Digitalization of Tamil Palm Leaf Manuscripts Using a Transfer Learning Approach

Balika J. Chelliah[1], B. Aarthi[2*] and B. Judy Flavia[3]

[1,3] Department of Computer Science and Engineering, SRM IST, Ramapuram, Chennai, India

[2]Assistant Professor Senior Grade I, School of Computer Science and Engineering, Vellore Institute of Technology, India

**aarthib@srmist.edu.in*

The Tamil language, renowned for its rich cultural heritage and historical significance, has been preserved for centuries through palm leaf manuscripts. These ancient records serve as invaluable repositories of knowledge, encompassing literary, religious, and historical texts. Tamil palm leaf manuscripts contain invaluable medicinal and scientific knowledge passed down through generations. However, due to aging, environmental degradation, and limited accessibility, digitizing these manuscripts has become essential for preservation and knowledge dissemination. This research proposes a transfer learning-based approach to digitize and translate ancient Tamil palm leaf manuscripts, ensuring their preservation and wider accessibility. The process begins with high-resolution image acquisition, followed by preprocessing techniques such as noise reduction, binarization, and morphological transformations to enhance text clarity. A custom-trained Optical Character Recognition (OCR) model, built on ResNet1.7 with Bidirectional LSTM, was employed to extract Tamil script characters from the processed images. The extracted text undergoes post-processing corrections using sequence modeling techniques to enhance readability. Experimental evaluations demonstrated that the transfer learning-based OCR model outperforms traditional methods in recognizing Tamil script from deteriorated manuscripts, significantly reducing character recognition errors. To improve accuracy, an attention mechanism was integrated into the OCR pipeline, ensuring robust recognition even in faded or degraded manuscripts. The transformer-based translation model effectively converts historical Tamil into a more accessible form, ensuring the preservation and dissemination of ancient knowledge. The system has also been designed to handle various dialects, handwritten scripts, and complex ligatures present in ancient Tamil texts. This research contributes to the preservation of Tamil cultural heritage while advancing the field of document image processing, OCR, and neural machine translation.

Keywords: *Digitalization; Neural machine translation; Optical character recognition; Tamil palm leaf manuscripts; Transfer learning*