**A Novel Approach for Newspaper Article Segmentation using Recurrent Convolutional Neural Network**

Shridevi SOMAa,[[1]](#footnote-1) and Shilpa b

a Computer Science and Engineering Department, PDACE,Klb,Krnataka,India

b Computer Science and Engineering Department, SUK, Klb,Karnataka,India

a shridevisoma@gmail.com , bs123shilpa@gmail.com

**Abstract.** On exploring the daily newspapers, they have information about various topics, local, national and international affairs, advertisements, etc. Every information found in the articles become important with different aspect of readers. As there are multiple news articles, segmenting and classifying whether it’s an article or not is a real challenge. Our novel approach helps in segmenting all news blocks and classify them into three major classes, articles, non-articles and titles using one of the most reliable deep learning algorithms, Mask-RCNN (Recurrent Convolutional Neural Network). This approach has been tested on dataset consisting of digital e-newspaper images of recent years from the TOI and Financial Express Newspapers with different layouts and complexities.

**Keywords.** Document understanding, Newspaper Segmentation, Article Classification, Deep Learning, Neural Networks.

# Introduction

Digital analysis of documents is challenging in most of the aspects and useful in understanding various design, patterns, layouts and structures of the document. In newspapers, articles are the building blocks and plays a vital role in representing any news. Apart from articles, other categories such as advertisements, weather reports, etc are included within the newspapers which completes daily news. These blocks are identified in the newspapers in a specific format according to the news media company. Different kinds of layout, grids and alignments are followed and types of attributes such as texts, graphics, shapes, etc are also included which brings a challenge to analyse their structure in a unified way. Hence there are many contributions made in newspaper layout analysis, article segmentation, block classification, etc in the recent years by applying the principles of digital image processing and artificial intelligence collectively.

In this work, we extensively focus on article segmentation within the digital newspapers along with multi-purpose results or outputs that can be utilised for further research in digital newspapers. The dataset we used for training and validation is developed by us with reference to multiple other existing datasets. There are very few datasets are found namely prima research’s Layout Analysis Dataset [15], RDCL2019 (Recognition of Documents with Complex Layouts) dataset [16] when it comes to newspaper digitalization or document layout analysis. No specific datasets are generated specifically for newspaper article semantic segmentation. Our dataset is a pre-liminary dataset where multiple improvements can be made in the future to bring the scope of segmenting articles whose layout are complex enough across the global present newspapers. Along with the labels that we created are simple enough to classify the articles such that our goal shall be satisfied.

Section III explains proposed methodology and gives details about each module of proposed approach. Section IV describes experimental results & discussions respectively. Section V provides the inferences of work and ensuing orientations.

# Related Work

Image segmentation problems in different kinds of documents have been studied and analyzed. Benjamin Meier et al. (2017) [1] proposed a newspaper article segmentation using Fully Convolutional Neural Networks. Here the input provided to the network are original image from dataset and its OCR processed image in the fixed shape of 256x256. They have used the dataset of 4135 images from ARGUS DATA INSIGHTS Schweiz AG and achieved Diarization Error Rate (DER) of 46% improvement over the average segmentation results. Hui-Yin Wu et al. (2019) [2] proposed a methodology for multilayered analysis of newspaper design. They provide a system of three different stages namely segmentation of newspaper using curvilinear delimiters, classification of segmented blocks to its corresponding labels, analysis of the above steps output is grouped into meaningful structure and design of the newspaper. Their method performs segmentation with 70% accuracy with test data using standard accuracy formula.

Sofia Ares Oliveira et al. (2019) [3] propose a method of pixel-wise predictor for document layout, pages, texts, images etc using Convolutional Neural Network (CNN). They utilized READ-BAD dataset for document layout segmentation of historical manuscripts. Anukriti Bansal et al. (2014) [4] proposes a technique where features for articles are extracted manually using certain rules for text and graphics segmentation. These blocks segmented are used as inputs to Support Vector Machine (SVM) algorithm to perform class label prediction of article/non-article. Roberto Beretta et al. (2011) [5] proposes a methodology to identify articles in the newspaper by segmenting each block in the newspaper and link them using a graph structure. These links are grouped as clusters and defined for articles. Yue Xu et al. (2017) [6] presented a fully convolutional network (FCN) to segment and identify blocks within handwritten documents. They utilize VGG 16-layer net for some FCN layers and segments the page into backgrounds and handwritten texts with an accuracy of 99.8 % for main text body prediction.

Liangcai Gao et al. (2012) [7] proposes a method to reconstruct articles using the graph frameworks. In spite of various layouts, articles are identified and reconstructed which brings the better understanding of article layouts. T. M. Breuel (2017) [11] proposes a implementation of convolutional neural networks and multi-dimensional LSTM for document layout analysis using text detection, column detection, text/image segmentation. Wonmin Byeon (2015) [13] addresses the problem of labelling various objects in the image as scene-labelling approach and achieved a decent accuracy score in both below mentioned datasets. They utilize LSTM networks as base architecture and trained with SIFT flow dataset and Stanford Background dataset with multi-labelled classes. J. Long (2016) [14] is used to understand how fully convolutional neural nets are implemented in order to segment objects using pixel-wise feature extraction.

The above-mentioned surveys determine the various deep learning approaches and graph-based approaches for document page segmentation and article classification. Specifically for newspaper, our approach investigates the purpose of segmentation needed and overriding the existing deep learning architecture with most reliable architecture.

# Proposed Methodology

The major methodology involved here covered in three major steps, dataset preparation, preprocessing, training and validation. In dataset preparation, collected e-newspapers are prepared for input annotations and class labelling. In preprocessing, process such as train-test split, masks and label loading, model configuring, etc. Further steps cover the training the dataset with the specified network architecture and validating them using general validation approaches.

## Dataset Preparation

Digital newspaper images of recent years are collected from <https://archive.org/details/TOIDELFEB18>, <https://archive.org/details/TOIDELJAN18>, and [https://dailyepaper.in/financial-express-newspaper](https://dailyepaper.in/finanmcial-express-newspaper). These collected images are analyzed for grouping the blocks into articles and other labels. After analysis, annotations of the segmented regions are made from the images using the open-source web tool called makesense.ai. It helps in annotating the regions, label them accordingly and export them into text files collective of all dataset images. An example representation of article segmentation using the tool with correctly noted annotations data as a ground-truth annotation,



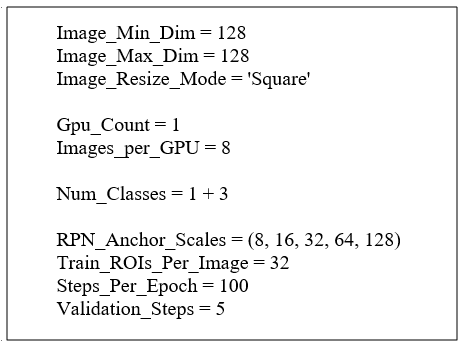
**Figure 1.** A sample annotation of articles during dataset preparation, Info: Image ID = TOI\_001, Image shape = (128, 128, 3), Mask shape = (128, 128, 14)

These annotation text data contain information related to the dataset images such as file name, file size, segmented region’s type, coordinate points, labels info, etc. following the format of VGG Image Annotator (VIA) json. Majorly rectangular and polygon shaped regions are found in annotations and these coordinate points are used as mask inputs to the neural network along with its respective images.

## 3.2. Preprocessing

After the dataset preparation, images are split into train, validation and test datasets for modelling as 60%, 30%, 10% of total dataset images respectively. Provided a folder for each with list of images and an annotation json file. With the help of json file, data such as image data, masks, bounding boxes (bboxes), class labels of each image will be loaded. Next step is model configuration which will be carried out as a major step by setting up some of the model specific parameters or hyper parameters defined as Num\_Classes, Train\_ROIs\_Per\_Image, Steps\_Per\_Epoch, Image\_Resize\_Mode, Images\_per\_GPU, Image\_Min\_Dim, Image\_Max\_Dim and RPN\_Anchor\_Scales. For resizing all images in to a fixed size by setting the parameters Image\_Min\_Dim and Image\_Max\_Dim as 128 which produces images of 128x128 size. At second step during training, no of ROIs (Region of Interests) per image is chosen as 32 using parameter Train\_ROIs\_Per\_Image. Altering this parameter leads to change in training time.

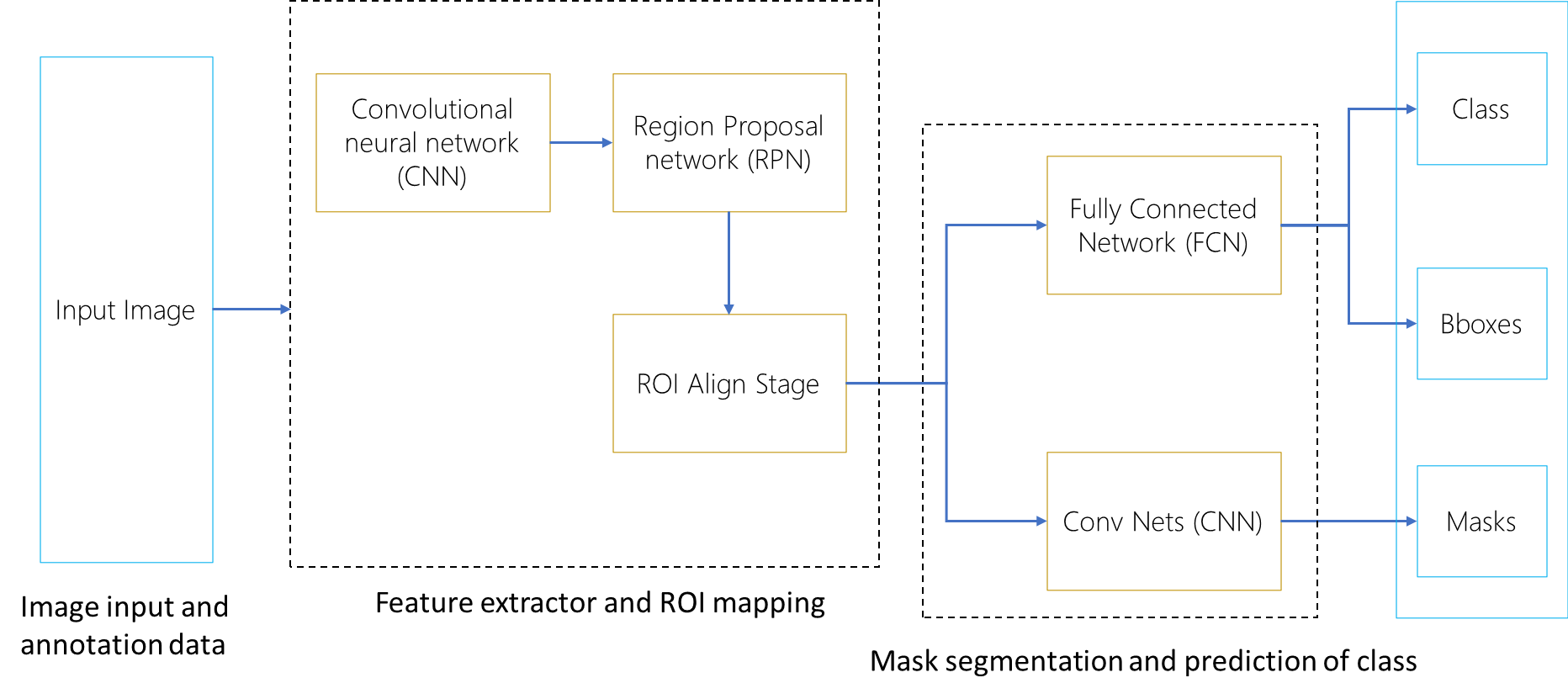
For overall network’s training, no of batch iterations for an epoch is chosen as 100 using the parameter Steps\_Per\_Epoch. Num\_Classes and Images\_per\_GPU are set as 4 and 8 respectively which defines the no of class labels (3 classes+background) and batch size for the training respectively. Similarly, there are multiple other hyperparameters which are kept in its default value to maintain the overall performance. Below figure (Figure 1.) represents the collective parameter changes for model configuration.



**Figure 2.** Model Configuration of Hyper Parameters.

## 3.3. Model Architecture

Our approach uses a network architecture based on state-of-the-art framework called as Mask-Recurrent Convolutional Neural Network (Mask-RCNN) which is preferably used for image segmentation along with region-based mask classification and it is utilized in order to segment and classify the different news blocks. Mask-RCNN follows the network structure of basic Convolutional Neural Network (CNN) architectures added with Region Proposal Network (RPN), Fully Connected Network (FCN) and another CNN at the end. Majorly CNN and RPN plays a huge role in feature extraction, region-based classification respectively. For determining the class labels and bounding box estimation, fully connected layers found in FCN are used and corresponding masks are segmented using the final CNN blocks. Initially RPN is used along with CNN for Image Classification which is said to be R-CNN and further changes in the architecture brings out other frameworks such as Fast R-CNN and Faster R-CNN. The idea of pixel-level segmentation of specific regions of interests (ROIs) using the CNN is introduced in Mask-RCNN. The overview of the network architecture is shown in the figure (Figure 2).



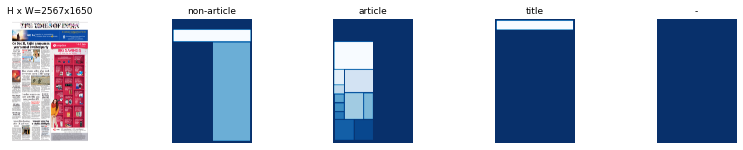
**Figure 3.** Overall architecture workflow.

Firstly, the architecture for CNN must be chosen and here ResNet-101 is used as network backbone for effective learning and manageable training timing. ResNet refers to residual network which contains 101 deep layers. As this is the first step, it helps in extracting features for the image by passing into successive convolutional networks. These features are in the image pixel level and they are utilized in the next step of RPN to segregate the objects in the image using the sliding window method. A fixed size of search slide boxes or anchors is provided according to our average object or mask size and the extracted features are mapped within those regions identified. During training, multiple anchors are considered for refining the ground truth object and a certain value is chosen as Intersection Over Union (IoU) to eliminate unwanted regions found among the ground truth object. These refined anchors or boxes are further considered for training and prediction for finding refined anchors will happen subsequently. To bring out the final ROIs, non-maximal suppression is applied over the multiple predicted detections of region boundaries of the object.

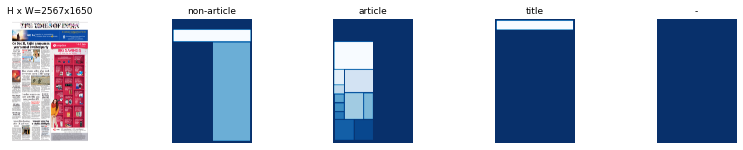
In ROI Align stage, ROIs identified for an image are aligned into fixed size decided at this stage to avoid differences in orientation which reduces the accuracy of the segmentation. Outcome of this step are considered to be final output to get trained in subsequent networks for various purposes. Using the extracted feature for the regions, the next fully connected layers learn them efficiently and classify them according to the class labels. Once the class prediction gets over, bounding boxes for the object detected will be generated, as a whole this stage provides two major outputs.

## 3.4. Training and Validation

Our custom dataset contains 100 images in training, 40 images in validation, 15 images in testing with mixing of newspaper images containing different layouts and structures. As per the format, significant data related to the dataset is found in the annotation json file under their respective folders. Hence whenever training or testing set is loaded, images are acquired with unique IDs, mask coordinate points and their labels. Below figure (Figure 3.) visualizes how the input mask applied data will be.



(a) (b) (c)



(d) (e)

**Figure 4.** Representation of one of the training images and their mask applied images, (a) Original image, (b) non-article class labelled masks, (c) article class labelled masks, (d) title labelled masks, (e) background

Above figure (Figure 3) represents the class labelled images which includes our three main classes i.e, article, non-article, title and background that are taken into account for the training phase of the model.

For model training, weights shall be initialized for the head CNN layer using the pre-trained COCO [8] weights which is more efficient for feature extraction part. These weights are loaded in the model instance and multiple configurations for tuning hyperparameters such as learning rate and no of epochs are made. Learning rate has been varied from 0.01 to 0.001 which is default and no of epochs has been varied from 5 to 40 steps. During training, these general neural network specific hyperparameters are tuned which brings changes at each step of training and their corresponding losses. As steps per epoch is selected as 100, for each epoch model will train the images 100 times throughout its architecture for the respective classes. Once after each epoch, according to the selected learning rate, weights will get adjusted and hence the losses.

Best optimal performance of the model can be found only at learning rate of 0.001 and epoch steps as 30. Training has been carried out in the system configuration of Intel i7 processor and 16 GB RAM with Windows OS. Implementation of model training is done using python language including frameworks such as TensorFlow, Keras and OpenCV. The scripts related to Mask R-CNN are used from the open-source implementation of Matterport GitHub Repository [9]. During both training and validation at each epoch, the losses such as loss, rpn\_class\_loss, mrcnn\_class\_loss are calculated at various stages are carefully analyzed as they play the vital role in determining the correct hyperparameters.

(1)

(2)

(3)

where is binary cross-entropy loss, is predicted class loss, is predicted bounding box loss, is the label of a cell (i, j), m2 is area of the region, is the predicted value of the cell, k is ground truth class.

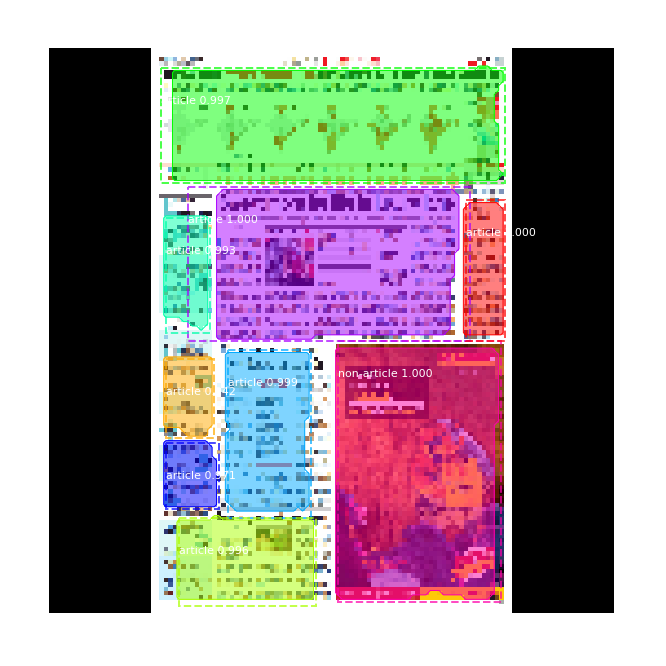
Reduction in these loss values is exponentially increasing over the epoch and training time and effective learning brings the better reduction of losses in the starting 5-10 epochs. After every epoch, to analyze or reuse the weights, model instances will be logged as .h5 file.

**4. Experimental Results & Discussions**

Experimented on a dataset consisting of digital e-newspaper images of recent years from The TOI and Financial express. Furthermore, from the results acquired, to determine how accurate the algorithm have segmented and classified the articles within the original data, certain calculations have been performed from output acquired. As a result, from the trained model some promising results are shown below in the figure.

(a) (b)



(c) (d)

**Figure 5.** Results of test image 1 overview, (a) Ground Truth image 1, (b) Predicted image 1, (c) Ground Truth image 2, (d) Predicted image 2

These results gave the understanding that our model performs well in identifying the article regions and predicting their classes with a note to improve the segmentation area overlapping. Mask overlapping can be improved from the anchor’s prediction during the RPN training. To provide better understanding over the prediction, some metrics other than model specific parameters are required. These metrics are calculated from the comparison between ground-truth classes and predicted classes.

Mean Average Precision (mAP), Precision and Recall calculations are the major factors to determine accuracy scores of our algorithm. In the Eq. (5) and (6) TP represents True Positives; FP represents False Positives and FN represents False Negatives. Additionally, other factors such as mean Average Precision (mAP) shown in Eq. (4) and overlaps of masks within an image are also considered.

(4)

(5)

(6)

where P represents Precision Value, R represents Recall Value and N represents of no of predictions.

Our architecture has achieved overall precision rate of 67.47%, overall recall rate of 44.61% and overall mAP rate of 57.18% where it performed a decent score in segmenting the articles and class labelling them.

**5. Conclusion**

This paper dealt with the problems of newspaper segmentation in finding out whether the blocks in newspaper belongs to three different classes articles, non-articles and page titles. Overall understanding of layout analysis and newspaper article segmentation has been proven experimentally with decent accuracy. Our work not only limited to training the model, but also preparing a custom dataset to bring out more challenges.

Furthermore, this can be improved in terms of specific document segmentation by defining their own class labels and masks. Our dataset can be extensively improved with multiple class labels for various research. After segmentation of article, a classifier model can be built to classify the segmented blocks to its respective topics or categories such as politics, sports, business, weather, etc, This can be a future work that can be added as an advantage to reduce misleads in article analysis.

**References**

[1] B. Meier, T. Stadelmann, J. Stampfli, M. Arnold, and M. Cieliebak, “Fully Convolutional Neural Networks for Newspaper Article Segmentation,”Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, vol.1, pp. 414–419, 2017, doi: 10.1109/ICDAR.2017.75.’

[2] H.-Y. Wu and P. Kornprobst, “Multilayered Analysis of Newspaper Structure and Design,” no. RR-92812019,[Online].Available:https://hal.inria.fr/hal-02177784.

[3] S. Ares Oliveira, B. Seguin, and F. Kaplan, “DhSegment: A generic deep-learning approach for document segmentation,” Proc. Int. Conf Front.Handwrit. Recognition, ICFHR, vol. 2018-Augus, pp.7–12, 2018, doi: 10.1109/ICFHR-2018.2018.00011.

[4] A. Bansal, S. Chaudhury, S. D. Roy, and J. B. Srivastava, “Newspaper article extraction using hierarchical fixed point model,” Proc. - 11th IAPR Int. Work. Doc. Anal. Syst. DAS 2014, pp. 257–

261, 2014, doi: 10.1109/DAS.2014.42.

[5] R. Beretta and L. Laura, 2011. Performance evaluation of algorithms for newspaper article identification. In 11th International Conference on Document Analysis and Recognition, ICDAR ’11, pages 394–398.

[6] Y. Xu, W. He, F. Yin and C. -L. Liu, "Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 541-546, doi: 10.1109/ICDAR.2017.94.

[7] L. Gao, Z. Tang, X. Lin and Y. Wang, "A graph-based method of newspaper article reconstruction," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 1566-1569.

[8] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1\_48.

[9] W. Abdulla, Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. Github, 2017. Available:https://github.com/matterport/Mask\_RCNN

[10] A. Almutairi and M. Almashan, “Instance segmentation of newspaper elements using mask R-CNN,” Proc. - 18th IEEE Int. Conf. Mach. Lear Appl. ICMLA 2019, pp. 1371–1375, 2019, doi:10.1109/ICMLA.2019.00223.

[11] T. M. Breuel, “Robust, simple page segmentation using hybrid convolutional mdlstm networks,” in Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on, vol. 1, pp. 733–740, IEEE, 2017.

[12] S. Ramesh C, Kumar Vinod V, “A Review on Instance Segmentation Using Mask R-CNN”, proc.–Int.Conf.OnSystems Energy and Environment(ICSEE)2021.

[13] Wonmin Byeon et al. “Scene labeling with lstm recurrent neural networks”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 3547–3555.

[14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440, 2015.

[15] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher, "A Realistic Dataset for Performance Evaluation of Document Layout Analysis," 2009 10th International Conference on Document Analysis and Recognition, 2009, pp. 296-300, doi: 10.1109/ICDAR.2009.271.

[16] C. Clausner, A. Antonacopoulos and S. Pletschacher, "ICDAR2019 Competition on Recognition of Documents with Complex Layouts - RDCL2019," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1521-1526, doi: 10.1109/ICDAR.2019.00245.

1. . Shridevi Soma, Computer Science and Engineering Department, PDACE, Karnataka, KLB

   E-mail: shridevisoma@gmail.com. [↑](#footnote-ref-1)