**Research Proposal**

|  |  |  |  |
| --- | --- | --- | --- |
| Student Name | H.S.K. Dilukshi | Student Index Number | 18APC3590 |
| Research Topic | Enhanced U-Net Architecture for Analyzing Complex Sinhala Document Layouts and Styles | | |
| Supervisor(s) | Prof. B.T.G.S. Kumara | | |

|  |
| --- |
| **Agreement** |
| I …H.S.K. Dilukshi……, willing to submit the following Proposal format and I am confirming that this topic is a novel research and I am aware of the consequences occurs in the case of plagiarism found.  ……………………………. ………19/08/2024………  Signature Date |

|  |
| --- |
| Background |
| Document layout analysis (DLA) plays an important role across various domains and sectors by enabling efficient digitization [1], data extraction [2], and information retrieval [3] from document images. Comprehensive DLA benefits various sectors such as government, healthcare, legal services, finance, and education [4-9]. Furthermore, DLA is essential for digitizing documents such as historical and cultural books, academic books and papers, old books, newspaper articles, etc. [10-15]. This automation increases data acquisition efficiency and ensures that valuable information can be accessed quickly and accurately. DLA is fundamental to the digitization process, as it preserves the structural integrity of documents, making them accessible for future use [16]. Moreover, A major challenge for low-resource languages (LRL) is the limited availability of high-quality, annotated datasets [7]. Unlike high-resource languages such as English, French, or Spanish, which benefit from extensive text and document data collections [17-19], LRLs often struggle with insufficient resources for developing and compiling adequate data for language processing models.  For the Sinhala language, accurate document layout analysis and font style identification are essential for effective digitization. Sinhala, with its unique script and complex structure, presents distinct challenges that make it critical to accurately capture and analyze the layout of documents [20]. Each element, such as titles, paragraphs, tables, figures, etc., must be precisely identified to ensure that the document's structure is protected during digitization [21]. Furthermore, recognizing font styles, such as bold, italic, and underlined text, is essential in maintaining the semantic nuances and emphasis intended by the original document creators. The ability to digitize Sinhala documents accurately will not only preserve cultural and historical content but also facilitate broader accessibility and research, ensuring that these documents can be used effectively in the digital age.  Document layout analysis and style identification present several challenges, particularly when it comes to capturing the various styles within document elements. One significant challenge in existing research is the lack of focus on identifying specific font and line styles, such as bold, italic, underlined, dashed, or dotted lines [1-19], which are essential for the accurate representation of document content. Additionally, the identification of numbering and bulleting styles is a complex task that has not been adequately addressed due to the variability of styles across different elements [21, 22]. |

|  |
| --- |
| Most studies on document layout analysis have limited their scope to a few components like text, images, and tables, with minimal attention given to more complex elements such as diagrams, charts, and multi-level headings [1-31]. This oversight makes digitizing documents, especially those with varied content such as diagrams and charts, particularly challenging. The lack of research on topic-level labeling categorized into topic levels (e.g., Level 1, Level 2, and Level 3), further complicates the process of creating a structured and accessible digital document [1-31].  The U-Net architecture, first introduced for biomedical image segmentation [32], has become a fundamental model in the field due to its ability to accurately outline the structures of medical images with limited training data [33]. Furthermore, U-Net's success in biomedical tasks such as tumor detection [34] and organ segmentation [35]stems 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 [36]. U-Net has potential in the field of document analysis to segment and analyze document layouts [37-39], as well as extract lines of text and other structural components [11]. However, its application to complex structured documents such as multi-columns, mixed fonts, and complex layouts remains limited and under-explored. The strength of the U-Net architecture lies in its structure, especially its skip connection model, which can hold spatial information across multiple layers. This allows U-Net to achieve high accuracy in tasks requiring precise localization, making it a powerful tool for image and document segmentation. Extending its use to more complex document structures can open up new possibilities in the digitization of diverse and complex document types [32]. Expanding its use to more complex document structures could unlock new possibilities in the digitization of diverse and intricate document types.  To overcome the challenges of character recognition and extraction, Optical Character Recognition (OCR) is an important component that makes digitization an approachable problem [40]. After analyzing document layouts and styles, applying OCR software is essential to extract characters. This process converts the characters into machine-readable format, enabling them to be edited, searched, and stored digitally. The most famous and accurate OCR software is Tesseract OCR and its applicability to different languages ​​is also important to address the Sinhala language [41]. For Sinhala documents, the application of OCR is very important, as it enables the recognition and translation of Sinhala characters from the fragmented text regions identified during the layout analysis stage [42-44]. This step is critical to creating complete digital documents that retain not only the visual and structural elements of the original but also the textual content in a searchable and editable format. OCR, combined with accurate layout analysis and style recognition, ensures that digitized documents are a faithful representation of the originals, facilitating their use in digital archives, academic research, and everyday access. |

|  |
| --- |
| Literature Review |
| **Document Layout and Style Analysis**  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. [31] 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. [14] 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. [11] 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. [28] 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 [18] 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. 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, DocLayNet, and PubTables benchmarks [45, 46]. Moreover, identify an optimal deep learning approach for document layout analysis (DLA) in low-resource and grapheme-based languages Md. Mutasim Billah, et al. [23] 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. [21] 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. 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 [15]. Yilun Huang, et al. [26] proposed a YOLO-based method to improve table detection in document analysis. Other than that, to improve the identification and classification of components in scientific papers Felipe Grijalva, et al. [4] 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. |

|  |
| --- |
| Furthermore, Sai Chandra, et al. [6] developed a texture-based CNN model for document layout classification, capable of identifying various document component blocks like text, images, tables, mathematical expressions, and line diagrams. In past decades, Graph-Based and GNN [24-27] methods were investigated to improve table extraction in PDF documents by addressing challenges like OCR errors, especially in text inside tables. Zhenrong Zhang, et al. [3] introduced a Graph Attention Network (GAN) based method to enhance document understanding across various formats and layouts using a multimodal approach.**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 [13, 16, 22, 29, 30], Abdullah Almutairi, et al. [2] used Mask R-CNN to enhance information extraction and develop a deep learning model to segment newspaper pages into articles, advertisements, and headers. Furthermore, to address the challenge of relying on labeled data Talha et, al. [19] used vision-based unsupervised pre-training to create object masks from unlabeled images and then train a detector for better object detection and segmentation. They have achieved significant improvements in accuracy and efficiency without the need for label data. Moreover, Ziyi Yang, et al. [9] 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. Achieves 89.3% accuracy in recognizing layout elements by enhancing practical applications. They have used a custom layout image dataset of Chinese academic papers. **U-net Architecture** The U-Net architecture, introduced by Ronneberger, Fischer, and Brox [32], 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. Furthermore, it has demonstrated its versatility across various domains. Nabiee, et al. [36] adapted U-Net to segment high-resolution satellite images, effectively detecting war-related destruction in Syria by introducing a multi-scale feature fusion approach. Moreover, Mechi, et al. [11] 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.  **Sinhala Language**  In the past decade, researchers have significantly contributed to the advancement of Sinhala script recognition. K.D. Thamarasee, et al. [44] have proposed a method using the Histogram of Oriented Gradients (HOG) for feature extraction and the Support Vector Machine (SVM) for classification. This research has contributed significantly to non-Latin script OCR and regional language digitization. In addition to that, Isuri Anuradha, et al. [42] enhanced Sinhala OCR by integrating deep learning with Tesseract by addressing complex curves and diacritical marks and exploring the cultural and global digitization significance of specialized OCR solutions. Sinhala's handwritten character recognition challenge was addressed by Janotheepan Mariyathas, et al. [20] by proposing a CNN-based method to identify complex characters. This approach achieved an accuracy of 90.27% for 100 characters and 82.33% for 434 characters. Furthermore, H. Waruna, et al. [43] improved Sinhala OCR by combining character geometry with ANNs. That method outperformed both standard methods and advanced local language digitalization. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Summary of prior research selected on document layout analysis and segmentation   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Ref** | **Focus** | **Technique** | **Improvements** | **Dataset** | | [1] | Digitization and subsequent text recognition | CTPN algorithm | Accuracy: horizontal row 0.98, vertical row 0.83 | Tibetan historical documents | | [18] | 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 | | [19] | 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 | | [21] | 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 | | [31] | 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 | | [14] | Multi-scale segmentation | CNN, MSNet | Pixel Accuracy 96.61%, Mean IoU 90.55% | Chinese document dataset | | [23] | 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) | | [12] | Text line segmentation | FCN | Success rate 64.3% | Arabic and Latin document images | | [27] | Segment complex document layouts and heterogeneous content | CNN, GAT, CRF | PRImA: Recall ~0.92, Precision ~0.85, F-score ~0.88 | PubLayNet dataset | | [11] | 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 | | [13] | 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 | | [26] | Improve table detection | YOLOv3, K-means clustering | ICDAR 2017 - IoU threshold of 0.6 and 0.8 | Datasets from ICDAR 2013 and ICDAR 2017 | | [22] | 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 | | [24] | 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 | | [25] | Table extraction in PDF documents | GNN | F1 Score: base 0.855, padding 0.832 | Merging PubLayNet and PubTables | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | [3] | Document understanding | GAN, OCR | Experiments on the publicly available datasets | 320k unlabeled documents | | [10] | Ancient Arabic manuscripts segmentation | One-time learning approach | F-score: main text 0.989, side text 0.991, average 0.990 | Ancient Arabic manuscripts | | [28] | Layout analysis and text recognition | CNN, U-net | F1-score 98.74, precision 98.88, recall 98.62 | Handwritten and printed text dataset | | [4] | identification of components in scientific papers | Deep CNN | Overall accuracy 96.2685% | 11,007 scientific papers | | [8] | 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 | | [6] | Document layout classification | CNN | Accuracy 96.27% | New English document image dataset | | [15] | Sub-line level segmentation | SOLOv2 | 91.18% and 81.35% on Marmot dataset, 89.51% and 80.20% on the GTDB | Kangyur documents | | [2] | Information extraction | Mask R-CNN | Accuracy 81.6% | Newspaper page images | | [29] | Historical Tibetan document segmentation | Block projection, graph model | Average recall 86.60% and precision 30.25% | Historical Tibetan document images | | [17] | Improve speed and efficiency in DLA | Fast 1D CNN | Overall accuracy 96.75% | English printed documents | | [16] | Enhance OCR results | CNN, Heuristics | Detects Layout Classes (e.g. initials, marginals) | historical document dataset | | [7] | Vietnamese document image understanding | CNN | AP scores: Table and figure > 85%, formula 50.1%, caption 76.2% | UIT-DODV dataset | | [9] | Improve automatic layout analysis | Mask R-CNN | Accuracy 89.3% | Chinese academic papers | | [30] | Improve optical character recognition | Deep CNN | Average accuracy 0.946 | Large public dataset (English) | | [5] | Improve text line detection | ARU-Net | F value increased from 0.859 to 0.922 on the cBAD dataset | Historical documents from ICDAR2017 Competition |   Table 1 summarizes the SLR on document layout analysis and segmentation details, highlighting the focus of each study, the main techniques employed, the improvements achieved, and the datasets used. The findings indicate that isolated languages have received less attention, with most researchers concentrating on English. In addition to that, Table 2 provides the summary of findings on U-Net, including evaluation metrics, implementation details, and dataset sizes. Table 2 shows that U-Net can achieve better results even with smaller datasets. However, within the domain of document layout analysis, U-Net architecture is not applied to segment a broader range of elements in complex documents. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2:Summary of prior research selected on U-Net architecture   |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **Dataset Size** | Training - 750 images, validation - 177 images | Training - 176, validation - 40, testing -539 | The dataset consists of 544 images | Training - 88, testing - 32, validation - 10 | From 252 images, training - 60%, validation - 10%, testing - 30% | Training - 240, testing - 80 | 130 total images | Japanese invoice images 261, Insurance medical receipts 200: testing - 70%, training - 30% | ICDAR2017 dataset: training - 1,600, testing - 817, validation - 300, PRimA dataset: training - 312, testing - 74, validation - 80 | Training samples - 1024 | | **Implementation** | PyTorch-1.5 framework, TITAN-RTX GPU with 24 GB memory | Keras framework, TITAN X GPU with 12 GB memory | Not mentioned | Python 3.6 and Pytorch 1.7.0, Nvidia V100 GPU with 32GB memory | Amazon Web Service (AWS) G4 instance with an NVIDIA T4 GPU with a 16 GB memory | Processor frequency - 3.00GHze, NVIDIA GPU 1050Ti | The network can run at 28 fps on a GTX 1080 Ti GPU | Nvidia Quadro M4000 with 8GB memory | Not mentioned | Titan X GPU, dual-core laptop (Intel Core i7-6600U with 16GiB RAM) | | **Evaluation Metrics** | pixel-wise accuracy (Pixel Acc), mean of the classes-wise Intersection over Union averaged (Mean IoU) | Precision (P), recall (R) and F-measure (F), intersection over union (IoU), the area under the curve (AUC) | 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 | Dice-Similarity coefficient (DSC), Hausdorff Distance (HD) | Unweighted mean IoU (mIoU), Frequency Weighed IoU (FWIoU), Mean Dice Similarity Coefficient (MDC), Jaccard index | Precision, Recall, F1-score, and Intersection over Union (IoU) | Multi-Scale Structural Similarity (MS-SSIM), SIFT flow | mean Intersection-over-Union (mIOU), mean pixel accuracy (pix acc) and the box F1-score | F-score and recall | Precision, Recall, F1-score | | **Purpose** | Multi-scale segmentation | Text line segmentation | Mathematical Expression detection | Medical image segmentation | High-resolution satellite images segmentation | Pixel-level segmentation | Document image unwraping | Character-Level Embedding | Page segmentation | Textline detection | | **Ref** | [14] | [11] | [13] | [33] | [36] | [28] | [39] | [37] | [38] | [5] | |

|  |
| --- |
| Gaps in Existing Research |
| Existing research on document layout analysis has mainly focused on the English language, with limited involvement given to the isolated languages. Also, no research is dedicated to the Sinhala language documents in this context. According to the analysis of language-specific results, Figure 1 shows that 46% of contributions are focused on English [3, 4, 6, 8, 14, 17-19, 21, 22, 31], 17% on Arabic [10, 12, 24], 13% on Tibetan [1, 15, 29], 8% on Chinese [9, 14, 28] and the remaining 17% on other languages [7, 23]. This exploration highlights a significant gap in document layout analysis for the Sinhala language. Additionally, there is no publicly available dataset of Sinhala document images related to Document Layout Analysis (DLA). Furthermore, digitization is a challenging task that cannot be accomplished alone with Optical Character Recognition (OCR) software without comprehensive DLA [42-44].  Figure 1:Classification of Language specific results  Regarding style capturing, researchers have not specifically focused on identifying styles within document elements, such as font styles (e.g. bold, italic, underlined) or line styles (e.g. dashed, dotted, solid, normal, and double) [1-19, 21-31]. Moreover, there is no research addressing the identification of numbering and bulleting styles because of the difficulty of the task. The reason is that there are different styles for various elements. When analyzing document image elements, most studies have focused on a limited number of components, primarily text, images, tables, formulas, headers, and footers [1-14, 17-19, 21, 22, 24-26, 31, 38, 39]. Only one study was found that considered a broader range of labels, including captions, footnotes, formulas, list items, footers, page headers, pictures, section headers, tables, text, and titles [22]. However, digitization still faces challenges in identifying diagrams and charts, as existing research on these elements is limited and fails to consider them in combination with a range of elements [6]. Most research tends to segment diagrams and charts as figures, the reason is that the diagrams and charts have different styles [7, 14, 21, 22]. Therefore, digitization is a challenging task without proper analysis of the diagrams and charts. Additionally, there is a lack of research on title-level labeling, which means identification of titles as level 1, level 2, and level 3 headings.  Furthermore, Complex document layout analysis shows several limitations in existing research. Although U-Net architecture has not yet been applied to complex document layouts that involve a wider range of elements [5, 11, 13, 14, 37-39]. Another thing is that it also faces challenges in handling various document structures and fine-grained segmentation tasks. Existing U-Net models often struggle with analyzing space arrangements and relationships between different layout elements which leads to less accurate segmentation results [39]. Therefore, developing an improved U-Net architecture adopts to complex document layouts is important to address these limitations and enhance the accuracy and effectiveness of the context. |

|  |
| --- |
| Problem Statement and Research Questions |
| **Problem Statement**  Digitizing document images, especially in low-resource languages like Sinhala, faces challenges due to the absence of document layout and style analysis. Current methods struggle with accurately identifying various document elements such as titles (Level 1, Level 2, and Level 3), paragraphs, figures, tables, captions, formulas, headers, footers, diagrams, bullet lists, and numbered lists, such as titles level-wise, diagrams, and font styles, line styles. To address these issues, this research proposes using an enhanced U-Net architecture for improved document layout and style analysis, even with limited datasets. This approach, supported by the Tesseract OCR for Sinhala character recognition, aims to provide accurate digitization of complex documents and could benefit other low-resource languages as well.  In detail, The primary reason for the absence of document layout and style analysis in Sinhala documents is the complexity of their structure and text. Additionally, there is no available benchmark dataset for this purpose. Title-level labeling poses a significant challenge across various languages due to the diverse sizes and styles used by different authors. Accurately identifying elements like diagrams in complex document images remains problematic. Style analysis within document elements, such as font and line styles, is also underexplored. Although Optical Character Recognition (OCR) research has been conducted for Sinhala, OCR alone is insufficient for comprehensive digitization.  **Research Questions**  The first research question focuses on layout and style analysis, aiming to identify elements such as titles (Level 1, Level 2, and Level 3), paragraphs, figures, tables, captions, formulas, headers, footers, diagrams, bullet lists, numbered lists, font styles, and line styles. The second question addresses the development of an enhanced U-Net architecture to tackle the challenges posed by complex document layouts. The third question pertains to the second step of digitization, specifically character identification for the analyzed text elements.  RQ1: How can the layouts and styles of complex document images be accurately analyzed?  RQ2: What improvements are required in the U-Net architecture for analyzing complex documents?  RQ3: How can Optical Character Recognition (OCR) technology be applied to analyzed Sinhala documents? |

|  |
| --- |
| Research Objectives and Goals |
| **Main Objective**  The main objective of this research is to digitize complex Sinhala document images by effectively analyzing their layouts and styles using an enhanced U-Net architecture, with support of Tesseract OCR for character recognition.  **Specific Objectives**  The first objective is to analyze elements such as titles (Level 1, Level 2, and Level 3), paragraphs, figures, tables, captions, formulas, headers, footers, diagrams, bullet lists, and numbered lists. Additionally, it involves analyzing font styles like bold, italic, normal, and underlined, as well as line styles such as dashed, dotted, solid, normal, and double. The second objective focuses on developing an enhanced U-Net architecture to address the challenges of complex Sinhala document images, with potential applicability to other low-resource languages. Furthermore, the third objective aims to complete the digitization process by applying optical character recognition (OCR) technologies to the analyzed elements, with the support of Tesseract OCR. These OCR technologies will be applied only to text-containing elements, excluding images. Additionally, this research aims to publish a comprehensive new benchmark dataset of Sinhala document images for future research. The relationships among the problem statement, main objective, research questions, and specific objectives are illustrated in Figure 2.  RO1: To effectively analyze the layouts and styles of complex Sinhala document images.  RO2: To develop an enhanced U-Net architecture for accurate analysis of complex document images.  RO3: To digitize various selected categories of Sinhala documents.    Figure 2: Research questions and objectives mapping |

|  |
| --- |
| Proposed Research Method |
| **Data Collection**  Sinhala document image datasets are not publicly available for research purposes, necessitating manual collection from various sources. To ensure the dataset's relevance and comprehensiveness, it must include all elements targeted for analysis in this research, such as titles (Level 1, Level 2, and Level 3), paragraphs, figures, tables, captions, formulas, headers, footers, diagrams, bullet lists, and numbered lists. The collection of these images must adhere to legal and ethical standards, particularly regarding copyright compliance. This involves obtaining proper permissions where necessary and ensuring that the use of materials from published sources respects intellectual property rights. The dataset will be sourced from diverse Sinhala materials, including old books, religious books, government publications, academic books, literary books, biographies and autobiographies, fictional novels and short story books, and cultural books. Additionally, images will be captured from materials available in public libraries, which offer a rich variety of documents critical for a comprehensive dataset. As demonstrated in the Systematic Literature Review (SLR), the U-Net architecture is capable of delivering robust results even with smaller datasets [5, 11, 13, 14, 37-39]. Therefore, this research aims to compile a dataset of 500-700 document images. To enhance the accuracy and generalizability of the model, images will be captured using five different mobile phone cameras, addressing variations in color correction and lighting conditions.  **Methodology**  The methodology starts with the data collection step and will follow the aforementioned data collection process to collect a complex Sinhala document image dataset. The next step is data annotation. At this stage, this research aims to carefully annotate each document image to identify and label various elements such as titles (Level 1, Level 2, and Level 3), paragraphs, figures, tables, captions, formulas, headers, footers, diagrams, bulleted lists, and numbered lists. Additionally, different font styles and line styles within the main elements will need to be labeled separately. These annotations are essential for the subsequent semantic segmentation process, and the U-Net architecture will be used to accurately analyze and segment these labeled elements [32]. Then conducting preprocessing of the collected data ensures consistency and quality across the dataset, including noise reduction, contrast enhancement, and normalization techniques. Image augmentation methods, such as rotation, flipping, and scaling, further enhance the dataset by increasing its diversity and robustness, which is important for effective model training.  The most important step in this research is to design an improved U-Net architecture, this is an iterative process until sufficient accuracy is found. To design the U-Net architecture, this research will follow the findings from a systematic literature review (SLR) related to U-Net architecture across different domains mainly focusing on document image tasks [5, 11, 13, 14, 37-39]. The encoder-decoder architecture of U-Net is very important to solve this task. Furthermore, the encoder and decoder capture the high-level features of the document and reconstruct these features to their original spatial dimensions [36]. The use of skip connections in U-Net allows the preservation of spatial information that is crucial for the correct segmentation of complex document elements. After preprocessing images, The dataset is divided into three subsets such as training, validation, and testing. The training subset is used for model training and the validation subset is used for hyperparameter tuning. |

|  |
| --- |
| Furthermore, the model performance evaluation and the test subset are reserved for reporting final performance metrics. Evolution metrics such as Precision (P), recall (R), F1-score (F1), and intersection over union (IoU) will follow to evaluation [5, 11, 28]. This methodology aims to perform optical character recognition (OCR) on text regions found after document image segmentation [40]. It is important to perform this step to correctly extract Sinhala characters while maintaining the font styles (bold, italic, underline). For OCR, this research aims to address the updated version of Tesseract OCR for the Sinhala language [41]. The result is converting Sinhala documents into editable digital formats to ensure the desired results. The methodology involves the creation and release of a benchmark Sinhala document dataset that can be used for future research purposes.    Figure 3: Proposed Methodology |

|  |
| --- |
| Expected Outcome |
| **Main Outcome**  The main outcome of this research is to digitize complex Sinhala document images by effectively analyzing their layouts and styles using an enhanced U-Net architecture, with support of Tesseract OCR for character recognition.  **Specific Outcomes**  This research aims to accurately identify and annotate various elements within Sinhala document images, including titles (Level 1, Level 2, and Level 3), paragraphs, figures, tables, captions, formulas, headers, footers, diagrams, bullet lists, and numbered lists. This comprehensive annotation process will help to analyze complex document structures.  In addition to layout identification, the research will focus on analyzing font styles such as bold, italic, normal, and underlined and line styles including dashed, dotted, solid, normal, and double, within the identified elements. This thorough style analysis will improve the precision of the digital conversion process, enabling a more accurate reproduction of the original document's visual and structural features.  By developing an improved U-Net architecture, this research will address the challenges associated with segmenting complex document layouts. The enhanced architecture will be specifically tailored to the unique characteristics of Sinhala documents, while also being adaptable for other low-resource languages. This will contribute to advancements in the field of document image analysis. Also, it will set a foundation for future research in similar domains.  The scope of this research extends to digitizing a wide range of Sinhala document images, including old books, religious books, academic books, literary books, biographies, autobiographies, fictional novels, short story books, and cultural books. This effort will play an important role in preserving important books.  Another significant outcome of this research will be the publication of a new benchmark dataset of Sinhala document images. This dataset will serve as a valuable resource for future research purpose in Optical Character Recognition (OCR) and document layout analysis for the Sinhala language. |

|  |
| --- |
| Research Plan and Time Table |
| Table 3: Research plan   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Activity | Time Duration | | | | | | | | | | | | | | | | | | | | | | | | | | | | | June | | | | July | | | | August | | | | September | | | | October | | | | November | | | | December | | | | | 1st Week | 2nd Week | 3rd Week | 4th Week | 1st Week | 2nd Week | 3rd Week | 4th Week | 1st Week | 2nd Week | 3rd Week | 4th Week | 1st Week | 2nd Week | 3rd Week | 4th Week | 1st Week | 2nd Week | 3rd Week | 4th Week | 1st Week | 2nd Week | 3rd Week | 4th Week | 1st Week | 2nd Week | 3rd Week | 4th Week | | Scope Review |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Systematic Literature Review (SLR) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Reporting SLR |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Proposal Preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Data Collection |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Data Annotation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Data Preprocessing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Model Architecture Design |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Model Training |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Evaluation and Benchmark dataset |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | Thesis Preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |

|  |
| --- |
| Proposed Budget |
| **Internet Connection Cost**  Monthly cost = 300 LKR  Total cost for 7 months = 300\*7 = 2100 LKR  **Publication Cost**  The cost will vary depending on the selected publication journal. |

|  |
| --- |
| References |
| [1] Y. Cuo, N. Tashi, Z. Liu, Q. Wei, L. Gadeng, and G. Trashi, "Layout Analysis of Tibetan Historical Documents Based on Deep Learning," presented at the Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence, Wenzhou, China, 2019. [Online]. Available: <https://doi.org/10.1145/3357777.3357790>.  [2] A. Almutairi and M. Almashan, "Instance segmentation of newspaper elements using mask R-CNN," in *2019 18th IEEE International conference on machine learning and applications (ICMLA)*, 2019: IEEE, pp. 1371-1375.  [3] Z. Zhang, J. Ma, J. Du, L. Wang, and J. Zhang, "Multimodal pre-training based on graph attention network for document understanding," *IEEE Transactions on Multimedia,* vol. 25, pp. 6743-6755, 2022.  [4] F. Grijalva, E. Santos, B. Acuña, J. C. Rodríguez, and J. C. Larco, "Deep learning in time-frequency domain for document layout analysis," *IEEE Access,* vol. 9, pp. 151254-151265, 2021.  [5] T. Grüning, G. Leifert, T. Strauß, J. Michael, and R. Labahn, "A two-stage method for text line detection in historical documents," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 22, no. 3, pp. 285-302, 2019.  [6] S. C. Kosaraju *et al.*, "DoT-Net: Document layout classification using texture-based CNN," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019: IEEE, pp. 1029-1034.  [7] K. Nguyen, A. Nguyen, N. D. Vo, and T. V. Nguyen, "Vietnamese document analysis: dataset, method and benchmark suite," *IEEE Access,* vol. 10, pp. 108046-108066, 2022.  [8] B. H. Phong, T. M. Hoang, and T.-L. Le, "A hybrid method for mathematical expression detection in scientific document images," *IEEE Access,* vol. 8, pp. 83663-83684, 2020.  [9] Z. Yang and N. Li, "Identification of Layout elements in Chinese academic papers based on Mask R-CNN," in *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2022: IEEE, pp. 250-255.  [10] A. De Nardin, S. Zottin, C. Piciarelli, E. Colombi, and G. L. Foresti, "A One-Shot Learning Approach to Document Layout Segmentation of Ancient Arabic Manuscripts," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 8127-8136.  [11] O. Mechi, M. Mehri, R. Ingold, and N. E. B. Amara, "Text line segmentation in historical document images using an adaptive u-net architecture," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019: IEEE, pp. 369-374.  [12] O. Mechi, M. Mehri, R. Ingold, and N. Essoukri Ben Amara, "A two-step framework for text line segmentation in historical Arabic and Latin document images," *International Journal on Document Analysis and Recognition (IJDAR),* vol. 24, no. 3, pp. 197-218, 2021.  [13] W. Ohyama, M. Suzuki, and S. Uchida, "Detecting mathematical expressions in scientific document images using a u-net trained on a diverse dataset," *IEEE Access,* vol. 7, pp. 144030-144042, 2019.  [14] B. Wang, J. Zhou, and B. Zhang, "MSNet: A multi-scale segmentation network for documents layout analysis," in *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*, 2021: Springer, pp. 225-235.  [15] P. Zhao, W. Wang, Z. Cai, G. Zhang, and Y. Lu, "Accurate fine-grained layout analysis for the historical Tibetan document based on the instance segmentation," *IEEE Access,* vol. 9, pp. 154435-154447, 2021.  [16] V. Rezanezhad, K. Baierer, M. Gerber, K. Labusch, and C. Neudecker, "Document Layout Analysis with Deep Learning and Heuristics," in *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, 2023, pp. 73-78.  [17] D. A. B. Oliveira and M. P. Viana, "Fast CNN-based document layout analysis," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017: IEEE, pp. 1173-1180.  [18] T. Shehzadi, D. Stricker, and M. Z. Afzal, "A hybrid approach for document layout analysis in document images," *arXiv preprint arXiv:2404.17888,* 2024.  [19] T. U. Sheikh, T. Shehzadi, K. A. Hashmi, D. Stricker, and M. Z. Afzal, "UnSupDLA: Towards Unsupervised Document Layout Analysis," *arXiv preprint arXiv:2406.06236,* 2024.  [20] J. Mariyathas, V. Shanmuganathan, and B. Kuhaneswaran, "Sinhala handwritten character recognition using convolutional neural network," in *2020 5th International Conference on Information Technology Research (ICITR)*, 2020: IEEE, pp. 1-6. |

|  |
| --- |
| [21] X. Wu, Y. Zheng, T. Ma, H. Ye, and L. He, "Document image layout analysis via explicit edge embedding network," *Information Sciences,* vol. 577, pp. 436-448, 2021.  [22] A. Banerjee, S. Biswas, J. Lladós, and U. Pal, "SemiDocSeg: harnessing semi-supervised learning for document layout analysis," *International Journal on Document Analysis and Recognition (IJDAR),* pp. 1-18, 2024.  [23] M. M. B. A. N. Akanda, M. Ahmed, A. S. A. Rabby, and F. Rahman, "Optimum Deep Learning Method for Document Layout Analysis in Low Resource Languages," in *Proceedings of the 2024 ACM Southeast Conference*, 2024, pp. 199-204.  [24] L. Aljiffry, H. Al-Barhamtoshy, A. Jamal, and F. Abukhodair, "Arabic Documents Layout Analysis (ADLA) using Fine-tuned Faster RCN," in *2022 20th International Conference on Language Engineering (ESOLEC)*, 2022, vol. 20: IEEE, pp. 66-71.  [25] A. Gemelli, E. Vivoli, and S. Marinai, "Graph neural networks and representation embedding for table extraction in PDF documents," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022: IEEE, pp. 1719-1726.  [26] Y. Huang *et al.*, "A YOLO-based table detection method," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019: IEEE, pp. 813-818.  [27] X.-H. Li, F. Yin, and C.-L. Liu, "Page segmentation using convolutional neural network and graphical model," in *Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Proceedings 14*, 2020: Springer, pp. 231-245.  [28] D. Liu and S. Zhou, "Pixel-Level Segmentation of Handwritten and Printed Texts in Document Images with Deep Learning," in *2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2023: IEEE, pp. 1-5.  [29] L. Ma, C. Long, L. Duan, X. Zhang, Y. Li, and Q. Zhao, "Segmentation and recognition for historical Tibetan document images," *IEEE Access,* vol. 8, pp. 52641-52651, 2020.  [30] M. Minouei, M. R. Soheili, and D. Stricker, "Document layout analysis with an enhanced object detector," in *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 2021: IEEE, pp. 1-5.  [31] S. Umer, R. Mondal, H. M. Pandey, and R. K. Rout, "Deep features based convolutional neural network model for text and non-text region segmentation from document images," *Applied Soft Computing,* vol. 113, p. 107917, 2021.  [32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015: Springer, pp. 234-241.  [33] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*, 2022: Springer, pp. 205-218.  [34] H. Dong, G. Yang, F. Liu, Y. Mo, and Y. Guo, "Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks," in *Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings 21*, 2017: Springer, pp. 506-517.  [35] M. Bardis *et al.*, "Deep learning with limited data: organ segmentation performance by U-Net," *Electronics,* vol. 9, no. 8, p. 1199, 2020.  [36] S. Nabiee, M. Harding, J. Hersh, and N. Bagherzadeh, "Hybrid U-Net: Semantic segmentation of high-resolution satellite images to detect war destruction," *Machine Learning with Applications,* vol. 9, p. 100381, 2022.  [37] T.-A. N. Dang and D.-T. Nguyen, "End-to-end information extraction by character-level embedding and multi-stage attentional u-net," *arXiv preprint arXiv:2106.00952,* 2021.  [38] J. Lee, H. Hayashi, W. Ohyama, and S. Uchida, "Page segmentation using a convolutional neural network with trainable co-occurrence features," in *2019 International conference on document analysis and recognition (ICDAR)*, 2019: IEEE, pp. 1023-1028.  [39] K. Ma, Z. Shu, X. Bai, J. Wang, and D. Samaras, "Docunet: Document image unwarping via a stacked u-net," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4700-4709.  [40] A. Chaudhuri *et al.*, *Optical character recognition systems*. Springer, 2017. |

|  |
| --- |
| [41] R. Smith, "An overview of the Tesseract OCR engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, 2007, vol. 2: IEEE, pp. 629-633.  [42] I. Anuradha, C. Liyanage, H. Wijayawardhana, and R. Weerasinghe, "Deep learning based sinhala optical character recognition (ocr)," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2020: IEEE, pp. 298-299.  [43] H. W. H. Premachandra, C. Premachandra, T. Kimura, and H. Kawanaka, "Artificial neural network based sinhala character recognition," in *Computer Vision and Graphics: International Conference, ICCVG 2016, Warsaw, Poland, September 19-21, 2016, Proceedings 8*, 2016: Springer, pp. 594-603.  [44] K. Thamarasee and R. Surendra, "Sinhala Character Identification Using Orientation and Support Vector Machine," in *2024 4th International Conference on Advanced Research in Computing (ICARC)*, 2024: IEEE, pp. 127-131.  [45] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. Staar, "Doclaynet: A large human-annotated dataset for document-layout segmentation," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 3743-3751.  [46] X. Zhong, J. Tang, and A. J. Yepes, "Publaynet: largest dataset ever for document layout analysis," in *2019 International conference on document analysis and recognition (ICDAR)*, 2019: IEEE, pp. 1015-1022. |

|  |
| --- |
| Any Other Relevant Information |
|  |