Learn how to use TextEvolve effectively and discover all its features.
English
தமிழ்
TextEvolve
Overview
TextEvolve is a cutting-edge solution for converting handwritten or scanned documents into searchable, editable digital formats. Using advanced OCR technology, it helps organizations digitize their archives, streamline data entry, and unlock valuable insights from historical records.
Getting Started
Upload Documents: Navigate to the "Upload" page and drag or select your files (JPEG, PNG, PDF, or DOCX).



9. Conclusion

TextEvolve is a comprehensive, AI-powered solution tailored to the digitization and preservation of historical handwritten documents. Its dual OCR approach—integrating a custom machine learning model with commercial OCR APIs—ensures superior accuracy in text extraction. Enhanced by post-processing, translation, and multi-format export capabilities, the platform serves as an invaluable tool for cultural preservation and academic research. With future plans to incorporate a community platform and advanced querying features, TextEvolve is well-positioned to become a cornerstone resource for preserving history and advancing research.