

Transforming Tamil Legal Transcriptions - Implementing AI-Powered Handwritten Character Recognition for Tamil Script

Prepared by:

Rajendiran, Parthasarathy
Ketharnathan, Ramsudhan
Khadka, Rohit
Prajapati, Shubham Maheshbhai

Prepared for:

Espinosa Carrillo, David
Conestoga College Institute of Technology and Advanced Learning

Submitted in partial fulfillment of the requirements for the “Case Studies in Artificial Intelligence and Machine Learning” (CSCN8040) course in the Applied Artificial Intelligence & Machine Learning program.

April 15, 2024

Abstract:

In the legal sphere of Tamil Nadu, the laborious task of manually transforming handwritten documents into digital form imposes a hefty toll on efficiency and resource allocation. Legal professionals, on average, dedicate upwards of 32 labor hours per month to the meticulous transcription of essential legal documents, a practice that not only overburdens the legal apparatus but also carries the inherent risk of human error with potentially grave legal ramifications (Madras High Court, n.d.; Advocate, personal communication, February 2024). Currently, this domain is bereft of any automated transcription solutions capable of addressing the intricate nature of handwritten Tamil characters, leading to a pressing call for technological ingenuity.

This paper introduces a novel initiative aimed at mitigating this gap through the development of an innovative AI-powered Handwritten Character Recognition technology for Tamil (HCRT). The proposed HCRT system boasts the capability to accurately decode the myriad of strokes characteristic of Tamil script, effectively halving the time required for transcription while substantially enhancing the reliability of digital records. The integration of this technology into the prevailing legal documentation workflow promises not only a surge in processing efficiency but also a significant elevation in the quality of legal records. This study envisions a fundamental alteration in the operational paradigm of legal practitioners, synchronous with the forward march of the digital age, and posits the HCRT system as a pivotal tool in this transformation.

Introduction:

In the digital era, the legal profession in Tamil Nadu is confronted with a pivotal challenge: the arduous conversion of handwritten documents to digital formats. As the world pivots increasingly towards digitalization, the existing manual methodologies used by legal practitioners to transcribe vital documents, such as wills and land agreements, have become glaringly antiquated. This labor-intensive process not only erodes efficiency but also invites the specter of human error, jeopardizing the integrity of legal outcomes and the judicial process at large (Garg, 2021).

The current paradigm is bereft of a sophisticated solution that addresses the intricate nuances of handwritten Tamil characters. This conspicuous gap in technological innovation is especially pronounced given the high stakes involved in legal documentation and the considerable membership of the Madras Bar Council, which stood at 60,000 as of 2018 ("Bar Council of Tamil Nadu and Puducherry," 2023).

The objectives of this study are thus twofold. Firstly, to meticulously dissect the inefficiencies rooted in the status quo of manual transcription, and secondly, to articulate and test the hypothesis that the introduction of an advanced AI-enabled Handwritten Character Recognition for Tamil (HCRT) system could dramatically slash transcription times by half, while simultaneously enhancing the precision of digitized records. The HCRT is an offline Optical Character Recognition (OCR) system that reads a document scanned from paper documents, segments it into individual characters, identifies the characters, and parses them to frame meaningful words and sentences (*Handwriting Recognition*, n.d.).

Problem Description:

The manual transcription of handwritten legal documents represents a bottleneck within the legal workflow, resulting in inefficient resource utilization and increased risk of errors. Quantitative studies reveal that legal practitioners spend an average of 32 hours per month transcribing documents such as wills and land records (Madras High Court, n.d.; Advocate, personal communication, February 2024), highlighting the substantial time investment required for this task. The root cause of this problem lies in the lack of access to automated solutions capable of effectively handling handwritten Tamil characters, underscoring the need for technological innovation within the legal community.

“Legal technology is not just about adopting new tools; it’s about embracing a future where efficiency, accuracy, and accessibility are paramount. It enables legal professionals to deliver faster, more reliable services while minimizing the risk of human error. For our clients, this means quicker resolutions to their legal matters, more transparent communication, and ultimately, better outcomes. (Group, 2024)”

Note: The 32-hour figure was derived based on information obtained from an advocate practicing in the Madras High Court during a personal communication in February 2024.

Targets:

The target is to reduce transcription times by half, aiming to streamline document processing and improve efficiency within the legal sector. By implementing automated solutions capable of handling handwritten Tamil characters, the goal is to alleviate the burden on legal practitioners and enhance overall workflow productivity. The automated solution is aimed at having a system that takes the scanned legal documents as feed, segments the characters, identifies them, parses them to frame meaningful words and sentences, and creates a digitized copy of the document (*Handwriting Recognition*, n.d.).

Root Cause Analysis:

The lack of access to automated solutions for handling handwritten Tamil characters is the root cause of the manual transcription challenge. This technological gap stems from limited advancements in Tamil character recognition technology, inadequate research and development efforts, and insufficient collaboration among stakeholders. Addressing these root causes requires innovative approaches to develop advanced Handwritten Character Recognition (HCRT) models tailored for Tamil script (Garg, 2021).

Objectives:

The objectives of this research are:

- To develop and implement advanced HCRT models capable of accurately transcribing handwritten Tamil legal documents.
- To evaluate the effectiveness of HCRT models in reducing transcription times and improving accuracy in legal document processing.
- To assess the impact of automated solutions on the efficiency and productivity of legal practitioners in Tamil Nadu.

User Acceptance Parameters

To evaluate the success of this project below are the crucial parameters for the end users to accept.

- **Accuracy of Transcription:** The degree to which the HCRT system correctly recognizes and transcribes handwritten Tamil characters into digital text. This can be measured by the percentage of accurately transcribed characters or words in test documents compared to manual transcriptions.
- **Time Efficiency:** The amount of time saved in the transcription process using the HCRT system compared to traditional manual transcription methods. This can be measured by a reduction in average transcription time per document when using the HCRT system.
- **User-Friendliness:** The ease with which legal professionals can use the HCRT system without extensive training or technical support. User ratings can be obtained from surveys or questionnaires assessing the system's interface, ease of learning, and overall usability.
- **Adaptability to Handwriting Variability:** The system's ability to accurately recognize and transcribe a wide range of handwriting styles, including those of individuals with unique or challenging scripts. This can be measured by accuracy rates across diverse handwriting samples, including those considered outliers or challenging.

Hypothesis:

Implementing an advanced Handwritten Character Recognition (HCRT) model, will significantly reduce the time required for manual transcription by 50% and increase the accuracy of information extraction from handwritten Tamil legal documents, thereby enhancing the efficiency and accuracy of legal proceedings.

Dependent Variable: The efficiency and accuracy of information extraction from handwritten Tamil legal documents. This is what the study aims to improve, and its change is observed as a result of implementing the HCRT model.

Independent Variable: The implementation of the advanced Handwritten Character Recognition (HCRT) model. This is the variable manipulated to observe its effect on the dependent variable.

Population: Legal professionals in Tamil Nadu who process handwritten Tamil legal documents.

Methodology:

The methodology of our project unfolds in a series of deliberate and structured stages, each meticulously designed to build towards an Optical Character Recognition (OCR) model adept at processing handwritten Tamil characters in Legal documents.

Data Collection and Image Preprocessing:

The foundation of our OCR model is a comprehensive dataset of handwritten Tamil characters. We commence by amassing a diverse range of handwritten documents, ensuring variations in style and stroke are captured. Each document is scanned and then subjected to a series of image preprocessing routines. This includes normalization to a consistent scale and contrast enhancement to ensure that the characters stand out for the upcoming segmentation process.

Character Segmentation:

Segmentation is the first critical step in our model's pipeline, where we dissect the scanned images of documents into individual characters. Employing advanced segmentation algorithms, we meticulously separate overlapped and connected characters. This step is crucial for isolating each character for individual recognition, laying the groundwork for accurate transcription.

Character Recognition and Sentence Parsing:

Following segmentation, we engage a Convolutional Neural Network (CNN) for the recognition of each isolated character. The CNN, renowned for its efficacy in image recognition tasks, is meticulously trained with our dataset, learning to identify the intricate patterns of the Tamil script. Once the characters are recognized, we apply parsing algorithms to assemble the characters into coherent sentences, mirroring the structure of the original document.

Model Evaluation:

To rigorously assess our OCR model's performance, we employ a paired t-test, a statistical method that compares the time efficiency of manual transcription versus the OCR system. By analyzing the time taken to transcribe documents by hand with the time our HCRT system requires, we can evaluate our hypothesis regarding the system's efficacy. This analysis will not only demonstrate the potential time savings but also provide a statistical measure of the improvement offered by the OCR system.

Hypothesis Testing:

The hypothesis postulates that the OCR model significantly reduces transcription time compared to manual methods. Through the paired t-test, we will establish whether the observed efficiency gains of the OCR system are statistically significant. A reduction in transcription time would corroborate the model's effectiveness and its potential as a transformative tool in legal documentation processes.

This approach strategically combines tried-and-tested computer vision techniques with state-of-the-art machine learning models, culminating in a robust analysis that validates the system's practical utility. The goal is not only to automate a labor-intensive process but to do so in a manner that is grounded in rigorous empirical testing, ensuring a solution that is both innovative and reliable.

Dataset:

The main dataset for the project is the unconstrained Tamil Handwritten Character Database (uTHCD). The samples were generated from around 850 native Tamil volunteers including school-going kids, homemakers, university students, and faculty. The database consists of about 91000 samples with nearly 600 samples in each of 156 classes (Shaffi & Hajamohideen, 2021).

The second dataset is Isolated Handwritten Tamil Character Dataset (HPL) was collected by HPLabs using HP TabletPCs and is in standard UNIPEN format. This dataset contains approximately 500 isolated samples each of 156 Tamil “characters” (details) written by native Tamil writers including school children, university graduates, and adults from the cities of Bangalore, Karnataka, India and Salem, Tamil Nadu, India (*HPL Isolated Handwritten Tamil Character Dataset*, n.d.).

In our project, we plan to amalgamate the HPLabs Tamil dataset with the uTHCD dataset (*Class Dataset*, n.d.) to harness the strengths of both. The fusion is set to enrich the OCR model's training, benefiting from uTHCD's vast array of authentic handwriting variations and HPLabs' precise digital input. This strategic combination aims to enhance the model's accuracy and versatility, equipping it to adeptly process an expanded spectrum of Tamil script handwriting, thereby improving its real-world application efficacy. Integrating these datasets will create a more comprehensive training environment, potentially leading to superior recognition performance.

uTHCD dataset: <https://kaggle.com/code/gauravduttakiit/class-dataset-uthcd>

HPLabs dataset: <https://lipitk.sourceforge.net/datasets/tamilchardata.htm>

Exploratory Data Analysis (EDA)

EDA is rigorously implemented to understand the data distribution across different character classes. Through visualizations and statistical metrics, we assess the completeness and quality of our data, addressing any imbalances or missingness. This process informs further preprocessing needs and potential model improvements.

HPL Dataset Observations:

Inconsistent Class Distribution: The HPL dataset is noted for having an inconsistent distribution of classes. This implies that some characters are over-represented while others are under-represented, which can pose challenges in training a balanced model that performs well across all character classes.

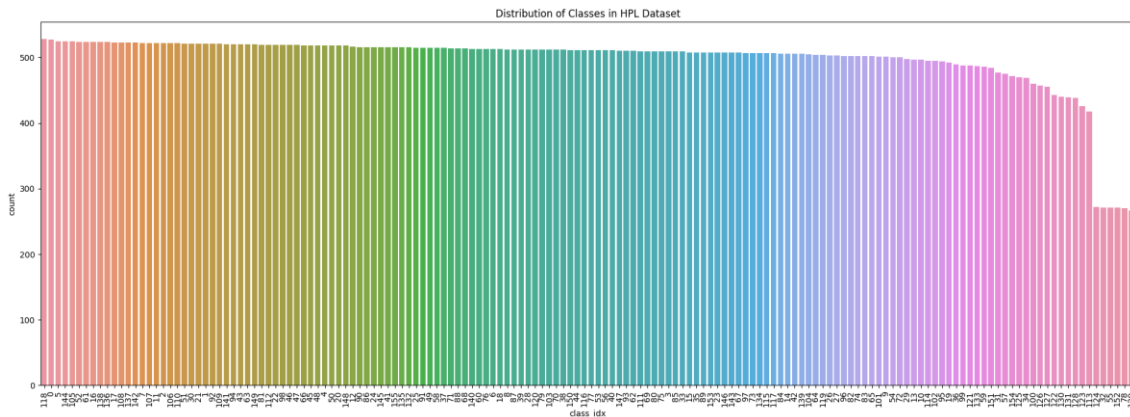


Figure 1: Class distribution of the HPL dataset.

Variable Image Sizes: Images within the HPL dataset vary in size, which introduces additional preprocessing requirements to standardize the images before they can be effectively used for training the HCRT model. Standardization is crucial for ensuring that the model learns from the character shapes rather than being influenced by the size differences.

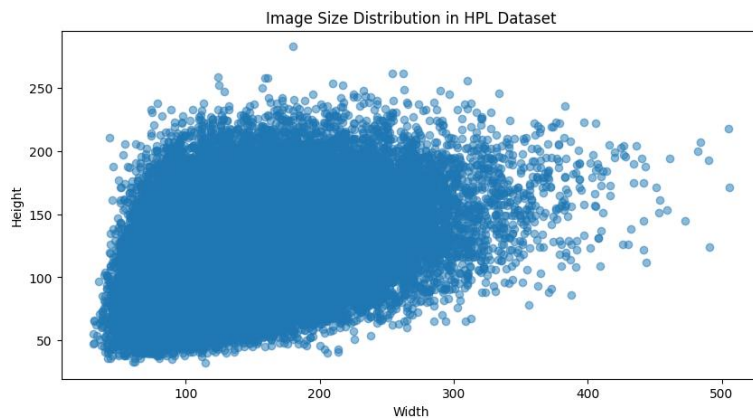


Figure 2: Size distribution of the HPL dataset.

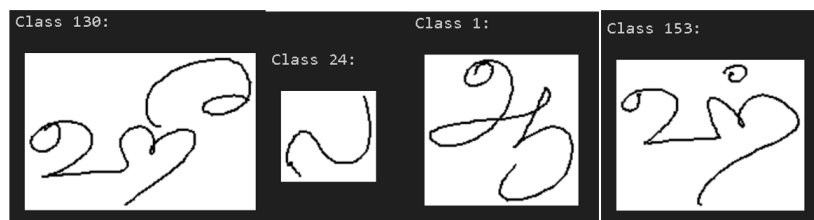


Figure 3: Samples from the HPL dataset.

uTHCD Dataset Observations:

Uniform Class Distribution: Contrary to the HPL dataset, the uTHCD dataset is highlighted for its uniformly distributed classes, meaning each character class has roughly the same number of samples. This uniformity is beneficial for model training as it allows the model to learn equally from each character, potentially leading to a more balanced recognition capability.

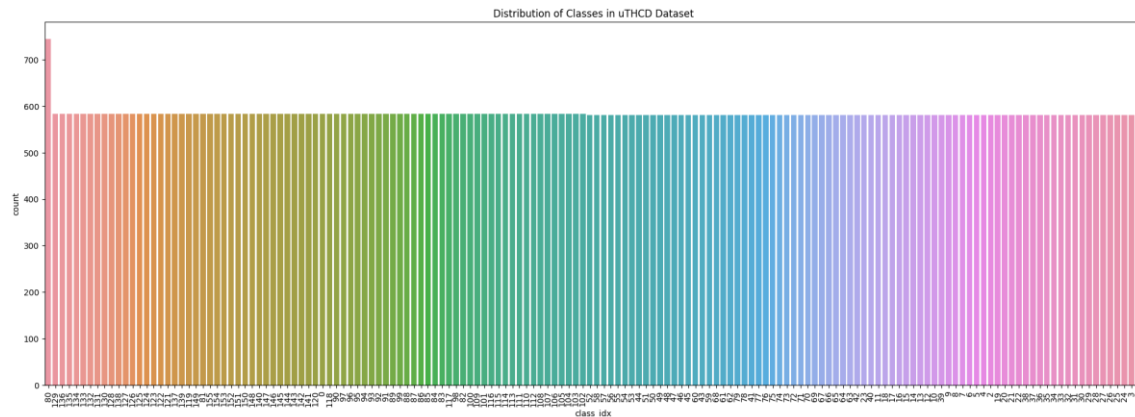


Figure 4: Class distribution of the uTHCD dataset.

Fixed Image Size: All images in the uTHCD dataset have a fixed size of 64x64 pixels. This consistency eliminates the need for size standardization, simplifying the preprocessing step and ensuring that the model can focus on learning character features rather than adjusting to varying image sizes.

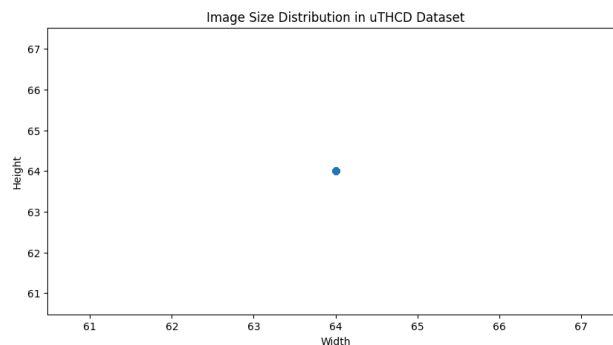


Figure 5: Size distribution of the uTHCD dataset.

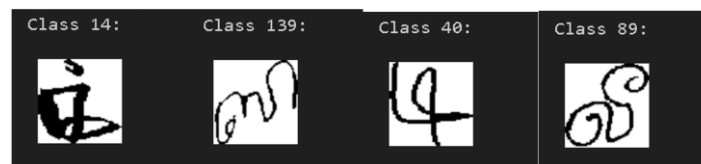


Figure 6: Samples from the uTHCD dataset.

These datasets cannot be simply combined because of the significant difference in the class distributions. A chi-squared test confirms the statistical significance of the difference in the class distributions with the p-value $4.88e-54$.

As a result of the EDA, we chose the uTHCD dataset for the model training after the necessary preprocessing steps like grayscale conversion and binarization.

Results:

Firstly, we developed and trained the CNN model for identifying individual characters, which is a crucial part of the project. The uTHCD dataset was split into train, test, and validation sets and then converted into grayscale in the preprocessing step.



Figure 7: Preprocessed images.

The preprocessed dataset was used to train the model after applying data augmentation steps like rescaling, rotation, and zoom. Below is the sample of preprocessed images and the performance of the model during training and validation after 24 epochs.

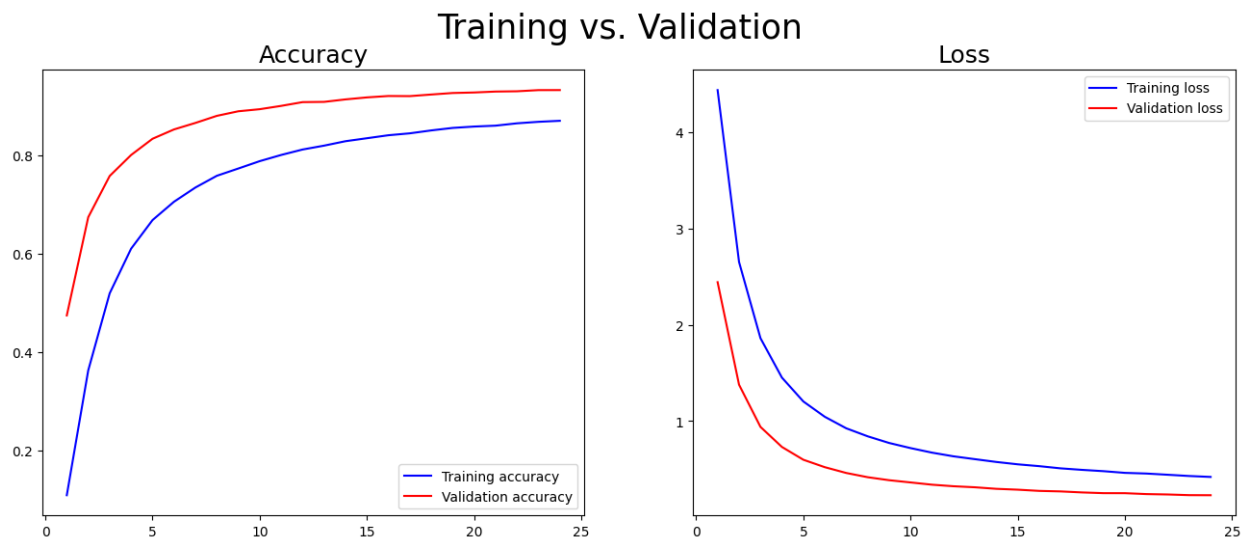


Figure 8: Model performance.

The model training achieved 87% accuracy and 0.42 categorical cross-entropy loss, and the validation achieved 93% accuracy and 0.23 loss, as depicted in Figure 9.

For testing further, we predicted some of the test images using the model and below are the results.











| | | | |
|--|------------|--|------------|
| Predicted character: க | class: 18 | Predicted character: ே | class: 154 |
|  | |  | |
| Predicted character: ஒ | class: 10 | Predicted character: ட | class: 63 |
|  | |  | |
| Predicted character: து | class: 52 | Predicted character: ா | class: 0 |
|  | |  | |
| Predicted character: ட | class: 69 | Predicted character: ரி | class: 82 |
|  | |  | |
| Predicted character: னா | class: 145 | Predicted character: ல் | class: 86 |
|  | |  | |

Figure 9: Testing.

Hypothesis Evaluation

To get the time taken for a character to be classified, the `timeit` package has been used, and the model is taking $646 \text{ ms} \pm 16.6 \text{ ms}$ to classify ten test images.

```
%timeit
[predict(img) for img in test_images]
```

646 ms \pm 16.6 ms per loop (mean \pm std. dev. of 7 runs, 1 loop each)

Figure 10: Time evaluation of model prediction.

The sample metrics for manual transcribing have been collected from a personal conversation with an advocate (Madras High Court, n.d.; Advocate, personal communication, February 2024), and the model time taken has been calculated based on the previous result and the average number of characters per page.

| Document Name | Pages | Time Taken for Manual (in mins) | Time Taken for Model (in mins) |
|---------------|-------|------------------------------------|-----------------------------------|
| Document 1 | 16 | 180 | 10.336 |
| Document 2 | 20 | 240 | 12.92 |
| Document 3 | 19 | 250 | 12.274 |
| Document 4 | 32 | 300 | 20.672 |

Table 1: Elapsed time comparison

With the information in Table 1, a paired T-test has been conducted to evaluate the performance of the HCRT model against manual transcription efforts. This test illuminates whether technological intervention significantly accelerates the transcription process. The test results with the T-Statistic of 10.105 and 0.002 as p-value.

Since the p-value is much less than the conventional threshold of 0.05, it would substantiate the hypothesis that the HCRT system provides a considerable improvement over traditional manual transcription methods in terms of speed.

Backlog

Implementing segmentation for the documents presents challenges due to the intricate and dense nature of the Tamil script. Tamil characters are complex with many curves and loops. Additionally, the script includes many compound characters and modifiers that appear above or below characters, further complicating the segmentation process. These factors require advanced algorithms capable of understanding the contextual and spatial relationships between components of the script.

Given the complexity of Tamil script, achieving effective segmentation has proven to be a formidable task. The challenges have led to delays in fully implementing this feature, necessitating its addition to the project backlog. Continuous research and development efforts are required to devise a robust solution that can handle the diverse nuances of Tamil handwriting.

Also, the integration of modules, internal testing, performance optimization, compliance, and security enhancements are added to the backlog.

Roadmap

Considering the remaining backlog items and future plans, a road map has been generated, and Figure 12 shows the roadmap for the HCRT system.

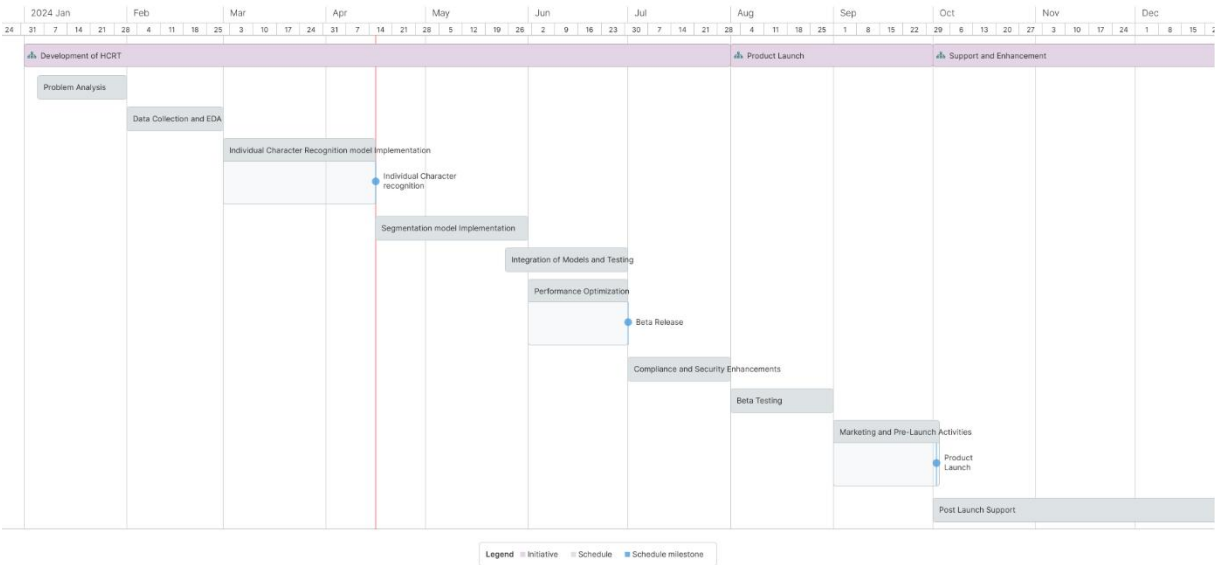


Figure 11: Roadmap.

This roadmap visualizes the project timeline for the development, testing, launch, and post-launch support for the HCRT system from January to December 2024. It outlines a phased approach to the project. So far, the milestone of individual character recognition has been reached. The milestone of releasing the beta version is planned for July, and the full product launch is planned for October 2024. After the product’s release, three months of post-launch support is planned to be provided.

Conclusion

In conclusion, the project to develop an AI-powered Handwritten Character Recognition system for Tamil script has made significant strides in bridging the digital gap within Tamil Nadu's legal documentation process. The journey began with a thorough analysis of the problem and meticulous data collection, leading to a sophisticated model capable of identifying individual characters. The model's effectiveness in reducing transcription time has been validated using a paired t-test, yielding promising results that underscore the potential of the HCRT system to enhance efficiency and accuracy in legal proceedings.

However, due to the intricate nature of Tamil script and the complexity involved in character segmentation, full document transcription remains a challenge and has been placed in the project's backlog. This highlights the need for ongoing research and innovation to refine the OCR system. Despite this, the project has reached a pivotal milestone with individual character recognition, setting the foundation for future enhancements.

As we look towards the future, the continued development of the HCRT system will focus on overcoming the segmentation hurdles, aiming for a holistic solution that fully automates the transcription process. The roadmap outlines a clear path forward, culminating in a product launch planned for October 2024. Post-launch, we will dedicate ourselves to supporting users and enhancing the system, ensuring that the HCRT continues to evolve and meet the dynamic needs of the legal community. This project stands as a testament to the transformative power of artificial intelligence in addressing real-world challenges and paves the way for digital transformation in legal document management.

References:

Madras High Court. (n.d.). Retrieved March 25, 2024, from <https://hcmadras.tn.gov.in/>

Bar Council of Tamil Nadu and Puducherry. (2023). In *Wikipedia*.

https://en.wikipedia.org/w/index.php?title=Bar_Council_of_Tamil_Nadu_and_Puducherry&oldid=1165693692#cite_note-Telag_Bar_Ref2-19

Handwriting Recognition: Definition, Techniques & Uses. (n.d.). Retrieved April 10, 2024, from

<https://www.v7labs.com/blog/handwriting-recognition-guide>,

<https://www.v7labs.com/blog/handwriting-recognition-guide>

Garg, R. (2021, November 24). Digitalization of the legal world. *iPleaders*.

<https://blog.ipleaders.in/digitalization-of-the-legal-world/>

Shaffi, N., & Hajamohideen, F. (2021). uTHCD: A New Benchmarking for Tamil Handwritten

OCR. *IEEE Access*, 9, 101469–101493. <https://doi.org/10.1109/ACCESS.2021.3096823>

Class Dataset: uTHCD. (n.d.). Retrieved February 19, 2024, from

<https://kaggle.com/code/gauravduttakiit/class-dataset-uthcd>

HPL Isolated Handwritten Tamil Character Dataset. (n.d.). Retrieved December 18, 2023, from

<https://lipitk.sourceforge.net/datasets/tamilchardata.htm>

Group, M. L. (2024, March 8). *A Comprehensive Guide to Understanding Legal Technology*.

Moton Legal Group. <https://motonlegallgroup.com/legal-technology/>