

An Enhanced U-Net-based Approach for Sinhala Document Layout Analysis

Hulathdoowage S.K.D.^{1*}, Kumara B.T.G.S.²

¹ Department of Computing & Information Systems, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Sri Lanka

² Department of Data Science, Faculty of Computing, Sabaragamuwa University of Sri Lanka, Sri Lanka

* skdilukshi@gmail.com

<https://doi.org/10.56427/jcbd.v4i3.767>

ARTICLE INFO

Article History

Received: May 29, 2025

Revised: August 27, 2025

Accepted: September 25, 2025

Keywords

Document Layout Analysis,
Sinhala Documents,
U-Net, Vision Transformer,
Semantic Segmentation

ABSTRACT

Document layout analysis plays a critical role in the digitization pipeline by identifying, segmenting, and classifying structural elements within documents to support accurate information extraction. This task becomes increasingly challenging when dealing with heterogeneous layouts that contain paragraphs, tables, figures, mathematical expressions, and other visual components. For Sinhala, a low-resource language with limited annotated datasets and specialized models, research in this area remains sparse. To address this gap, this study proposes an enhanced U-Net architecture that integrates convolutional neural networks with vision transformer blocks to improve semantic segmentation performance. The model leverages convolutional layers to capture fine-grained local features while employing transformer components to model long-range dependencies and global contextual relationships across document regions. A manually annotated dataset of 750 Sinhala document images covering 14 distinct element categories was developed to train and evaluate the model. Experimental results demonstrate that the proposed architecture significantly outperforms standard U-Net and attention U-Net variants, achieving 93.06% pixel accuracy, 64.37% mean IoU, and 77.32% mean F1-score. This research represents the first comprehensive document layout analysis framework tailored specifically for Sinhala documents and provides a strong foundation for future digitization, archival, and text processing initiatives within Sri Lankan academic, governmental, and cultural institutions.



JCBD is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).

1. Introduction

Document Layout Analysis (DLA) is the process of identifying, segmenting, and analyzing various elements within a document, such as paragraphs, images, tables, lists, charts, and diagrams [1]. Converting document images into editable digital formats while maintaining their layouts is a critical step in the digitization process. In industries like government, healthcare, legal services, finance, and education, effective DLA facilitates easier data extraction, better information retrieval, and increased accessibility to vital information [2][3][4].

DLA has been extensively researched for high-resource languages such as English, French, and Chinese, with approaches ranging from traditional rule-based methods to modern deep learning architectures [5][6][7][8]. Recent advances include CNN-based approaches like Faster R-CNN for region detection, transformer-based models for document understanding, and hybrid architectures combining multiple neural network paradigms. However, low-resource languages remain significantly underrepresented in this field. Sinhala, the primary language of Sri Lanka spoken by approximately 87% of the country's population, represents a critical gap in DLA research. As the dominant language in school education, government documentation, and official services in Sri Lanka, developing effective DLA solutions for Sinhala documents is essential for advancing digital accessibility and preservation of cultural heritage [9].

Current state-of-the-art DLA approaches predominantly utilize either CNN-based architectures (U-Net, Mask R-CNN) [8][10], transformer-based models (LayoutLM, DocFormer) [11][12], or hybrid methods

combining multiple techniques [13][14][15][16][17][18][19][20]. While U-Net architectures have proven effective for semantic segmentation tasks, they face limitations in capturing long-range dependencies due to their reliance on local convolutional operations. Transformer-based approaches excel at modeling global context but may struggle with fine-grained spatial details. This presents an opportunity to develop hybrid architectures that leverage the complementary strengths of both approaches. The U-Net architecture, first introduced for biomedical image segmentation [21], is a highly effective approach compared to other image processing methods [22].

This research provides an approach for Sinhala document layout analysis using a Vision Transformer (ViT) [23] integrated U-Net architecture. The proposed method addresses current limitations by combining CNN-based local feature extraction with transformer-based global context modeling. This hybrid approach enhances segmentation performance through improved feature fusion via transformer-refined encoder-decoder connections, robust handling of scale variations through multi-scale feature understanding, and superior capture of long-range dependencies essential for complex document layouts.

The study guided by following research questions:

- RQ1: How can a comprehensive Sinhala document dataset with diverse layouts be collected and annotated?
- RQ2: What architectural improvements are required in U-Net for analyzing complex multi-element documents?
- RQ3: How can hybrid CNN-transformer architectures achieve accurate layout analysis for complex document images?

2. Research Methodology

2.1 Data Collection and Annotation

In the context of Sinhala document layout analysis, the unavailability of a collected and annotated dataset presents a significant challenge. To address this gap, the primary objective was to collect a comprehensive dataset including a variety of Sinhala document images, making this the first dataset specifically designed for Sinhala document layout analysis. Prior to the data collection step, a set of inclusion criteria was developed to ensure dataset quality, adaptability, and relevance. The research considered books published from 1950 onward, avoiding older publications due to their different layouts, font styles, and linguistic structures. Additionally, to create a comprehensive dataset with diverse elements, different categories of books were included such as religious books, government publications, school textbooks, and literary works, etc.

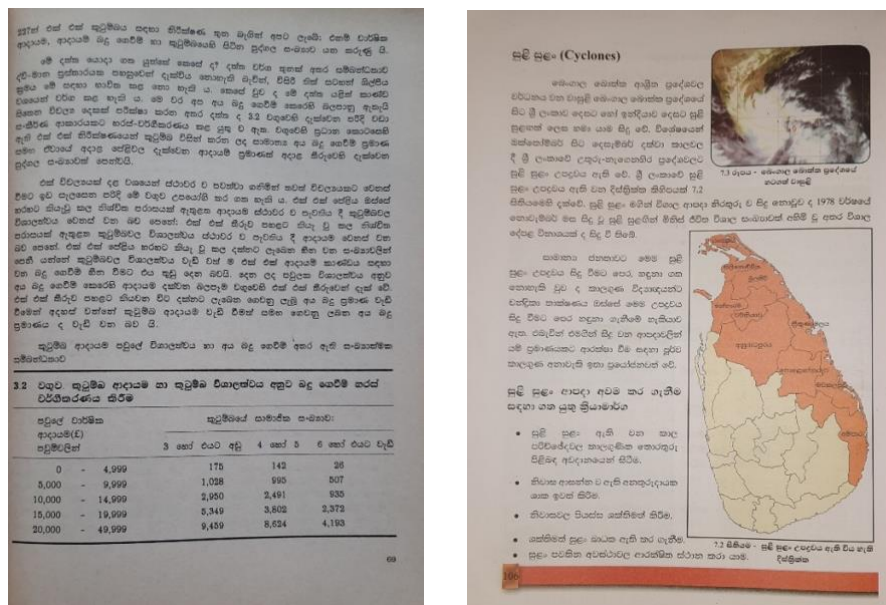


Figure 1. Sinhala Document Image Dataset Samples

To enhance the dataset's comprehensiveness, the collected document images were ensured to contain a diverse range of elements. These included Paragraphs, Topic – Level1, Topic – Level2, Topic – Level3, Image, Diagram, Table, Mathematical Expression, List, Header, Footer, Caption, Poem, and Charts. A total of 750 Sinhala document images were captured using five different smartphone models: Samsung Galaxy M21, Apple

iPhone 12 Pro Max, Huawei Y9 Prime, Vivo S1, and Vivo V9. Copyright and ethical considerations were strictly followed by capturing only a limited number of pages per book to prevent intellectual property violation issues. Images were manually annotated using Label Studio for semantic segmentation by drawing masks for each label .

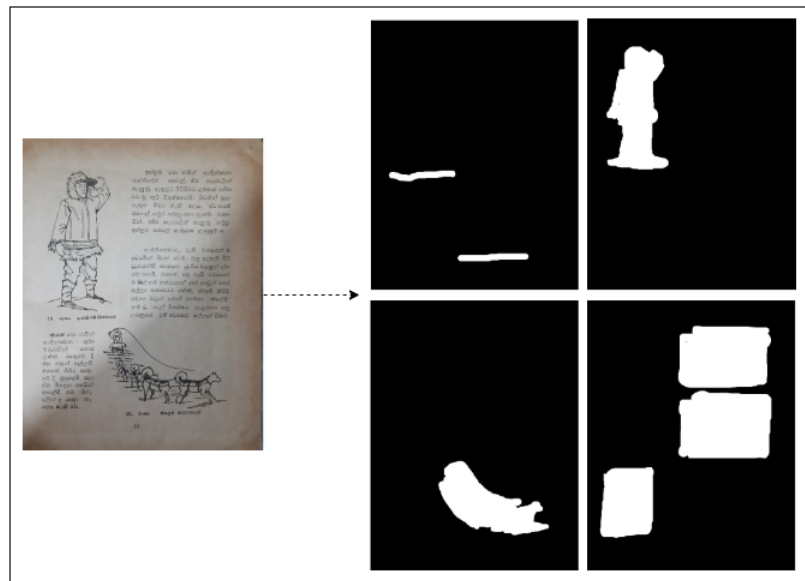


Figure 2. Annotation Output Sample from Label Studio

2.2 Create Segmentation Mask

To generate segmentation masks, an empty mask matching the dimensions of the document image is initialized with all pixels set to the background class. Annotation filenames are parsed to extract document IDs, linking each label to the corresponding image. Label masks, representing layout classes, are loaded in grayscale and resized to match the document dimensions. A binary threshold ensures clean separation between foreground and background elements. Foreground pixels in the label mask are assigned a class index in the segmentation mask. Finally, the mask is converted into tensor format suitable for deep learning model training.

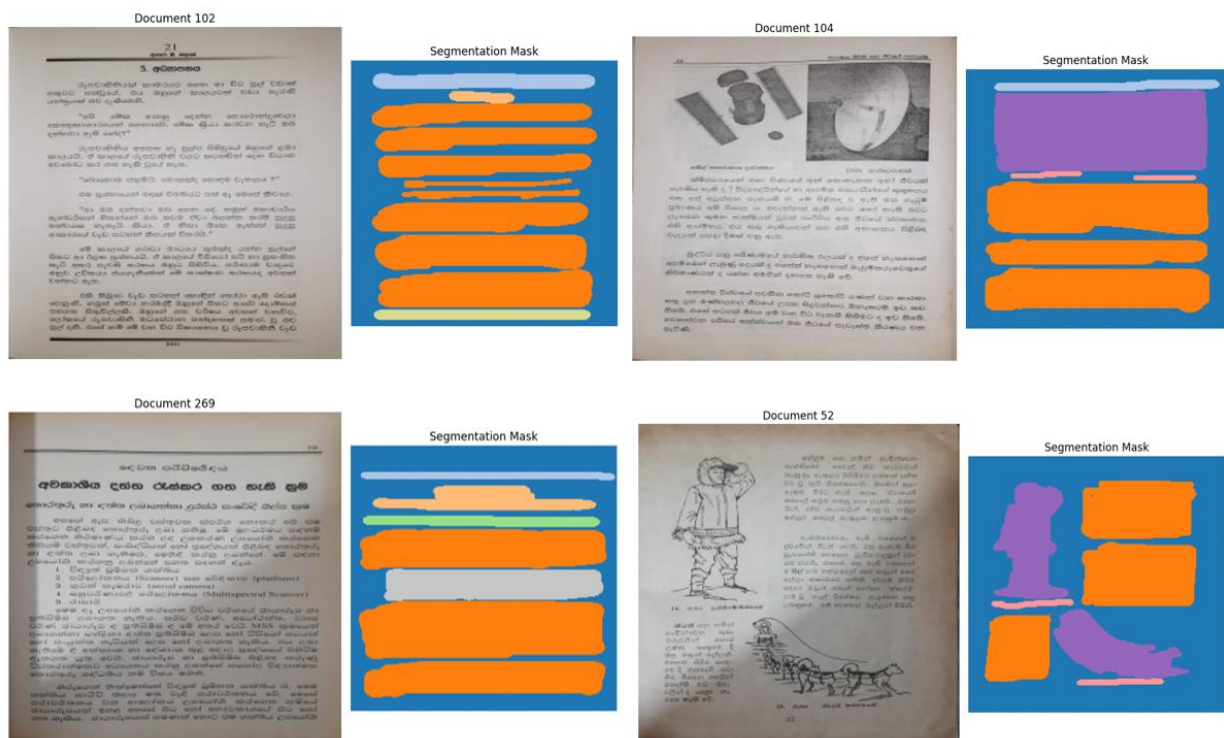


Figure 3. Segmentation Masks with Raw Images

2.3 Enhanced U-Net Model Building

The proposed architecture enhances the traditional U-Net by incorporating Vision Transformer (ViT) components, which use attention mechanisms to understand relationships between different parts of an image simultaneously, rather than processing them sequentially like traditional methods.

- Encoder:** The encoder integrates convolutional layers with Vision Transformer blocks, enabling both local detail extraction and global context understanding. Unlike traditional U-Net encoders that rely solely on convolutional operations, this approach uses attention mechanisms to enhance feature representation across different spatial scales.
- Feature extraction:** The combination of convolutional layers and attention mechanisms allows efficient multi-scale feature extraction. Convolutions capture fine-grained local details, while ViT blocks preserve long-range relationships, leading to more robust representations.
- Skip connections:** Instead of direct feature combination as in standard U-Net, this research employs Transformer-refined skip connections. Encoder features are first processed through ViT blocks before being merged with the decoder, ensuring better feature integration and reducing information loss.
- Downsampling:** This approach incorporates MaxPooling with ViT-enhanced feature compression, where traditional pooling layers are followed by attention-based refinement. This enables the model to maintain global context even at lower resolutions.
- Bottleneck:** A deeper Vision Transformer module at the bottleneck stage enhances the ability to aggregate global contextual information through multi-head attention mechanisms, enabling better scale-invariant feature encoding.
- Upsampling:** The upsampling strategy combines bilinear interpolation with ViT-based feature refinement. The decoder restores spatial resolution while reprocessing features through Transformer blocks, leading to more precise boundary reconstruction.

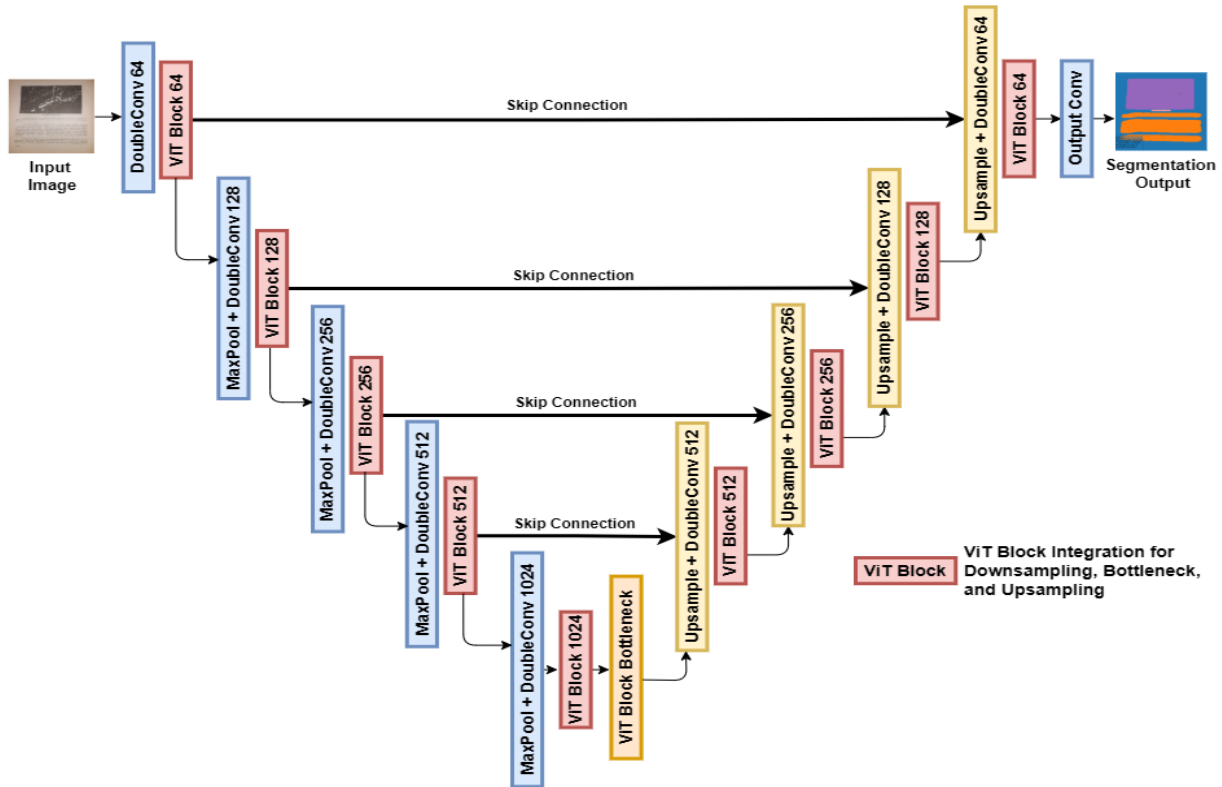


Figure 4. Proposed U-Net Architecture

Within the input image $X \in \mathbb{R}^{H \times W \times C}$, where H and W represent height and width, and C represents the number of channels. Then, the encoding process can be defined as:

$$F_l = \text{ReLU}(\text{BN}(\text{Conv}_l(F_{l-1})))$$

Unlike conventional decoders that rely solely on convolutions, introduced transformer-based refinement, allowing the network to enhance long-range feature dependencies and improve segmentation accuracy, a 1x1 convolution layer generates the segmentation mask.

$$Y = \text{Softmax}(\text{Conv}_{1 \times 1}(O_L))$$

In this research, the Cross-Entropy Loss function was used to train the document segmentation model. Cross-Entropy Loss is a widely used loss function for classification tasks, particularly in multi-class segmentation problems.

$$L = - \sum_{i=1}^N y_i \log(\hat{y}_i)$$

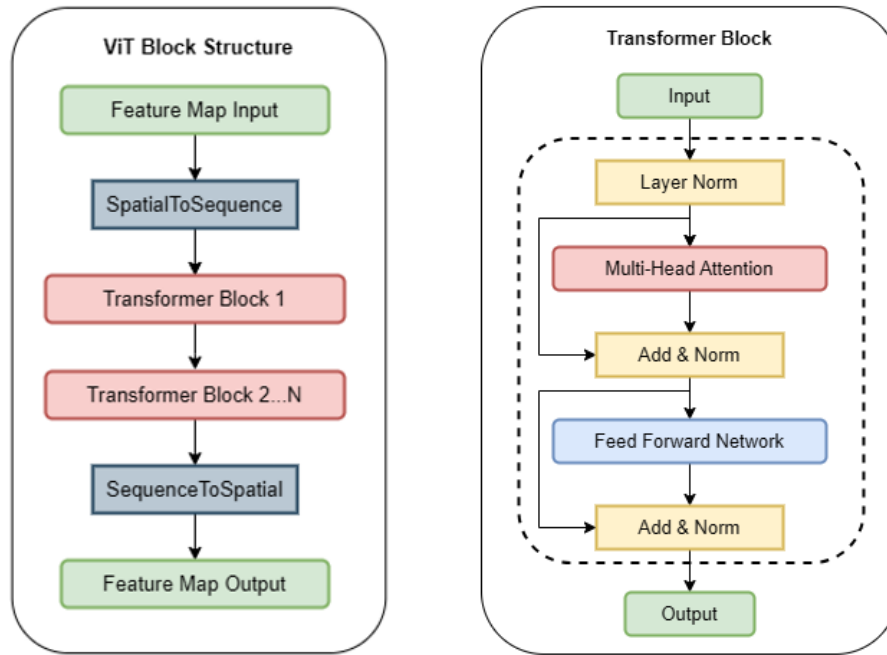


Figure 5. ViT Block Structure and Transformer Block Structure

The ViT block is the key component that significantly enhances feature representation at various levels of the Enhanced U-Net. Unlike conventional approaches that process image regions independently, the ViT block uses attention mechanisms to understand relationships between distant regions simultaneously. This allows the model to effectively capture dependencies across the entire document layout, making it highly efficient for complex segmentation tasks. At the bottleneck of the U-Net, a deep transformer block enhances high-level feature representations before decoding. This component processes and refines the most abstract features extracted from the input document by operating on sequences of feature patches, allowing it to model complex global dependencies across the entire document.

2.4 Training and Validation

The training process incorporates effective learning rate adjustments, model checkpointing, and performance tracking. Cross-Entropy Loss is employed as it is well-suited for multi-class segmentation tasks. The Adam optimizer is selected with an initial learning rate of 0.001 for stable learning convergence. A learning rate scheduler (ReduceLROnPlateau) enhances learning efficiency by monitoring validation loss and reducing the learning rate when improvements plateau. The patience parameter is set to five epochs, with the learning rate halved when needed. During training, mini-batches of image-mask pairs are processed through the model to generate predictions, followed by loss calculation and weight updates via backpropagation. To prevent overfitting, model checkpoints are saved, with the best-performing model determined by lowest validation loss retained. Additional checkpoints are saved every five epochs for recovery and analysis purposes.

The dataset is divided into training, validation, and test sets using an 80:10:10 ratio to ensure independent evaluation. The training set (600 images) is used for model optimization, validation set (75 images) for hyperparameter tuning and early stopping, and test set (75 images) for final performance evaluation. All images are resized to 512×512 pixels for consistency.

Experiments were conducted on NVIDIA RTX 3080 GPU with 10GB VRAM, 32GB RAM, and Intel Core i7-10700K processor. Training was performed using PyTorch framework with CUDA acceleration. Each model was trained for a maximum of 100 epochs with early stopping based on validation loss. Average training time was approximately 6-8 hours per model. Data loading utilized PyTorch's DataLoader with batch size of 4 to balance memory usage and training efficiency.

3. Results and Discussion

3.1 Results Comparison

To evaluate the proposed enhanced U-Net model for Sinhala document layout analysis, standard evaluation metrics including Pixel Accuracy, IoU, Precision, Recall, and F1-score were employed. Standard U-Net and Attention U-Net served as baseline models for comparison.

Table 1. Overall Results Comparison

	Pixel Accuracy	Mean Accuracy	Mean IoU	Mean Precision	Mean Recall	Mean F1-Score
Proposed U-Net	93.06	95.17	64.37	80.68	74.24	77.32
Attention U-Net	84.11	89.50	54.71	78.91	67.25	70.86
Standard U-Net	81.26	86.68	53.89	70.73	61.72	65.57

The proposed model achieves 93.06% pixel accuracy with 77.32% F1-score, substantially outperforming Attention U-Net (70.86% F1-score) and Standard U-Net (65.57% F1-score). The mean IoU of 64.37% shows improvements of 9.66% and 10.48% over baseline models, though remains lower than other metrics due to IoU's requirement for exact boundary matching. This indicates better segmentation accuracy through the combined CNN and transformer approach. Tables (89.0% F1-score) and headers (86.7% F1-score) show strong results due to their clear, structured layouts that are easier to identify and segment. Diagrams (48.4% IoU) and mathematical expressions (54.3% IoU) show lower performance. Diagrams are difficult because of their varied shapes and complex internal details. Mathematical expressions have high precision (98.1%) but low recall (51.0%), meaning the model correctly identifies them when found but often misses smaller math elements mixed within text. Topic-level elements (Level 2: 52.6%, Level 3: 52.4% IoU) struggle because they look similar to regular paragraphs, making it hard to distinguish between them.

Table 2. Element Wise Results Comparison

	Proposed U-net				Attention U-net				Standard U-net			
	IoU	Pre.	Re.	F1	IoU	Pre.	Re.	F1	IoU	Pre.	Re.	F1
Header	76.6	91.1	82.8	86.7	68.9	87.3	76.1	81.3	55.2	73.8	64.2	68.6
Paragraph	68.8	71.4	95.0	81.5	60.5	62.3	89.8	74.9	58.3	62.8	87.5	72.6
TopicL1	70.6	88.5	77.7	82.8	62.5	81.8	69.5	75.1	59.8	81.2	68.0	74.5
TopicL2	52.6	72.2	65.0	68.4	48.8	69.0	61.8	62.7	46.5	55.5	59.0	57.2
TopicL3	52.4	98.1	53.2	68.8	48.0	93.7	49.1	64.5	45.9	70.4	47.0	62.3
Caption	68.2	91.2	73.0	81.1	60.2	86.4	65.2	74.6	57.8	73.5	43.5	54.6
Image	68.7	75.8	88.0	81.4	58.3	70.8	77.5	74.1	56.0	68.5	75.4	71.3
Diagram	48.4	65.9	64.6	65.3	45.8	65.1	64.0	63.6	43.2	61.5	51.3	60.1

Table	80.1	86.1	92.0	89.0	64.2	82.0	86.5	84.3	71.3	79.3	84.2	81.1
Math. Ex.	54.3	98.1	51.0	67.5	50.3	89.8	47.1	63.2	48.0	87.1	45.9	61.5
List	50.6	56.5	82.8	67.2	42.3	50.0	74.5	59.9	40.5	47.5	71.3	57.8
Footer	72.1	92.5	76.5	83.8	69.1	88.5	74.1	80.7	66.3	85.7	71.2	78.1
Poem	58.1	85.3	64.6	73.5	51.2	74.4	48.3	58.6	49.1	76.5	56.3	65.8
Charts	53.6	94.5	55.3	69.8	49.8	93.9	50.4	66.5	47.0	91.0	48.3	63.7

The sample results obtained from the proposed model illustrate different stages of the segmentation process, including the original document, the model predictions, and the corresponding error analysis. In this analysis, the original document layouts are compared with their ground truth segmentation masks, in which different regions such as text blocks and images are highlighted with different colors. The model's predictions indicate how effectively the network segments various components of the document, while the error analysis visualizes misclassified pixels by showing areas where the model's predictions deviate from the ground truth. The error patterns show that most mistakes occur at the boundaries between different parts of the document. However, the results demonstrate accurate segmentation for the majority of document regions, with clear identification of major elements such as paragraphs, images, and tables. Smaller elements exhibit more errors than larger ones, but the model successfully captures most of the document structure. The model occasionally confuses similar elements, particularly mislabeling Topic Level 2 elements as paragraphs. The transformer component helps the model understand the overall document structure but can sometimes make boundaries less precise when elements are visually similar. This reflects a trade-off between capturing the global context of the document and achieving precise boundary segmentation, although the overall segmentation quality remains high across diverse Sinhala document types. These findings demonstrate clear improvements in Sinhala document layout analysis while indicating specific areas that require further refinement.

3.2 Key Contributions

The key contributions of this study can be summarized as follows, highlighting the novelty and significance of the proposed work:

- This research introduces the first annotated Sinhala document image dataset.
- This work establishes the first document layout analysis approach for Sinhala language.
- This study presents an enhanced U-Net architecture integrating Vision Transformer components at multiple stages.
- The proposed model achieves superior segmentation performance with 93.06% pixel accuracy and 77.32% F1-score.
- The architecture demonstrates robust generalization capability across diverse Sinhala document structures.
- This research introduces hierarchical topic classification and novel element categories including poems, charts, diagrams, and images as individual classes.
- This research presents the first multiclass document layout analysis using U-Net-based architecture.

3.3 Limitations

One of the primary limitations encountered in this research was the size of the dataset. The study was conducted using a dataset comprising 750 Sinhala document images. Furthermore, the dataset did not include Sinhala newspapers or handwritten documents.

3.4 Future Works

Build a larger, more diverse Sinhala dataset to improve training and evaluation. Test the Enhanced U-Net on Sinhala newspapers to handle complex layouts. Extend evaluation to English and other low-resource languages to assess generalizability.

4. Conclusion

The research presents the first comprehensive document layout analysis solution for Sinhala documents through a U-Net architecture enhanced with Vision Transformer components. The proposed model demonstrates substantial performance improvements, achieving 93.06% pixel accuracy and 77.32% F1-score,

and consistently outperforming standard U-Net and Attention U-Net baselines across 14 document element categories. The study's key contributions include the development of the first annotated Sinhala document dataset, the introduction of hierarchical topic classification, and the effective integration of convolutional and transformer-based mechanisms to improve global context modeling. By addressing crucial challenges in low-resource document processing, this work establishes a strong foundation for multilingual document digitization and contributes directly to academic, governmental, and cultural preservation efforts within Sri Lankan communities.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] L. Aljiffry, H. Al-Barhamtoshy, A. Jamal, and F. Abukhodair, "Arabic Documents Layout Analysis (ADLA) using Fine-tuned Faster RCNN," in *2022 20th International Conference on Language Engineering (ESOLEC)*, 2022, vol. 20: IEEE, pp. 66-71.
- [6] 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.
- [7] T. Shehzadi, D. Stricker, and M. Z. Afzal, "A hybrid approach for document layout analysis in document images," *arXiv preprint arXiv:2404.17888*, 2024.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [12] J. Wang, K. Hu, and Q. Huo, "DLAFormer: An End-to-End Transformer For Document Layout Analysis," in *International Conference on Document Analysis and Recognition*, 2024: Springer, pp. 40-57.
- [13] 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.
- [14] 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>.
- [15] Z. Gao and S. Li, "YOLOLayout: Multi-Scale Cross Fusion Former for Document Layout Analysis," *International Journal of Emerging Technologies and Advanced Applications*, vol. 1, no. 2, pp. 8-15, 2024.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] H. M. Zangana, Z. B. Sallow, and B. A. Salih, "The Impact of Artificial Intelligence on Healthcare: A Systematic Review of Innovations, Challenges, and Ethical Considerations," *Journal of Computers and Digital Business*, vol. 4, no. 1, pp. 1-9, 2025.
- [23] K. Han *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87-110, 2022.