**Paper No: SC-24** **Smart Computing**

# A Comparative Evaluation of PDF-to-HTML Conversion Tools

Pramodya Pathirana[1*], Asini Silva[2], Thenuka Lawrence[3], Thushani Weerasinghe[4], Roshan Abeyweera[5]

[1] *University of Colombo School of Computing, Colombo, Sri Lanka, 2018is054@stu.ucsc.cmb.ac.lk*

[2] *University of Colombo School of Computing, Colombo, Sri Lanka, 2018is079@stu.ucsc.cmb.ac.lk*

[3] *University of Colombo School of Computing, Colombo, Sri Lanka, 2018is044@stu.ucsc.cmb.ac.lk*

[4]*University of Colombo School of Computing, Colombo, Sri Lanka, taw@ucsc.cmb.ac.lk*

[5]*University of Colombo School of Computing, Colombo, Sri Lanka, rns@ucsc.cmb.ac.lk*

PDF (Portable Document Format) is a popular file format used for sharing and storing documents across different platforms. However, there are occasions when the content of a PDF document needs to be re-purposed for online use. PDF-to-HTML conversion is a common method used to achieve this goal. This research paper presents a comparative evaluation of existing PDF-to-HTML conversion tools for their suitability in extracting text and images. These tools were tested using school textbooks in Sri Lanka, which contain complex text formatting and non-textual elements. The evaluation was based on various criteria, such as the accuracy of the output, handling of complex text formatting, and non-textual elements. Comparisons were drawn based on the performance of each of these tools with respect to the criteria. The study provides useful insights for individuals and organizations looking to re-purpose PDF content for online use in the HTML format, particularly in the education sector.

**Keywords:** *e-learning, educational design research, text extraction, PDF to HTML conversion*