International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Exploring Techniques for Analyzing Sinhala Document Layouts and Styles: A Systematic Literature Review

***Hulathdoowage S.K.D*[a], *Kumara B.T.G.S*[b]**

[a]Department of Computing & Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka.

[b]Department of Data Science, Faculty of Computing, Sabaragamuwa University of Sri Lanka.

ABSTRACT:

Document layout analysis is an important part of digitization, information retrieval, and document understanding processes. However, effective digitization cannot be achieved with optical character recognition technology alone, as layouts and styles vary from document to document, which highlights the need for separate analysis of different document types and languages. This study conducted a comprehensive literature review to examine current technologies for document layout analysis and assessed the applicability of the U-Net architecture in the domain. It also examined existing technologies for Sinhala optical character recognition to convert Sinhala document images into editable digital formats. The process involved database searches, manual searches, researchers and research groups, and snowball methods. Through this process, 1257 articles were examined, and 39 relevant articles were identified. Furthermore, convolutional neural network based methods outperformed traditional methods and confirmed the applicability of the U-Net architecture. Additionally, Tesseract based Sinhala optical character recognition technologies outperformed other methods. The findings underscore the need for comprehensive layout segmentation to accurately identify different elements in document images. Finally, this review contributed to identifying a comprehensive approach to digitization of Sinhala documents.

Keywords: Document Layout Analysis, Optical Character Recognition, Semantic Segmentation, Sinhala, u-net

Introduction

Document layout analysis (DLA) is a process of identifying various elements (e.g. text, figures, tables, formulas, headers, and footers) within a document image (Akanda et al., 2024). It helps to convert unstructured digital documents into a format that computers can understand and process effectively. Furthermore, DLA is critical for digitizing and archiving large volumes of paper documents, making them searchable and easily accessible in original formats (Aljiffry et al., 2022; Almutairi & Almashan, 2019; Banerjee et al., 2024; Cao et al., 2022; Cuo et al., 2019). It also plays a pivotal role in information extraction and retrieval, and document understanding tasks (Dang & Nguyen, 2021). Furthermore, Document layout analysis achieves significant improvement in various sectors, such as banking & finance, education, government, and healthcare (De Nardin et al., 2024; Gemelli et al., 2022; Grijalva et al., 2021; Grüning et al., 2019).

Optical character recognition (OCR) is the technology that enables to extraction of characters from images in accessible and editable format (Huang et al., 2019; Kosaraju et al., 2019). However, document digitization cannot be done using OCR technologies alone and it is essential to conduct layout analysis to identify objects on document images in their exact formats including size, style, and position. Therefore, the attention of researchers has been focused on DLA but there are several challenges such as the diversity of documents, complexity of layouts, varying scales of images, and document-wise changes of the elements (Lee et al., 2019).

Research on document layout and style analysis for the Sinhala language is critical for preserving and digitizing complex documents, enabling efficient data extraction and accessibility (Ma et al., 2018). When converting images into editable digital documents, this analysis significantly reduces human effort by automating the process of recognizing and organizing layout and styles (Ma et al., 2020).

The U-Net architecture, first introduced for biomedical image segmentation (Mechi et al., 2019), has become a fundamental model in the field due to its ability to accurately outline the structures of medical images with limited training data. Furthermore, U-Net's success in biomedical tasks such as tumor detection (Mechi et al., 2021) and organ segmentation stems (Minouei et al., 2021) from its unique design that combines a contracting path for context capture and a symmetric expansion path for precise localization. Beyond biomedical applications, U-Net has been widely used for a variety of tasks including satellite image segmentation where it effectively identifies environmental observations to identify geographic features and land use changes (Nabiee et al., 2022). U-Net has potential in the field of document analysis to segment and analyze document layouts (Nguyen et al., 2022), as well as extract lines of text and other structural components (Ohyama et al., 2019; Oliveira & Viana, 2017).

This systematic literature review was conducted according to the Kitchenham et al methodology (Phong et al., 2020), which is a structured approach for conducting systematic literature reviews (SLRs) in software engineering and related fields, developed by Barbara Kitchenham and her colleagues. This SLR is involved in collecting, synthesizing, and analyzing research articles from manual searches, database searches, researchers & research groups and snowballing methods. Furthermore, IEEE Explore, ACM, Springer Link, and Science Direct were selected mainly as databases. Removal of duplicates and selection criteria based on titles, abstracts, and full text were applied to select the most relevant studies.

Review Methodology

A comprehensive review was conducted to identify the most relevant research papers in the domain of document layout and style analysis, U-Net architecture and related fields. Additionally, this study aims to explore language-specific results to demonstrate the need for document layout and style analysis in low-resource languages like Sinhala. This systematic literature review (SLR) adhered to the methodology proposed by Barbara Kitchenham in 2007 (Kitchenham et al., 2009). It incorporated predefined research questions and objectives, inclusion and exclusion criteria, search strategy, study selection process, and quality assessment criteria.

Review Protocol

A review protocol serves as a detailed guide outlining the methods and steps for conducting a review. For this SLR, a well-structured protocol was followed to ensure that the process was organized, transparent and repeatable. By following this protocol, the review minimizes bias and increases the reliability and accuracy of the results. The review protocol is essential when identifying the background of the literature and search strategy. It also assists in clearly understanding and documenting the research questions and quality assessment criteria. Furthermore, repositories were selected based on their Q-index and H-index to enhance the accuracy of the results.

Research Questions and Objectives

Research questions and objectives are critical for guiding a systematic literature review. These questions help define keywords and search strings to identify relevant research papers within the field. The primary goal of this study is to effectively analyze the layouts and styles of complex Sinhala document images and convert them into editable digital documents using an enhanced U-Net architecture. To accomplish this, research questions have been formulated to explore various technologies in the domain of document layout analysis, conduct comprehensive experiments with the U-Net architecture, and address Sinhala language character recognition, which is a subsequent step in the digitization process.

Problem statement: The digitization of document images faces significant challenges due to the lack of document layout and style analysis. Current methods struggle with segmenting and recognizing document elements, which limits the effectiveness of OCR, particularly for low-resource languages like Sinhala.

Research questions:

- RQ1: How to collect a dataset?
- RQ2: What improvements are required in the U-Net architecture for analyzing complex documents?
- RQ3: How can the layouts and styles of complex document images be accurately analyzed?
- RQ4: How can Optical Character Recognition (OCR) technology be applied to analyzed Sinhala documents?

Main objective: To digitize Sinhala document images by enhancing document layout and style analysis through an improved U-Net architecture integrated with transformer inspired feature learning blocks in vanilla U-Net and to improve Optical Character Recognition (OCR) performance through element wise OCR application.

Specific Objectives:

- RO1: Collect and annotate dataset of Sinhala document images.
- RO2: Enhance the U-Net architecture by integrating transformer inspired feature learning blocks.
- RO3: Develop and evaluate methods for detecting and categorizing document layouts and styles.
- RO4: Integrate with OCR technologies at the element level to improve text recognition accuracy.

Inclusion and Exclusion Criteria

Inclusion and exclusion criteria are essential for achieving the predefined objectives and identifying the most relevant studies. Studies were included if they directly addressed the main research question and were published between January 2019 and August 2024. Studies were excluded if they lacked an abstract or were published as an abstract, as well as non-primary studies such as editorials, keynotes, workshops, and dissertations. Additionally, studies that were not written in English, had significant methodological problems, or did not provide sufficient information about the procedures used were excluded. The initial search yielded 1257 results, and the final selection was made after a careful review of the title, abstract, and full text.

Quality Assessment Criteria

The pre-defined Quality Assessment Criteria (QAC) guided the selection of higher-quality research papers as the final outcome of the SLR. Six questions were used to conduct the quality assessment, applied during the reading of titles, abstracts, and full texts.

These questions were:

- Does the title of the paper relate to the research objectives?
- Does the abstract is clear, concise, meaningful, and relate to the research objectives?
- Does the dataset come from valid sources?
- Does the article clearly describe the research background and methodology?
- Does the research provide a clear explanation of the final outcomes?
- Does it conduct a proper evaluation process?

Search Strategy

This study employed a comprehensive search strategy based on relevant keywords. Three search strings were created for document layout analysis with object detection, document layout analysis and semantic segmentation with U-Net, and Sinhala with optical character recognition.

Table 1 - Search terms of the document layout analysis for object detection

Areas	Search terms
Document layout analysis	“document layout analysis”, “document structure analysis”, “document processing”, “document layout understanding”, “document layout recognition”, “Document region classification”
Image data	“image data”, “image dataset”, “document images”
Object detection	“object detection”, “element detection”, “object recognition”, “object identification”, “object tracking”
Search string: (“document layout analysis” OR “document structure analysis” OR “document processing” OR “document layout understanding” OR “document layout recognition” OR “Document region classification”) AND (“image data” OR “image dataset” OR “document images”) AND (“object detection” OR “element detection” OR “object recognition” OR “object identification” OR “object tracking”)	

Table 2 - Search terms of the document layout analysis by applying semantic segmentation and u-net

Areas	Search terms
Document layout analysis	“document layout analysis”, “document structure analysis”, “document processing”, “document layout understanding”, “document layout recognition”, “Document region classification”
Semantic segmentation	“semantic segmentation”, “pixel-wise classification”, “region-based segmentation”, “image segmentation”
U-net	“u-net”, “U-shaped network”
Search string: (“document layout analysis” OR “document structure analysis” OR “document processing” OR “document layout understanding” OR “document layout recognition” OR “Document region classification”) AND (“semantic segmentation” OR “pixel-wise classification” OR “region-based segmentation” OR “image segmentation”) AND (“u-net” OR “U-shaped network”)	

Table 3 - Search terms of the Sinhala language by applying optical character recognition

Areas	Search terms
Sinhala	“Sinhala”, “Sinhala language”, “Sinhala documents”
OCR	“ocr”, “optical character recognition”, “character recognition”, “character identification”
Search string: (“Sinhala” OR “Sinhala language” OR “Sinhala documents”) AND (“ocr” OR “optical character recognition” OR “character recognition” OR “character identification”)	

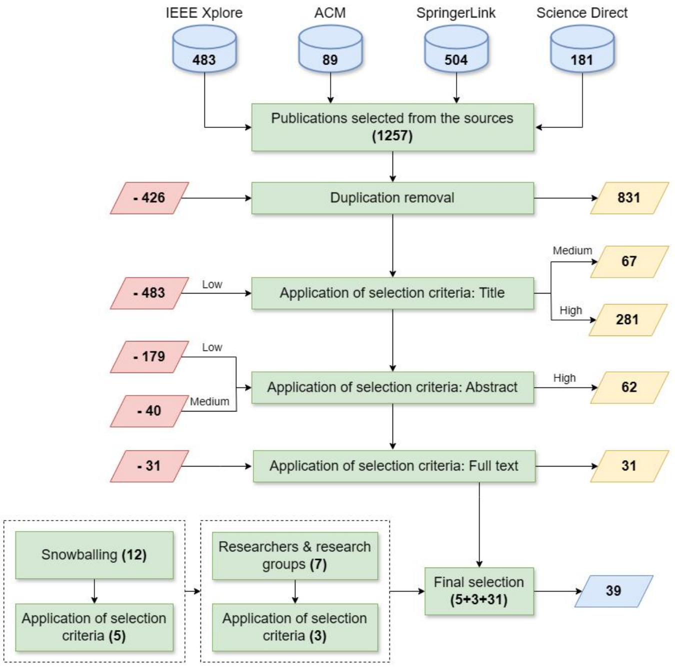
IEEE Xplore
483
ACM
89
SpringerLink
504
Science Direct
181
Publications selected from the sources (1257)
- 426
Duplication removal
831
Medium
67
- 483
Low
Application of selection criteria: Title
High
281
- 179
Low
- 40
Medium
Application of selection criteria: Abstract
High
62
- 31
Application of selection criteria: Full text
31
Snowballing (12)
Application of selection criteria (5)
Researchers & research groups (7)
Application of selection criteria (3)
Final selection (5+3+31)
39

Figure 1: Systematic literature review methodology

This systematic literature review (SLR) was conducted primarily using four scientific databases: IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect. Additionally, papers were selected through snowballing and recommendations from researchers and research groups. A total of 1,257 papers were identified using search strings, with an additional 12 papers found through snowballing and 7 papers sourced from researchers and research groups. Finally, 39 papers were selected from this pool for the study.

Study Selection Process

The study selection process involved several steps, including duplication removal, title screening, abstract screening, and full-text reading, to select the final papers. As illustrated in Figure 1, 426 duplicate studies were removed from the initial 1,257 results. The title screening process was then conducted for 831 studies, which were categorized as low, medium, or high relevance. Abstract screening was performed on the medium and high relevance papers, resulting in 62 papers being selected for full-text reading. After a careful full-text review, the most relevant 31 papers were selected. This rigorous process ensured the inclusion of most relevant studies to the research.

Table 4 - Illustrates the number of studies selected from each scientific database at each stage of the selection process.

Database	After duplication removal	Selection Criteria: Title	Selection Criteria: Abstract	Selection Criteria: Full Text
IEEE Xplore	326	127	42	19
ACM	87	35	6	2
Springer	255	83	9	7
Science Direct	163	103	5	3
Total	831	348	62	27

Outcomes of the Survey

Classification based on Methods

This systematic literature review (SLR) aims to investigate various deep learning approaches used in document layout analysis. Researchers have primarily employed deep learning-based methods and introduced hybrid and combined approaches to address specific challenges. Additionally, several segmentation and detection techniques have been developed to accurately identify and segment elements within documents. Moreover, some researchers have proposed specialized methods for tackling unique problems in document layout analysis.

- Deep Learning based Methods

Significant progress has been made in document layout analysis over the past few decades, yet several limitations remain in this context. Saiyed Umer, et al. (Umer et al., 2021) developed segmentation techniques to extract text regions from complex document layouts, distinguishing them from non-text regions such as symbols, logos, and graphics. The authors proposed a deep CNN model that includes three phases such as pre-processing with different patch sizes, post-processing with recursive partitioning, and text/non-text region prediction to address the goal. Furthermore, they used a collection of complex layout magazine images from Google Sites and the ICDAR 2015 database. Furthermore, Bo Wang, et al. (Wang et al., 2021) have proposed a method to address challenges in documents with sparse structures and fine typologies at variable scales. They introduced a novel U-net-based approach called MSNet for multi-scale segmentation. The aim is to identify and classify regions of interest, especially in complex and heterogeneous documents.

The development of a reliable method for text line segmentation of historical document images is a challenge in document image transcription, indexing, and retrieval systems. To mitigate this challenge Olfa Mechi, et al. (Mechi et al., 2019) proposed a novel deep learning-based method using an adaptive U-Net architecture. They evaluated the performance in text line segmentation with qualitative and numerical results using historical document images from the Tunisian National Archives and recent benchmark datasets from the ICDAR and ICFHR competitions.

Furthermore, to address the challenges of layout analysis and text recognition and distinguishing between handwritten and printed text in document images. Axel De Nardin, et al. (De Nardin et al., 2024) proposed a U-Net deep learning model for pixel-level segmentation combined with image enhancement methods such as median filtering and connected component analysis. They achieved high accuracy in segmentation with a Mean Intersection over Union (MIoU) score of 97.54%.Tahira, et al. proposed a hybrid approach for document layout analysis in document images by correctly identifying different graphical elements such as text, images, tables, and headings in document images (Shehzadi et al., 2024). This approach involves a transformer-based object detection network and query encoding mechanism. Then hybrid matching scheme to enhance object detection. Furthermore, they evaluated the model using PubLayNet (Zhong et al., 2019), DocLayNet (Pfitzmann et al., 2022), and PubTables benchmarks. Moreover, identify an optimal deep learning approach for document layout analysis (DLA) in low-resource and grapheme-based languages Md. Mutasim Billah, et al. (Akanda et al., 2024) investigated DiT, LayoutLMv3, and YOLOv8 to determine the most effective method for low-resource languages like Bengali. After experiments, YOLOv8 achieved an 8.95% better IoU score than DiT and 38.48% better IoU score than LayoutLMv3.

Xingjiao Wu, et al. proposed a method to improve document layout analysis by focusing on important edge details and high-frequency structures in images. They used an Explicit Edge Embedding Network (E3 Net), which contains edge embedding blocks, dynamic skip connection blocks, and a lightweight full translation subnet as the backbone (Wu et al., 2021). Evaluated on three document layout analysis benchmarks including synthetic document data to address data scarcity. Researchers introduced novel deep-learning frameworks to address the challenges in the context of document layout analysis and segmentation (Yang & Li, 2022). Yilun Huang, et al. proposed a YOLO-based method to improve table detection in document analysis (Huang et al., 2019). Other than that, to improve the identification and classification of components in scientific papers Felipe Grijalva, et al. (Grijalva et al., 2021) build a three-stage system using spectrograms of intensity histograms, a deep convolutional neural network (CNN) for classification, and Bag of Visual Words (BOVW) with Zernike moments for identifying isolated equations.

- Hybrid and Combined Approaches

Yong Cuo et al. conducted a study focused on performing layout analysis of complex Tibetan historical documents to facilitate digitization and subsequent text recognition (Cuo et al., 2019). By adopting an integrated deep learning approach, they implemented rule-based layout analysis for horizontal and vertical text regions. Their method provided reliable text regions for further text recognition and achieved high accuracy in detecting various document layouts, contributing significantly to the digitization of Tibetan historical texts.

Olfa Mechi et al. developed an effective framework for the segmentation of text lines in historical Arabic and Latin script images. Their approach applied various deep fully convolutional networks (FCNs) to extract the main text areas, followed by topological structure analysis to refine the FCN results. This combination allowed for the complete extraction of text lines, including ascender and descender components, resulting in a refined and detailed segmentation of text lines (Mechi et al., 2021).

Bui Hai Phong developed a system for the automatic detection of mathematical expressions in scientific document images, targeting both inline and isolated expressions. The system utilized a two-stage hybrid method, beginning with layout analysis to enhance text line and word segmentation, followed by detection using a combination of hand-crafted and deep learning features (Phong et al., 2020). This approach achieved detection accuracies of 91.18% and 81.35% for isolated and inline expressions, respectively, on the Marmot dataset, and 89.51% and 80.20% on the GTDB dataset.

- Segmentation and Detection techniques

Segmentation techniques involve dividing documents into separate regions for analysis. After that, the detection techniques identify and locate specific elements such as tables, paragraphs, formulas, and figures within those regions. In past decades, researchers have achieved significant improvement in segmentation (Almutairi & Almashan, 2019; Banerjee et al., 2024; Ma et al., 2020; Ohyama et al., 2019), Wataru Ohyama and others proposed a method for recognizing mathematical expressions in scientific document images that can complement traditional mathematical OCR processes without relying on mathematical and linguistic grammars. They used the U-Net framework is used to recognize mathematical symbols. This study shows that dataset diversity and additional training samples enhance performance and achieved superior performance compared to InftyReader.

Vahid Rezanezhad et al. introduced an approach with deep learning and heuristics to improve layout analysis for historical document images by enhance Optical Character Recognition (OCR) results (Rezanezhad et al., 2023). Uses pixel-wise segmentation with convolutional neural networks combined with heuristic methods to detect marginal and determine the reading order of text regions. Achieves higher accuracy and detects more layout classes (e.g., initials, marginal) compared to competitive approaches.

Abdullah Almutairi, et al. used Mask R-CNN to enhance information extraction and develop a deep learning model to segment newspaper pages into articles, advertisements, and headers (Almutairi & Almashan, 2019). Furthermore, to address the challenge of relying on labeled data Talha et al. used vision-based unsupervised pre-training to create object masks from unlabeled images and then train a detector for better object detection and segmentation (Sheikh et al., 2024). They have achieved significant improvements in accuracy and efficiency without the need for label data. Moreover, Ziyi Yang, et al. improve automatic layout analysis of Chinese academic papers using Mask R-CNN with a weighted anchor box mechanism for recognizing and locating nine layout elements (Yang & Li, 2022). Achieves 89.3% accuracy in recognizing layout elements by enhancing practical applications. They have used a custom layout image dataset of Chinese academic papers.

- Specialized and Novel Methods

Latifa Aljiffry et al. focused on improving layout analysis for Arabic documents, addressing a significant research gap in this field. They optimized the Fast R-CNN model for the layout analysis of Arabic printed and original handwritten documents, utilizing two distinct datasets with different regions of interest (RoI) (Aljiffry et al., 2022). Their approach achieved better F1 scores and accuracy compared to existing models, such as LABA and FFRA, highlighting the effectiveness of their model in enhancing Arabic document layout analysis.

Axel De Nardin et al. addressed the challenge of segmenting document layouts in ancient Arabic manuscripts, which often feature irregular layouts and damage due to age. Their approach simplified information extraction by introducing a one-shot learning method for document layout segmentation, requiring only one labeled page per manuscript for training (De Nardin et al., 2024). Despite minimal label data, their method achieved state-of-the-art performance in document layout segmentation, demonstrating its effectiveness in handling complex and damaged historical documents.

Table 5 - Summary of prior research selected on document layout analysis and segmentation.

Ref	Focus	Technique	Improvements	Dataset
(Cuo et al., 2019)	Digitization and subsequent text recognition	CTPN algorithm	Accuracy: horizontal row 0.98, vertical row 0.83	Tibetan historical documents
(Shehzadi et al., 2024)	Identify different graphical elements	CNN, transformer-based model	Average precision scores of 97.3% on PubLayNet, 81.6% on DocLayNet, and 98.6% on PubTables	PubLayNet, DocLayNet, and PubTables benchmarks
(Sheikh et al., 2024)	Address the challenge of labeled data scarcity	Self-supervised DINO	PubLayNet: 28.7 (box), 29.3 (mask) TableBank: 88.6 (box), 88.8 (mask) DocLayNet: 22.4 (box), 24.2 (mask)	Unlabeled document images
(Wu et al., 2021)	Edge details and high-frequency structures in images	E3 Net, FCN	DSSE-200 - 0.82, CS-150 - 0.96, ICDAR2015 -90.59	DSSE-200, CS-150, and ICDAR2015
(Umer et al., 2021)	Segmentation of text and non-text regions	CNN	Average accuracy 89.78%, F1-Score 0.8945	Magazine images from Google Sites and the ICDAR 2015 database
(Wang et al., 2021)	Multi-scale segmentation	CNN, MSNet	Pixel Accuracy 96.61%, Mean IoU 90.55%	Chinese document dataset
(Akanda et al., 2024)	Low-resource and graphene-based languages	DiT, LayoutMv3, YOLOv8	YOLOv8 achieved an 8.95% better IoU score than DiT and 38.48% better IoU score than LayoutLMv3.	low-resource and graphene-based language datasets (Bengali)
(Mechi et al., 2021)	Text line segmentation	FCN	Success rate 64.3%	Arabic and Latin document images
(Li et al., 2020)	Segment complex document layouts and heterogeneous content	CNN, GAT, CRF	PRImA: Recall ~0.92, Precision ~0.85, F-score ~0.88	PubLayNet dataset
(Mechi et al., 2019)	text line segmentation	CNN, U-net	cBAD 79%, DIVA-HisDB (CB55) 76%, ANT (Arabic) 76%	Tunisian National Archives images and datasets from the ICDAR and ICFHR competitions
(Ohyama et al., 2019)	Recognizing mathematical expressions	CNN, U-net	The highest F measure among results is 0.947±0.016	New scientific documents datasets GTDB-1 and GTDB-2
(Huang et al., 2019)	Improve table detection	YOLOv3, K-means clustering	ICDAR 2017 - IoU threshold of 0.6 and 0.8	Datasets from ICDAR 2013 and ICDAR 2017
(Banerjee et al., 2024)	Reduce dependence on large labeled datasets	Semi-supervised learning, SEN network	PRIMA: text 81.2, image 70.5, table 40.6, math 53.3, separator 26.1 accuracies.	PRIMA, DocLayNet, and Historical Japanese (HJ) datasets
(Aljiffry et al., 2022)	Improve layout analysis for Arabic documents	Faster R-CNN	Early printed dataset AvgF1 99.5%, printed dataset AvgF1 99.4%	Two distinct Arabic language datasets
(Gemelli et al., 2022)	Table extraction in PDF documents	GNN	F1 Score: base 0.855, padding 0.832	Merging PubLayNet and PubTables
(Zhang et al., 2022)	Document understanding	GAN, OCR	Experiments on the publicly available datasets	320k unlabeled documents
(De Nardin et al., 2024)	Ancient Arabic manuscripts segmentation	One-time learning approach	F-score: main text 0.989, side text 0.991, average 0.990	Ancient Arabic manuscripts
(Liu & Zhou, 2023)	Layout analysis and text recognition	CNN, U-net	F1-score 98.74, precision 98.88, recall 98.62	Handwritten and printed text dataset
(Grijalva et al., 2021)	identification of components in	Deep CNN	Overall accuracy 96.2685%	11,007 scientific papers

	scientific papers			
(Phong et al., 2020)	Automatic detection of mathematical expressions	CNN, Random Forest (RF)	Accuracy: Marmot dataset - 91.18% and 81.35% , GTDB dataset - 89.51% and 80.20%	Marmot and GTDB
(Kosaraju et al., 2019)	Document layout classification	CNN	Accuracy 96.27%	New English document image dataset
(Zhao et al., 2021)	Sub-line level segmentation	SOLOv2	91.18% and 81.35% on Marmot dataset, 89.51% and 80.20% on the GTDB	Kangyur documents
(Almutairi & Almashan, 2019)	Information extraction	Mask R-CNN	Accuracy 81.6%	Newspaper page images
(Ma et al., 2020)	Historical Tibetan document segmentation	Block projection, graph model	Average recall 86.60% and precision 30.25%	Historical Tibetan document images
(Oliveira & Viana, 2017)	Improve speed and efficiency in DLA	Fast 1D CNN	Overall accuracy 96.75%	English printed documents
(Rezanezhad et al., 2023)	Enhance OCR results	CNN, Heuristics	Detects Layout Classes (e.g. initials, marginals)	historical document dataset
(Nguyen et al., 2022)	Vietnamese document image understanding	CNN	AP scores: Table and figure > 85%, formula 50.1%, caption 76.2%	UIT-DODV dataset
(Yang & Li, 2022)	Improve automatic layout analysis	Mask R-CNN	Accuracy 89.3%	Chinese academic papers
(Minouei et al., 2021)	Improve optical character recognition	Deep CNN	Average accuracy 0.946	Large public dataset (English)
(Grüning et al., 2019)	Improve text line detection	ARU-Net	F value increased from 0.859 to 0.922 on the cBAD dataset	Historical documents from ICDAR2017 Competition

U-Net Architecture.

The U-Net architecture, introduced by Ronneberger, Fischer, and Brox, represents an important advancement in biomedical image segmentation. It combines a contracting path for context capture with a symmetric expanding path for precise localization, it enables efficient training from limited annotated data. Another thing is, that the U-Net architecture outperforms other methods. The architecture has achieved remarkable results in ISBI challenges and demonstrated rapid segmentation capabilities on contemporary GPUs (Ronneberger et al., 2015).

Furthermore, it has demonstrated its versatility across various domains. Bo Wang et al. developed MSNet, a deep learning-based multi-scale segmentation network designed to identify and classify regions of interest in complex and heterogeneous documents (Wang et al., 2021). MSNet effectively addresses challenges posed by documents with sparse structures and variable-scale typologies, highlighting the significance of advanced segmentation methods in improving document layout analysis in complex scenarios. Nabiee, et al. adapted U-Net to segment high-resolution satellite images, effectively detecting war-related destruction in Syria by introducing a multi-scale feature fusion approach (Nabiee et al., 2022). Moreover, Mechi, et al. proposed an adaptive U-Net for text line segmentation in historical document images. The method has shown the best performance in text line segmentation with qualitative and numerical results (Mechi et al., 2019).

Table 6 - Summary of results finding on U-Net architecture.

Ref	Purpose	Evaluation Metrics	Implementation	Dataset Size
(Mechi et al., 2019)	Text line segmentation	Precision (P), recall (R) and F-measure (F), intersection over union (IoU), the area under the curve (AUC)	Keras framework, TITAN X GPU with 12 GB memory	Training - 176, validation - 40, testing -539
(Ohyama et al., 2019)	Mathematical Expression detection	ME-based recall (Re), precision (Pe) and F-measure (Fe) as supplemental performance measures, mathematical symbol (character) recall Rs, precision Ps and F-measure Fs as the performance measures, Pixel-level majority voting for the symbol-level	Not mentioned	The dataset consists of 544 images
(Cao et al., 2022)	Medical Image Segmentation	Dice-Similarity coefficient (DSC), Hausdorff Distance (HD)	Python 3.6 and Pytorch 1.7.0, Nvidia V100 GPU with 32GB memory	Training - 88, testing - 32, validation - 10
(Nabiee	High-resolution	Unweighted mean IoU (mIoU), Frequency	Amazon Web Service	From 252 images, training -

et al., 2022)	satellite images segmentation	Weighed IoU (FWIoU), Mean Dice Similarity Coefficient (MDC), Jaccard index	(AWS) G4 instance with an NVIDIA T4 GPU with a 16 GB memory	60%, validation - 10%, testing - 30%
(Liu & Zhou, 2023)	Pixel-level segmentation	Precision, Recall, F1-score, and Intersection over Union (IoU)	Processor frequency - 3.00GHze, NVIDIA GPU 1050Ti	Training - 240, testing - 80
(Ma et al., 2018)	Document image unwraping	Multi-Scale Structural Similarity (MS-SSIM), SIFT flow	The network can run at 28 fps on a GTX 1080 Ti GPU	130 total images
(Dang & Nguyen, 2021)	Character-Level Embedding	Mean Intersection-over-Union (mIOU), mean pixel accuracy (pix acc) and the box F1-score	Nvidia Quadro M4000 with 8GB memory	Japanese invoice images 261, Insurance medical receipts 200: testing - 70%, training - 30%
(Lee et al., 2019)	Page segmentation	F-score and recall	Not mentioned	ICDAR2017 dataset: training - 1,600, testing - 817, validation - 300, PRimA dataset: training - 312, testing - 74, validation - 80
(Grüning et al., 2019)	Textline detection	Precision, Recall, F1-score	Titan X GPU, dual-core laptop (Intel Core i7-6600U with 16GiB RAM)	Training samples - 1024

Limitations and Conclusion

Limitations.

The limitations of this systematic literature review (SLR) are primarily related to the scope and focus of the research. The review was conducted using four specific scientific databases IEEE Xplore, ACM, Springer Link, and Science Direct which may limit the breadth of available studies.

Additionally, the analysis is intentionally narrowed to align with the specific research objective, and not all areas of the document layout analysis domain are covered. The search strings were also crafted with a targeted purpose, potentially excluding relevant studies that fall outside the defined scope. These limitations suggest that while the review provides valuable insights, it may not fully represent the broader field of document layout analysis.

Conclusion.

In conclusion, this systematic literature review (SLR) underscores the critical need for document layout analysis research, particularly for isolated languages like Sinhala, which have unique document structures that differ significantly from well-studied languages such as English. The review highlights that the absence of dedicated research on Sinhala document layout analysis poses a significant challenge, emphasizing the necessity for focused studies in this area. Furthermore, style analysis plays a vital role in accurately converting documents into editable digital formats, particularly when dealing with diverse font styles, lines, bullets, and numbering styles. These stylistic elements must be preserved to maintain the document's integrity during digitization.

This SLR also explores the potential of enhancing the U-Net architecture to better analyze complex document layouts and styles an approach not yet applied in existing research. The adaptability of U-Net in handling intricate and diverse layouts remains underexplored, making it a promising area for future development. Additionally, the review reveals that the performance of Optical Character Recognition (OCR) systems can be significantly improved through comprehensive document layout analysis, as OCR accuracy is heavily dependent on the correct identification and segmentation of individual document elements.

Accurate detection of all document elements is essential, yet many existing studies focus on a limited range of elements, often overlooking more complex components like diagrams and charts. These elements, frequently misclassified as simple images, possess distinct layouts, styles, and embedded texts that require specialized analysis techniques. Addressing these gaps will not only improve the fidelity of document digitization but also enhance the overall performance of OCR technologies. Therefore, advancing document layout and style analysis, particularly for underrepresented languages like Sinhala, is crucial for the future of accurate and comprehensive digital document processing.

REFERENCES:

1. Akanda, M. M. B. A. N., Ahmed, M., Rabby, A. S. A., & Rahman, F. (2024). Optimum Deep Learning Method for Document Layout Analysis in Low Resource Languages. Proceedings of the 2024 ACM Southeast Conference,
2. Aljiffry, L., Al-Barhamtoshy, H., Jamal, A., & Abukhodair, F. (2022). Arabic Documents Layout Analysis (ADLA) using Fine-tuned Faster RCN. 2022 20th International Conference on Language Engineering (ESOLEC),

3. Almutairi, A., & Almashan, M. (2019). Instance segmentation of newspaper elements using mask R-CNN. 2019 18th IEEE International conference on machine learning and applications (ICMLA),
4. Banerjee, A., Biswas, S., Lladós, J., & Pal, U. (2024). SemiDocSeg: harnessing semi-supervised learning for document layout analysis. International Journal on Document Analysis and Recognition (IJDAR), 1-18.
5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. European conference on computer vision,
6. Cuo, Y., Tashi, N., Liu, Z., Wei, Q., Gadeng, L., & Trashi, G. (2019). Layout Analysis of Tibetan Historical Documents Based on Deep Learning Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence, Wenzhou, China. https://doi.org/10.1145/3357777.3357790
7. Dang, T.-A. N., & Nguyen, D.-T. (2021). End-to-end information extraction by character-level embedding and multi-stage attentional u-net. arXiv preprint arXiv:2106.00952.
8. De Nardin, A., Zottin, S., Piciarelli, C., Colombi, E., & Foresti, G. L. (2024). A One-Shot Learning Approach to Document Layout Segmentation of Ancient Arabic Manuscripts. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,
9. Gemelli, A., Vivoli, E., & Marinai, S. (2022). Graph neural networks and representation embedding for table extraction in PDF documents. 2022 26th International Conference on Pattern Recognition (ICPR),
10. Grijalva, F., Santos, E., Acuña, B., Rodríguez, J. C., & Larco, J. C. (2021). Deep learning in time-frequency domain for document layout analysis. IEEE Access, 9, 151254-151265.
11. Grüning, T., Leifert, G., Strauß, T., Michael, J., & Labahn, R. (2019). A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR), 22(3), 285-302.
12. Huang, Y., Yan, Q., Li, Y., Chen, Y., Wang, X., Gao, L., & Tang, Z. (2019). A YOLO-based table detection method. 2019 International Conference on Document Analysis and Recognition (ICDAR),
13. Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. Information and software technology, 51(1), 7-15.
14. Kosaraju, S. C., Masum, M., Tsaku, N. Z., Patel, P., Bayramoglu, T., Modgil, G., & Kang, M. (2019). DoT-Net: Document layout classification using texture-based CNN. 2019 International Conference on Document Analysis and Recognition (ICDAR),
15. Lee, J., Hayashi, H., Ohyama, W., & Uchida, S. (2019). Page segmentation using a convolutional neural network with trainable co-occurrence features. 2019 International conference on document analysis and recognition (ICDAR),
16. Li, X.-H., Yin, F., & Liu, C.-L. (2020). Page segmentation using convolutional neural network and graphical model. Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14,
17. Liu, D., & Zhou, S. (2023). Pixel-Level Segmentation of Handwritten and Printed Texts in Document Images with Deep Learning. 2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI),
18. Ma, K., Shu, Z., Bai, X., Wang, J., & Samaras, D. (2018). Docunet: Document image unwarping via a stacked u-net. Proceedings of the IEEE conference on computer vision and pattern recognition,
19. Ma, L., Long, C., Duan, L., Zhang, X., Li, Y., & Zhao, Q. (2020). Segmentation and recognition for historical Tibetan document images. IEEE Access, 8, 52641-52651.
20. Mechi, O., Mehri, M., Ingold, R., & Amara, N. E. B. (2019). Text line segmentation in historical document images using an adaptive u-net architecture. 2019 International Conference on Document Analysis and Recognition (ICDAR),
21. Mechi, O., Mehri, M., Ingold, R., & Essoukri Ben Amara, N. (2021). A two-step framework for text line segmentation in historical Arabic and Latin document images. International Journal on Document Analysis and Recognition (IJDAR), 24(3), 197-218.
22. Minouei, M., Soheili, M. R., & Stricker, D. (2021). Document layout analysis with an enhanced object detector. 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA),
23. Nabiee, S., Harding, M., Hersh, J., & Bagherzadeh, N. (2022). Hybrid U-Net: Semantic segmentation of high-resolution satellite images to detect war destruction. Machine Learning with Applications, 9, 100381.
24. Nguyen, K., Nguyen, A., Vo, N. D., & Nguyen, T. V. (2022). Vietnamese document analysis: dataset, method and benchmark suite. IEEE Access, 10, 108046-108066.
25. Ohyama, W., Suzuki, M., & Uchida, S. (2019). Detecting mathematical expressions in scientific document images using a u-net trained on a diverse dataset. IEEE Access, 7, 144030-144042.
26. Oliveira, D. A. B., & Viana, M. P. (2017). Fast CNN-based document layout analysis. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW),
27. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A. S., & Staar, P. (2022). Doclaynet: A large human-annotated dataset for document-layout segmentation. Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining,
28. Phong, B. H., Hoang, T. M., & Le, T.-L. (2020). A hybrid method for mathematical expression detection in scientific document images. IEEE Access, 8, 83663-83684.
29. Rezanezhad, V., Baierer, K., Gerber, M., Labusch, K., & Neudecker, C. (2023). Document Layout Analysis with Deep Learning and Heuristics. Proceedings of the 7th International Workshop on Historical Document Imaging and Processing,
30. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18,
31. Shehzadi, T., Stricker, D., & Afzal, M. Z. (2024). A hybrid approach for document layout analysis in document images. arXiv preprint arXiv:2404.17888.

32. Sheikh, T. U., Shehzadi, T., Hashmi, K. A., Stricker, D., & Afzal, M. Z. (2024). UnSupDLA: Towards Unsupervised Document Layout Analysis. arXiv preprint arXiv:2406.06236.
33. Umer, S., Mondal, R., Pandey, H. M., & Rout, R. K. (2021). Deep features based convolutional neural network model for text and non-text region segmentation from document images. Applied Soft Computing, 113, 107917.
34. Wang, B., Zhou, J., & Zhang, B. (2021). MSNet: A multi-scale segmentation network for documents layout analysis. Learning Technologies and Systems: 19th International Conference on Web-Based Learning, ICWL 2020, and 5th International Symposium on Emerging Technologies for Education, SETE 2020, Ningbo, China, October 22–24, 2020, Proceedings 5,
35. Wu, X., Zheng, Y., Ma, T., Ye, H., & He, L. (2021). Document image layout analysis via explicit edge embedding network. Information Sciences, 577, 436-448.
36. Yang, Z., & Li, N. (2022). Identification of Layout elements in Chinese academic papers based on Mask R-CNN. 2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE),
37. Zhang, Z., Ma, J., Du, J., Wang, L., & Zhang, J. (2022). Multimodal pre-training based on graph attention network for document understanding. IEEE Transactions on Multimedia, 25, 6743-6755.
38. Zhao, P., Wang, W., Cai, Z., Zhang, G., & Lu, Y. (2021). Accurate fine-grained layout analysis for the historical Tibetan document based on the instance segmentation. IEEE Access, 9, 154435-154447.
39. Zhong, X., Tang, J., & Yepes, A. J. (2019). Publaynet: largest dataset ever for document layout analysis. 2019 International conference on document analysis and recognition (ICDAR),