ABSTRACT

Optical Character Recognition (OCR) is a technology that enables computers to recognize and convert text in digital images into editable and processable text data. This study aims to implement the Long Short-Term Memory (LSTM) algorithm in an OCR system to extract text from scanned images of high school exam documents and evaluate its recognition accuracy. The developed system consists of two main stages: training the LSTM model and testing the model on scanned document images. The training process utilized 2,342 training data with 73 character classes, while the testing phase involved 30 test images scanned from high school exam books in Times New Roman font. The testing results showed an average recognition accuracy of 88.50%, with the highest accuracy reaching 92.89%. Additionally, external validation was conducted by involving five high school teachers as respondents to evaluate the OCR system based on the aspects of output accuracy, text readability, and feasibility for practical use. The questionnaire results indicated that the system is feasible to use, achieving an average score of 4 (Agree) for accuracy, 3.8 (Fairly Agree to Agree) for readability, and 4.2 (Agree) for feasibility. This OCR system is beneficial for accelerating the digitalization process of exam documents and supports teachers' efficiency in managing digital question banks. However, character recognition accuracy can still be improved by optimizing the segmentation stage to address issues with touching characters and noise in input images. This study provides an initial contribution to developing deep learning-based OCR technology in the field of education.

Keywords — *OCR, LSTM, character recognition, image processing.*