CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

Notice for the PhD Viva-Voce Examination

Mr Abhijit Guha (Registration Number: 1981901), PhD scholar at the School of Sciences, CHRIST (Deemed to be University), Bangalore will defend his PhD thesis at the public viva-voce examination on Saturday, 30 September 2023 at 10.30 am in Room No. 044, Ground Floor, R & D Block, CHRIST (Deemed to be University), Bengaluru - 560029.

Title of the Thesis	**:**	**Intelligent Information Retrieval Model for Digital Documents in Title Insurance**
Discipline	**:**	**Data Science**
External Examiner (Outside Karnataka, Maharashtra and Uttar Pradesh)	**:**	**Dr Jitesh Pradhan** Associate Professor Department of Computer Science and Engineering National Institute of Technology (NIT) Jamshedpur Jharkhand
External Examiner (Within Karnataka, Maharashtra and Uttar Pradesh)	**:**	**Dr Sandeep Saxena** Professor and HOD-CSE Greater Noida Institute of Technology (GNIOT) B-1302, ACE PLATINUM Society, Zeta-1 Greater Noida Uttar Pradesh - 201310
Supervisor	**:**	**Dr Debabrata Samanta** Associate Professor Department of Statistics and Data Science School of Sciences CHRIST (Deemed to be University) Bengaluru - 560029 Karnataka

The members of the Research Advisory Committee of the Scholar, the faculty members of the Department and the School, interested experts and research scholars of all the branches of research are cordially invited to attend this open viva-voce examination.

Registrar

Place: Bengaluru
Date: 29 September 2023

ABSTRACT

Documents have been pivotal in shaping human history by preserving knowledge and enabling the transmission of ideas across generations and cultures. They have facilitated the establishment of legal systems, institutions, and governance, fostering societal order and progress. Additionally, documents serve as a collective memory, chronicling the achievements and lessons learned, enriching the human experience.

Transforming documents from physical to digital format has revolutionized how we access, store, and share information in the digital age. This transition, enabled by technological advances, began with the invention of the scanner, which allowed for the digital capture of images and text. Optical Character Recognition (OCR) technology that can convert scanned documents into searchable, editable digital texts further streamlined this process. As the storage capacity and internet speeds have increased, digitization has become more accessible and widespread. Cloud-based storage solutions, such as Google Drive and Dropbox, now allow users to store, access, and share digital documents from anywhere with an internet connection. This has improved collaboration and communication and reduced the need for physical storage space. The digitization of documents has also significantly impacted the environment, with paper consumption decreasing and many industries' carbon footprint reducing. Libraries and archives have transformed digitally, making vast information more easily accessible and preserving vital historical records for future generations. This digital shift has democratized knowledge, granting people worldwide access to resources that were once limited to those with physical proximity to the material. Moreover, digital documents have enabled the creation of innovative tools, such as Machine Learning (ML), Deep Learning (DL) algorithms and Natural Language Processing (NLP) or computational linguistics, that can analyze and process vast amounts of data leading to the development of new technologies and accelerating scientific research across various fields. Digitization has also facilitated the growth of the e-book market, revolutionizing the publishing industry and providing readers with greater access to literature. In addition, digitizing documents has transformed industries such as healthcare, where electronic health records have improved patient care and streamlined medical data management. In the legal sector, e-discovery and digital forensics have emerged as essential practices, enabling professionals to find relevant information more efficiently. The Title Insurance industry has been undergoing a drastic change where the underwriting and examination of property titles are being automated through Artificial Intelligence (AI).

The proposed Intelligent Document Processing (IDP) system has been thoroughly validated, demonstrating its effectiveness and reliability. The validation process consisted of two critical aspects: verifying individual task models and evaluating the overall system in real-world scenarios. The system was tested using telecommunications, public, and banking datasets, indicating its broad applicability and effectiveness in various contexts.

The research study introduced a novel Page Stream Segmentation (PSS) module that employs a bimodal approach. Results showed improved overall segmentation through-put precision (STP) for three distinct datasets: Technical Instructions (TI), Banking, and Tobacco800. The study also focused on developing a ground breaking hybrid approach for Document Anomaly Detection (DAD) that combines classical statistical techniques with state-of-the-art DL methods. Using autoencoders for dimensionality reduction led to an accuracy improvement of 2.66% and a 99.76% reduction in training and inference times.

A novel Data Leak Prevention (DLP) module was presented, employing a binary classification framework with composite n-gram features and artificial neural networks (ANN). The module demonstrated a significant improvement in the F1 score by 4.82% compared to the baseline. Furthermore, a new approach for information extraction from structured digital documents was introduced, utilizing traditional digital image processing techniques. The system extracted 17 out of 32 entities with 100% accuracy, 14 with an accuracy greater than 80%, and 1 with an accuracy greater than 70%.

For Non-Textual Information Extraction, a novel approach for margin detection in digital image documents was developed using the classical DLP (Dynamic Line Pro-filing) technique without OCR. The approach demonstrated a 16.2% higher accuracy compared to current state-of-the-art algorithms. Finally, the overall system validation for the IDP system showcased an impressive Straight Through Processing (STP) rate of 87% when processing 40,000 packages, compared to the baseline system with an STP rate of 0

The proposed IDP system has shown remarkable effectiveness in automating and streamlining various aspects of business processes. With an impressive 87% STP rate, the system significantly outperforms the baseline, highlighting its potential for transforming how organizations manage their document processing tasks.

Keywords: Intelligent Document Processing, Natural Language Processing, Title Insurance, Machine Learning, Deep Learning, Computer Vision, Artificial Intelligence.

Publications:

1. **Guha, Abhijit**, and Debabrata Samanta. "Hybrid approach to document anomaly detection: an application to facilitate RPA in title insurance." International Journal of Automation and Computing 18.1 (2021): 55-72.
2. **Guha, Abhijit**, et al. "A deep learning model for information loss prevention from multi-page digital documents." IEEE Access 9 (2021): 80451-80465.
3. **Guha, Abhijit**, et al. "A Multi-Modal Approach to Digital Document Stream Segmentation for Title Insurance Domain." IEEE Access 10 (2022): 11341-11353.
4. **Guha, Abhijit**, Debabrata Samanta, and S. K. Islam. "IIRM: Intelligent Information Retrieval Model for Structured Documents by One-Shot Training Using Computer Vision." Arabian Journal for Science and Engineering (2022): 1-17