Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 7, No. 1, January 2025, pp: 154-164; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received October 5, 2024; Revised December 1, 2024; Accepted December 12, 2024; date of publication January 30, 2025
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i1.616

How to cite: Ravichandra Sriram, Siva Sathya S, and LourduMarie Sophie S, "De-identification of Protected Health Information in Clinical Document Images using Deep Learning and Pattern Matching", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 1, pp. 154-164, January 2025.

De-identification of Protected Health Information in Clinical Document Images using Deep Learning and Pattern Matching

Ravichandra Sriram, Siva Sathya S, and Lourdumarie Sophie S

Department of Computer Science, Pondicherry University, Pondicherry, India.

Corresponding author: Ravichandra Sriram (e-mail: ravichandrasriram@gmail.com).

ABSTRACT Clinical documents that include lab results, discharge summaries, and radiology reports of patients are generally used by doctors for diagnosis and treatment. However, with the popularization of AI in healthcare, clinical documents are also widely used by researchers for disease diagnosis, prediction, and developing schemes for quality healthcare delivery. Though huge volumes of clinical documents are produced in various hospitals every day, they are not shared with researchers for study purposes due to the sensitive nature of health records. Before sharing these documents, they must be de-identified, or the protected health information (PHI) should be removed for the purpose of preserving the patient's privacy. If the documents are stored digitally, this PHI can be easily identified and removed, but finding and extracting PHI from old clinical documents that are scanned and stored as images or other formats is quite a daunting task for which machine learning models have to be trained with a large number of such images. This work introduces a novel combination of deep learning and pattern matching algorithms for the efficient de-identification of scanned clinical documents, distinguishing it from previous methods, which can primarily work only on text documents and not on scanned clinical documents or images. Thus, a comprehensive de-identification technique for automatically extracting protected health information (PHI) from scanned images of clinical documents is proposed. For experimental purposes, we created a synthetic dataset of 700 clinical document images obtained from various patients across multiple hospitals. The de-identification framework comprises two phases: (1) Training of YoloV3- Document Layout Analysis (Yolo V3-DLA) which is a Deep learning model to segment the various regions in the clinical document. (2) Identifying regions containing PHI through pattern-matching techniques and deleting or anonymizing the information in those regions. The proposed method was implemented to identify regions based on content structure, facilitating the de-identification of PHI regions and achieving an F1 score of 0.97. This system can be readily adapted to accommodate any form of clinical document.

INDEX TERMS Clinical De-identification, Biomedical data sharing, Document Image Analysis, Object Detection, Pattern Matching.

I. INTRODUCTION

Patients' medical records have the medications and treatments they have previously received from various hospitals. Medical practitioners use these data in subsequent meetings to improve the patient's diagnosis and care, as well as in clinical research to gain valuable insights into the patient's illness. Large number of AI models are being trained with patient's data for automating the process of disease classification and prediction to provide better quality health care. But, before sharing a patient's clinical record with researchers for clinical studies, it must eliminate or anonymize protected health information (PHI), such as the patient's name, age, and contact number. HIPAA [1] defines eighteen distinct types of PHI that require de-identification before sharing with third parties in the United States, due to the sensitive nature of patient information.

In several countries, healthcare professionals extensively utilize electronic health records (EHR) to store and organize patient information. Many electronic health records (EHRs), built by various companies, are incompatible, and healthcare facilities lack a standard protocol to control EHR interoperability. Consequently, patients are required to possess physical copies of their clinical reports on visits to new medical facilities. Owing to the various format of the papers and the overheads in terms of time and cost associated with integrating them into electronic health records, the majority of hospitals preserve patients' previous clinical documentation as scanned images. Several healthcare practitioners manually de-identify clinical record images with image editing software. Manual de-identification is inefficient in terms of time and cost. As a result, a large number of digital medical record copies remained unutilized.

This work introduces a system that employs deep learning and pattern-matching methodologies to exclude personally identifiable information from healthcare document images autonomously. Medical documents have inconsistent structures, featuring various data formats such as prominent headings, logos, single- and dual-column key-value pairs, and multicolumn tables with or without borders. This variability complicates the accurate identification of Protected Health Information (PHI) within the document, making de-identification a challenging task. Traditional image-to-text conversion methods fail to produce legible text from these images. Segmenting images by their structural composition enhances data retrieval from each segment or region according to its format. This will enhance our understanding of the images. This research seeks to automate the extraction of pertinent data from scanned clinical document images and its subsequent integration into electronic health records (EHRs). This technique primarily aims to de-identify the patient's personal information in these reports, enabling its use in clinical research.

In [2], a full study of Document Layout Analysis (DLA) is presented, along with an evaluation of cutting-edge object detectors using publicly available benchmark datasets and standard evaluation methods. YoloV3 [3] and Faster-RCNN [4] are popular object detection models for analyzing handwritten documents [5], detecting tables in documents [6], and detecting layouts in the scientific literature [7].

Over the past 25 years, researchers have developed various machine learning methods, such as CRF, HMM, decision trees, and SVM, to automate healthcare record de-identification. Recently, deep learning models, particularly in NLP, have made significant advancements in improving de-identification tasks. Deep learning models like LSTM, GRU, and BERT have outperformed traditional machine learning, showing improved accuracy. These techniques have been extensively evaluated on public clinical datasets such as MIMIC-II [8], MIMIC-III [9], i2b2 Corpus [10, 11], and CEGS N-GRID Corpus [12]. These datasets primarily consist of free-text EHR, making them suitable for NLP tasks aimed at de-identifying sensitive information. Despite the progress in textual EHR de-identification, there has been little research on automating the de-identification of clinical document images. This gap in research on image based de-identification is a key focus of this paper, which seeks to address the challenge of extracting and anonymizing the PHI from clinical document images in varied formats.

The majority of de-identification models focus on NLP tasks where PHI must be identified and removed from text data. This involves sequence labeling, where words are categorized based on whether they contain PHI. Early systems used CRF and HMM, but deep learning models have now become the preferred approach. Unlike traditional methods that require handcrafted rules and features, deep learning models learn from large datasets, making them more effective in recognizing complex patterns within text.

In [13, 14], a detailed survey on de-identifying healthcare data highlights techniques like rule-based, machine learning, and hybrid approaches. It emphasizes challenges such as balancing privacy with data utility for research. Recurrent neural networks (RNNs) enhance de-identification processes, particularly LSTM and GRU. A significant study by Dernoncourt et al. [15] utilized a bidirectional LSTM (Bi-LSTM) model. Bi-LSTM, effective in named entity recognition (NER), has shown its effectiveness in identifying and removing PHI from medical records. Bi-LSTM is a commonly employed deep learning methodology for natural language processing, demonstrating efficacy in numerous named entity recognition applications. Multiple studies [16, 17, 18] employed Bi-LSTM for clinical de-identification, providing the likelihood of all possible PHI labels for each token in the sequence.

The introduction of the BERT model by Devlin et al. [19] further revolutionized NLP tasks. BERT is a pre-trained, bidirectional transformer-based model that uses surrounding context to create word embeddings, enhancing contextual understanding. Johnson et al. [20] adapted BERT to enhance de-identification in EHR, showing that it could accurately remove sensitive information like patient names, ages, and social security numbers. This model was evaluated using established criteria like positive predictivity and F1 score, achieving impressive results in de-identifying PHI from healthcare data.

Despite these advances, most research has been limited to de-identifying text-based EHR, leaving a gap in the study of clinical documents in the form of images. The development of systems that can automatically de-identify PHI from document images is an emerging area that needs further exploration to ensure privacy across all forms of healthcare records. This review of de-identification models highlights the success of deep learning in text-based tasks while pointing at the need for new methods to handle clinical images effectively. Identifying the research gaps that exist in the domain of de-identification of clinical documents, this research focuses on the following issues not addressed in the literature:

a. Devising a methodology combining deep learning and pattern matching algorithms for effective de-identification of clinical records
b. Unlike most existing work that focuses on textual documents, this research focuses on scanned images of clinical documents
c. Identifying regions containing PHI and removing them.

II. MATERIALS AND METHODS

FIGURE 1 depicts the process model for the proposed system. We segment the entire methodology into two major components. The upper section uses clinical document images to train a deep-learning model, specifically YoloV3-DLA, for document layout analysis. The lower section uses a trained deep-learning model and pattern-matching techniques to anonymize the images of clinical documents. The PHI fields in these images are mostly condensed in a single location. Segmenting the images based on their structure allows us to identify the regions containing PHI fields. We first examined the document structure through layout analysis and then applied the de-identification method to eliminate personally identifiable information.

A. CLINICAL DOCUMENT IMAGES DATASET

A dataset consisting of clinical document images from patients from various hospitals has been created for experimental purposes. The dataset created for this research purpose consist

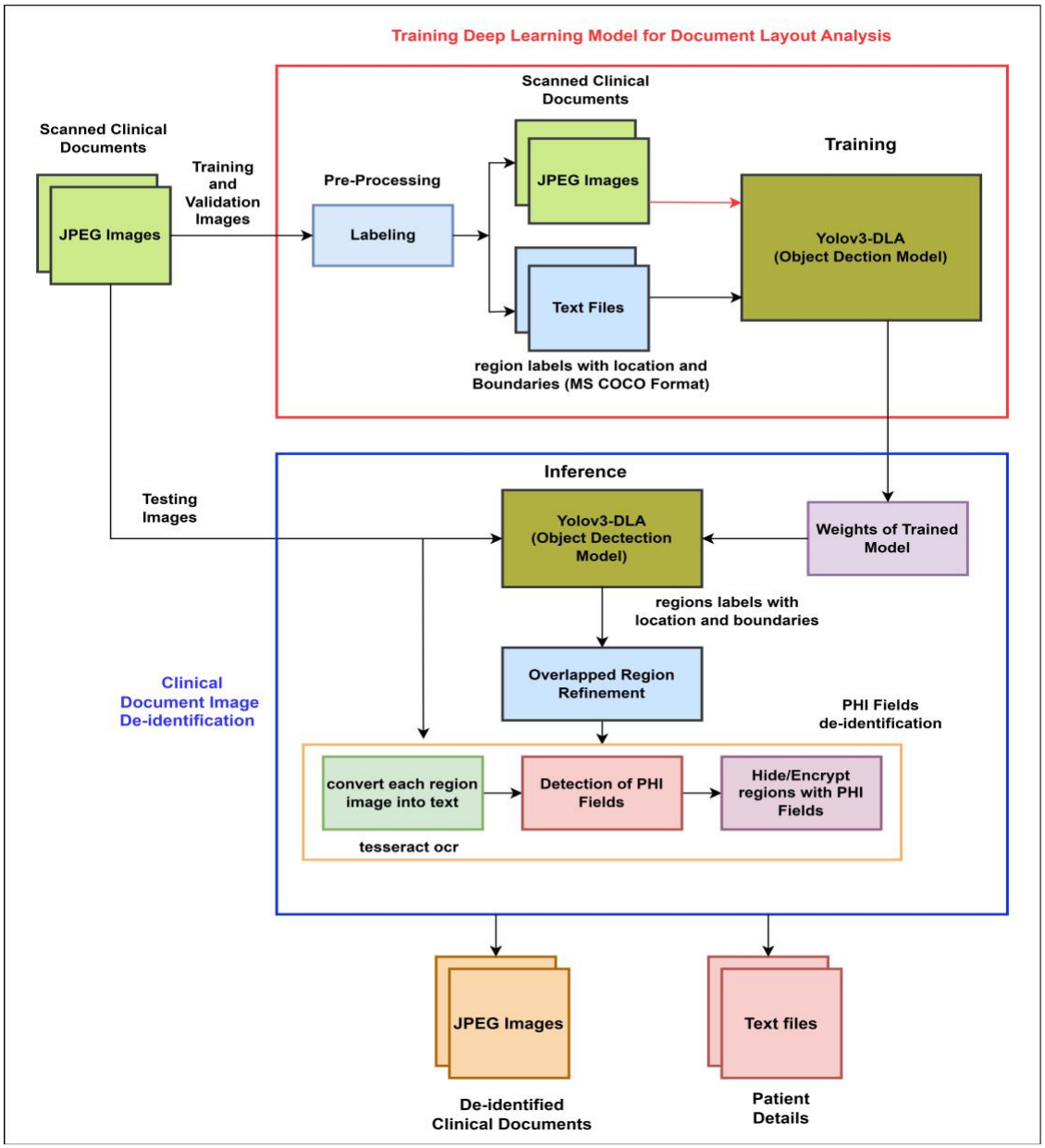


FIGURE 1. Framework of Clinical Document Image De-identification

of images drawn from different hospitals with different report formats. The reports are collected with the patient's consent; however, the dataset remains inaccessible to the public due to their sensitive nature., For this research purpose, 700 images were gathered from 54 individuals. They include discharge summaries, radiological reports, biochemistry, hematology, and laboratory findings. The report formats vary from plain text to embedded logos to single column tables, multi-column tables etc. Hence the 700 images are very diversified justifying the variety of clinical document images that will figure out in real time. Images are stored in the Microsoft COCO [21] format after manually annotating with bounding boxes and class names. The files can be accessed as RGB-formatted JPEG images. TABLE 1 illustrates that the clinical document image layout is complicated, consisting of ten different class labels. There are 700 annotated images in the dataset; 600 are used for training, 50 for validation, and 50 for testing. Each image has a different structure, and reports from the same institution can have different patterns based on the type of report. The dataset's suitability for automated de-identification in NLP research is improved by HIPAA's substitutions and deletions of PHI identifiers. The dataset's PHI distributions are illustrated in TABLE 2.

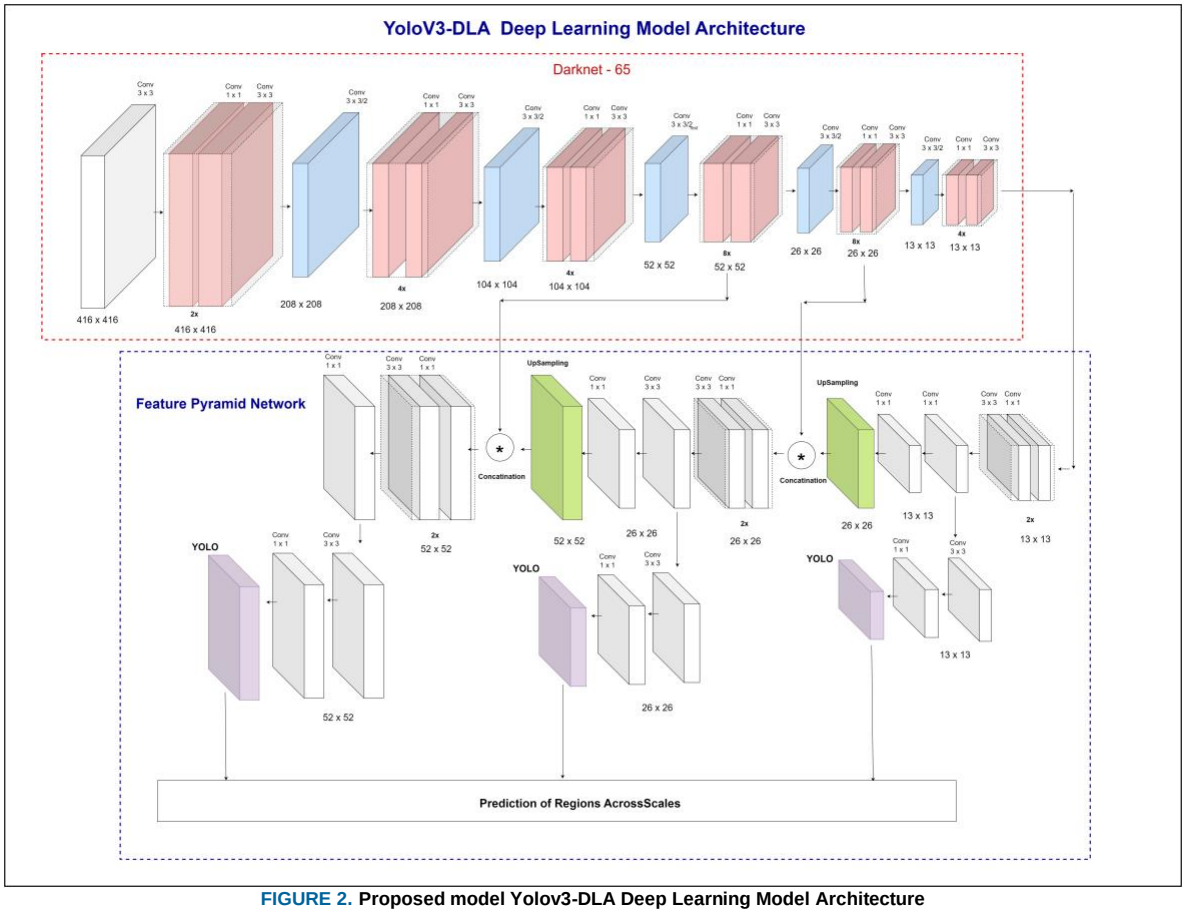


FIGURE 2. Proposed model Yolov3-DLA Deep Learning Model Architecture

TABLE 1
Class Labels of Clinical Document Image Dataset

Class Label	Description
Logo	Logo of an institute/hospital
Table	Multi-column tabular data with border
One_Col_Key_Value	One-column key-value pairs
Two_Col_Key_Value	Two column key-value pairs
Multi_Col_Text	Multi-column tabular data with-out or partial border
QR_Bar_Code	QR code or Barcode
Seal_Stamp	Seal or Stamp of institute/hospital
Signature	Signature of doctors or members of institute/hospital
Image	Photos of patients' organs/ultrasound scans.

TABLE 2
PHI distributions on clinical document images dataset

Category	Sub-Category	No of Tokens	Category	Sub-Category	No of Tokens
Name	Patient	700	Age	Age	700
	Doctor	615	Date	Date	851
	Hospital	425	Contact	Phone	295
	Organization	270		Mobile	610
Location	Dno	226		Fax	65
	Street	482		Email	466
	City	685	ID	Patient ID	485
	State	540		UHID	354
	Country	241		IPNO	583
	Pincode	654			

B. DOCUMENT LAYOUT ANALYSIS USING YOLOV3-DLA

The proposed YoloV3-DLA model enhances the established YoloV3 [3] model. In this process, the dataset is first annotated by creating bounding boxes around relevant regions in the input images and assigning class labels to each region. This annotation step is critical for helping the model identify and accurately classify objects within the images.

Once all the images in the dataset are labeled, the YoloV3-DLA model is trained on this annotated data. The labeled dataset enables the model to learn the relationships between image regions, bounding boxes, and their corresponding classes, allowing it to predict objects and their locations in new, unseen images. The proposed Yolov3-DLA (Document Layout Analysis) structure further improves the model's performance, enhancing its ability to detect objects more precisely. It leads to better object detection outcomes, particularly in complex scenarios, compared to the original YoloV3 [3] model. YoloV3 [3] stands out as a widely utilized deep learning model renowned for its effectiveness in real-time object detection, enabling the identification of specific objects within videos, live feeds, or images. YoloV3 [3] demonstrates superior performance in identifying smaller objects. In document layout analysis, YoloV3 [3] has a higher accuracy in detecting smaller regions, whereas its performance in detecting larger regions is comparatively less precise. The proposed YoloV3-DLA framework specifically introduces the Darknet-65 model for feature extraction. It captures a greater number of features at a higher level. This facilitates the straightforward identification of larger regions. Like YoloV3 [3], YoloV3-DLA is a cohesive neural network that processes the complete image in a single evaluation, efficiently generating predictions for bounding boxes and class probabilities. Figure 2 depicts the network architecture of YoloV3-DLA, comprising two main components: the feature extractor and the feature detector.

1) FEATURE EXTRACTOR

FIGURE 2 illustrates the architecture of the YoloV3-DLA network. YoloV3-DLA employs Darknet-65 as its feature extractor. Darknet-65 comprises of 66 fully convolutional layers, with 60 arranged into six separate residual units. Each module employs a series of 1x1 and 3x3 convolutional layers to link the interactions of adjacent features. These residual units effectively address the problem of vanishing gradients.

The images were adjusted to 416 × 416 while maintaining their original aspect ratio. A convolution layer with a stride of 2 reduces the input images five times before forwarding them to the backbone network, which extracts features across various levels. The convolutional layer applies filters to the input images, generating multiple feature maps. Darknet 65 incorporates more layers with residual connections compared to Darknet 53. Extracting advanced features at higher levels enhances the identification of larger objects.

2) FEATURE DETECTOR

The YoloV3-DLA model divides the input image into a grid of SxS cells, with S varying among different versions of Yolo. In this study, the model generates a 13x13 grid by setting S to 13. Each cell in the grid is responsible for detecting objects whose centers fall within it. The model predicts each cell's bounding box, objectness score, and class confidence level. The objectness score reflects the likelihood that the cell contains an object, while the class confidence represents the probability that the object belongs to a particular category. To predict bounding boxes, YoloV3-DLA uses anchor boxes based on dimension clusters. Each bounding box in an object detection model encapsulates several key attributes that provide essential information about the detected object. These attributes include the object's center coordinates, which indicate its position within the image, along with the dimensions of the bounding box itself, describing its width and height. Additionally, the bounding box contains an objectness score, which reflects the likelihood that the box contains an object, and a class confidence score, indicating how confidently the model predicts the class of the detected object. A log-space transformation is applied to the data to effectively estimate the bounding box's size. This transformation allows for better handling of the varying scales of objects in the dataset. After this transformation, the size of the bounding box is further refined by multiplying it with a predefined anchor, which serves as a reference size. Logistic regression is employed to predict the objectness score. This statistical method provides a probability estimate for the presence of an object within the bounding box. Moreover, binary cross-entropy loss is used to assess the accuracy of class predictions, which quantifies the difference between the predicted class confidences and the actual classes, enabling optimization of the model's performance.

YoloV3-DLA improves object detection by predicting bounding boxes across three distinct scales, enhancing its ability to detect objects of varying sizes. This multi-scale prediction method resembles the feature pyramid networks (FPN) [22] approach, which extracts features at different levels. The detection network generates feature maps from the backbone network with dimensions of 13x13, 26x26, and 52x52, forming a feature pyramid that improves detection across scales. For the initial scale (13x13), bounding boxes are predicted using multiple convolutional layers applied to smaller feature maps. Up-sampling combines smaller feature maps with larger ones to predict bounding boxes at larger scales (26x26 and 52x52). Further convolutional layers are applied to the aggregated feature maps to refine bounding box predictions. This process is repeated for all three scales to ensure accurate detection at multiple levels.

The final detection layer processes input from all three scales, using a regression approach to predict bounding boxes. An objectness score is used to refine predictions, with bounding boxes below a certain threshold being excluded from consideration. To handle overlapping bounding boxes and reduce duplicate detections of the same object, YoloV3-DLA applies non-maximum suppression (NMS). NMS operates using a method called "Intersection over Union" (IoU), which compares the overlap between predicted boxes and suppresses redundant ones.

YoloV3-DLA demonstrates notable strengths in its ability to detect objects accurately, offering flexibility in configuring detection thresholds and a high processing speed. By integrating multi-scale predictions, dimension clustering, and advanced suppression techniques, the model enhances the

precision and reliability of object detection, making it suitable for a wide range of applications where detecting objects of different sizes is critical. These improvements over the original YoloV3 [3] model contribute to its effectiveness in complex detection tasks.

C. DE-IDENTIFICATION USING PATTERN MATCHING

Clinical document images exhibit inconsistent structures, incorporating diverse data formats, including prominent headings, logos, single- and dual-column key-value pairs, and multicolumn tables with or without borders. It limits the precise identification of Protected Health Information (PHI) within the document, making de-identification challenging. The PHI in these clinical documents is primarily concentrated in the top and bottom sections. The proposed de-identification algorithm initially segments the image according to its structural content through document layout analysis. The algorithm subsequently checks if each segment comprises personal health information (PHI). If the segment or region contains PHI, the algorithm will extract the PHI, store it separately for future re-identification, delete the segment containing PHI from the image, and assign a unique identification number to the image for future reference.

ALGORITHM 1. De-identification of Clinical Document Images

Require: Clinical Document Image and Trained Yolov3-DLA Model Weights
Ensure: De-identified Clinical Document Image and Patient Details as Text File

```
p_regions ⇐ Predict(inputImage,TrainedModelWeights)
regions ⇐ regionRefinement(p_regions)
for region in regions do
    tokens ⇐ ImagetoText(region)
    Initialize PTokens = []
    for token in tokens do
        if isPHIToken(token) then
            PTokens.add(token)
        end if
        if isConsiderablePHI(PTokens) then
            patientDetails ⇐ getPatientDetails(tokens)
            PHIregions.add(region)
        end if
    end for
end for
outputImage ⇐ removeRegions(PHIregions,inputImage)
savePatientDetailsAsTextFile(patientDetails)
```

In this process, the document layout analysis of the clinical document image is performed using a trained YoloV3-DLA model. Next, pattern-matching techniques will be applied to identify the PHI tokens in each identified region. Finally, those regions with PHI tokens were removed from the clinical document image, as depicted in the lower half of Figure 1. The clinical document image de-identification method is described in ALGORITHM 1. The de-identification process starts by figuring out the layout of the input image. The Predict function is invoked, utilizing the weights of the YoloV3-DLA trained model. The refinement function then enhances the identified regions by adjusting overlapping area boundaries to segment the image precisely. Next, The ImagtoText function transforms each region in the image into text using Tesseract-OCR. Subsequently, confirm all tokens by invoking the isPHIToken function. The isPHIToken function determines if the provided input token corresponds to a PHI field by utilizing regular expressions to identify standard keywords (such as name, age, date, patient ID, etc.) and key-value pairs. The isConsiderablePHI function recommends removing the region after it has evaluated the PHI tokens that have been detected. Finally, the identified regions will be removed and stored with separate PHI tokens, assigning them a unique identifier for future re-identification.

III. RESULTS

The proposed model was evaluated in two stages. We first analyzed clinical document image layouts using YOLOv3-DLA, achieving an F1 score of 0.92. This stage accurately identified key regions within the documents. Secondly, we applied a de-identification method to eliminate regions that contained Protected Health Information (PHI). The model identified standard keywords, dates, and numbers using pattern-matching techniques like regular expressions. This process effectively removed sensitive data, achieving an F1 score of 0.97. Combining object detection with de-identification provides a robust solution for protecting patient privacy in clinical documents. These results demonstrate the model's effectiveness in document layout analysis and PHI removal, which supports secure document processing.

A. YOLOV3-DLA PERFORMANCE

The YoloV3-DLA model was trained and validated on a single NVIDIA Quadro GV100 for 250 epochs. The batch size was set to 8, and the weights were initialized arbitrarily. The optimization method used was stochastic gradient descent (SGD). Initially, the learning rate is 0.003, decreasing by 10% after 200 and 225 epochs. The F1-score, precision, and recall of the Yolov3-DLA model were maximized at an IoU threshold of 0.5 and achieved an F1-score of 0.92. TABLE 3 presents the performance of YoloV3-DLA and other object detection techniques on clinical document image datasets. FIGURE 3 compares the performance of document layout analysis using object detection models. FIGURE 4 shows the Precision, Recall, mAP@0.5, and F1-score graphs of YoloV3-DLA on the validation dataset of clinical document images. Sample outputs of YoloV3-DLA are shown in APENDIX-1.

TABLE 3
Performance of Yolo-V3 on clinical documents images dataset.

Model	Precision	Recall	F1-Score
Faster-RCNN- VGG17 [4]	0.74	0.78	0.76
Faster-RCNN- ResNet [4]	0.8	0.79	0.79
YoloV3- Tiny [3]	0.63	0.68	0.65
YoloV3 [3]	0.89	0.88	0.88
YoloR [23]	0.82	0.9	0.86
YoloV3-DLA (Proposed)	**0.91**	**0.93**	**0.92**

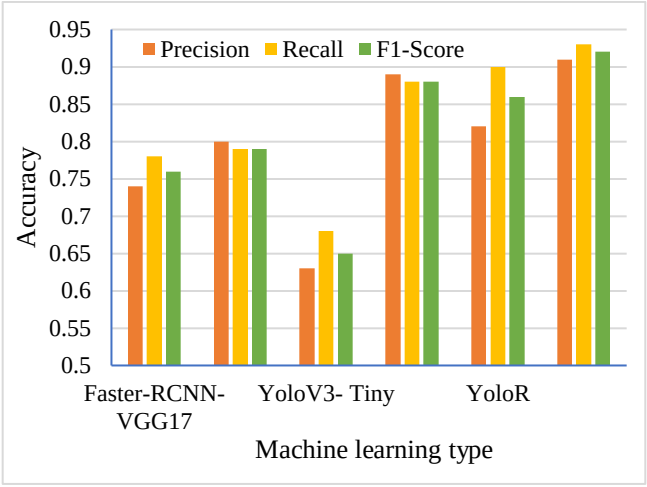


FIGURE 3. Performance of Document Layout Analysis

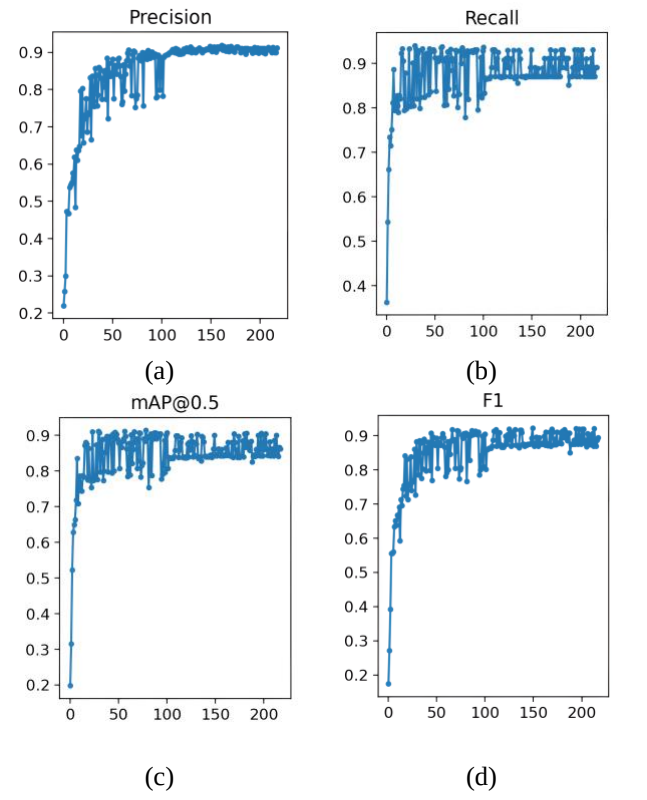


FIGURE 4. Performance Graph of YoloV3-DLA on Validation dataset (a) Precision, (b) Recall, (c) mAP, (d) F1-Score.

B. DE-IDENTIFICATON PERFORMANCE

The literature review revealed no findings regarding de-identifying clinical document images. Many researchers have investigated electronic health records and unstructured clinical notes. A comparative analysis was conducted to evaluate the efficacy of the proposed model in relation to existing de-identification approaches. Experiments were conducted utilizing various pre-trained BERT models to identify protected health information (PHI) tokens within clinical records. We employed Tesseract OCR to convert the image into text. We subsequently pre-processed the text and input the tokens into pre-trained BERT models to determine their classification as PHI or non-PHI categories. PHI (Protected Health Information) is removed from the clinical document image using the stored location during the image-to-text conversion. Table 4 presents the performance results of the proposed de-identification model as specified in ALGORITHM 1 and existing BERT models on clinical document images dataset. The proposed algorithm achieved an F1-score of 0.97. Figure 5 compares the performance of de-identification techniques on the clinical document images dataset. Sample outputs of de-identification are shown in APENDIX-1.

TABLE 4
Performance of de-identification on clinical documents images dataset

Model	Precision	Recall	F1-Score
$BERT_{base}$ [19]	0.84	0.83	0.84
BioBERT [24]	0.86	0.85	0.86
SciBERT [25]	0.91	0.87	0.89
ClinicalBERT [26]	0.88	0.92	0.9
RoBERTa [27]	0.9	0.89	0.9
BERT by Johnson [20]	0.93	0.91	0.92
Proposed De-identification Algorithm	**0.97**	**0.96**	**0.97**

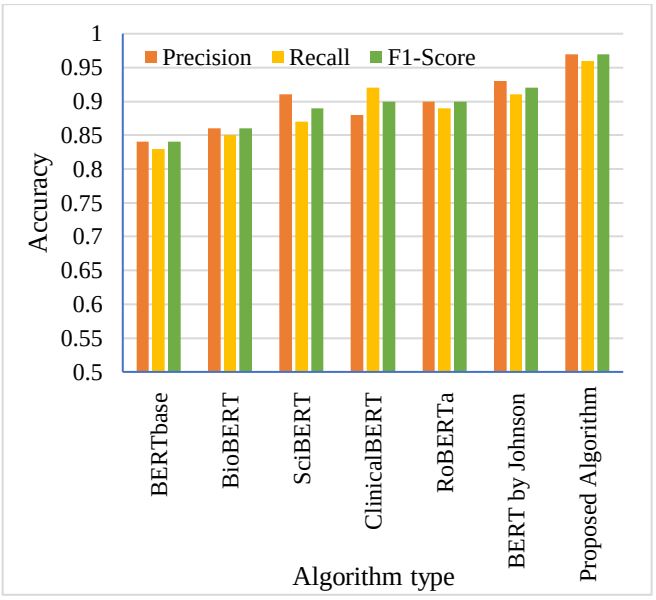


FIGURE 5. Performance of de-identification

IV. DISCUSSION

The YoloV3-DLA model achieved an F1 score of 0.92 in document layout analysis. The footer of clinical document images displays the addresses and other institution details in a small font with a variety of styles. Due to the low quality of the document images and the presence of text in areas with inconsistent font styles and sizes, the model struggles to identify the contents at the bottom. Furthermore, the model misclassified some regions due to similar structural representation; this may be overcome by increasing the training data.

In the de-identification process, the proposed algorithm performs exceedingly with a high precision and recall., whereas the overall performances of all pre-trained BERT models, except for Johonson's BERT [19], are less, differing by 1% across all criteria. Johnson [20] presents the BERT model, attaining a notable F1 score of 92.45%. The BERT model, as developed by Johnson [20], is an advanced method that enhances the original BERT framework. Training was conducted using de-identification datasets sourced from the i2b2 2006 corpus [10], i2b2 2014 corpus [11], Physionet corpus [28], and Dernoncourt-Lee corpus [15]. BERT provides supplementary training datasets that are comparable to other pre-trained models. This specialized clinical training produced the most favorable outcomes. BERT achieved an F1 score of 98.82 on the i2b2 2014 [11] challenge test set, as noted in reference [20] & [29] and confirmed in reference [11] & [30]. These document images contain inconsistent clinical information. The formatting and types of reports in these documents differ among various health institutions. The translation of images into text fails to represent the information accurately. The text in these images may display different font styles and sizes. The text exhibits a lack of uniform alignment. A significant challenge in digitizing physical documents is the degradation of certain materials over time. The document includes the health institute's name, address, contact details, and logo in the header and footer sections, employing diverse font styles and sizes. The proposed de-identification algorithm tackles all the above-mentioned challenges by analyzing the document and identifying the PHI fields with high precision and recall. The YoloV3-DLA model attained an F1 score of 0.92 in document layout analysis.

In some cases, the model fails to recognize the upper and lower sections of the clinical documents due to the poor quality images and the presence of text in those regions, marked by inconsistent font styles and sizes. Additionally, due to the close proximity and similar structural features, the model incorrectly labels some regions as a multi_col_text instead of a One_Col_Key_Value. This slightly impacts the accuracy. Increasing the training dataset size with different image formats from various hospitals will definitely reduce these issues. Due to exclusive training on clinical document images, the proposed model can de-identify only these clinical images. In future, the model can be trained on other scanned documents to identify and extract data from all types of documents.

V. CONCLUSION

The proposed work presents a comprehensive framework for de-identifying clinical document images using deep learning-based object detection and pattern matching techniques. The process consists of two stages: Initial document layout analysis is conducted using YoloV3-DLA, followed by the de-identification of regions containing PHI fields through the application of regular expressions and common keywords. The proposed Yolov3-DLA method utilizes Darknet-65 and Feature Pyramid Network (FPN) to capture and integrate deeper features, resulting in advanced performance outcomes. Contextual factors, such as the document structure, its location, and the size of the involved regions, influence the model's effectiveness. The experiments were performed on a synthetic dataset of clinical document images, indicating that hyper-parameter optimization yielded positive outcomes in document layout analysis. Yolov3-DLA demonstrated superior performance relative to alternative object detection models. The proposed de-identification algorithm uses a pattern-matching method to achieve higher accuracy by exploiting the characteristic features of these clinical document images. Subsequent research will focus on automatically extracting literal information from these documents. In terms of real-world applications, this approach could significantly improve privacy and data security in healthcare settings, enabling more efficient sharing of clinical documents for research purposes while maintaining patient confidentiality. Further research will focus on enhancing the extraction of other information from these documents, such as disease diagnoses and treatment history, and exploring the potential of deploying this framework in real-world clinical environments.

REFERENCES

[1] States. U., "Health insurance portability and accountability act of 1996," 1996.

[2] Nguyen TT, Le H, Nguyen T, et al.," A brief review of state-of-the-art object detectors on benchmark document images datasets." *International Journal on Document Analysis and Recognition (IJDAR),*2023. doi.org/10.1007/s10032-023-00431-0

[3] Redmon, Joseph and Ali Farhadi., "YOLOv3: An Incremental Improvement," *ArXiv abs/1804.02767*, 2018.

[4] Ren S, He K,Girshick R, et al., "Faster RCNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, 2015 doi.org/10.1109/TPAMI.2016.2577031

[5] Ravichandra S, Siva Sathya S, Lourdu Marie Sophie S., "Deep learning based document layout analysis on historical documents." *In: Advances in Distributed Computing and Machine Learning. Springer Nature* Singapore, pp 271–281, 2022. doi.org/https://doi.org/10.1007/978-981-19-1018-0 23

[6] Zhang D, Mao R, Guo R, et al., "Yolo-table: disclosure document table detection with involution," *International Journal on Document Analysis and Recognition (IJDAR)* vol. 26(1), pp. 1–14, 2023 doi.org/10.1007/s10032-022-00400-z

[7] Yang H, Hsu W., "Vision-based layout detection from scientific literature using recurrent convolutional neural networks." *25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, pp. 6455–6462, 2021.doi.org/10.1109/ICPR48806.2021.9412557

[8] Saeed M, Villarroel M, Reisner AT, et al., "Multiparameter intelligent monitoring in intensive care ii: A public-access intensive care unit database," *Critical Care Medicine*, vo. 39, pp. 952–960, May. 2011. doi: 10.1097/CCM.0b013e31820a92c6.

[9] Johnson, Alistair E W et al. "MIMIC-III, a freely accessible critical care database," *Scientific data,* vol. 3 160035, 24 May. 2016. doi:10.1038/sdata.2016.35

[10] Uzuner, Ozlem et al. "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*: JAMIA , vol. 14, no. 5, pp. 550-63, 2007. doi:10.1197/jamia.M2444

[11] Stubbs, Amber, and Özlem Uzuner. "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus," *Journal of Biomedical Informatics,* vol. 58 Suppl, pp. S20-S29, 2015. doi:10.1016/j.jbi.2015.07.020

[12] Stubbs, Amber, et al. "De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1," *Journal of Biomedical Informatics,* vol. 75S, pp. S4-S18, 2017. doi:10.1016/j.jbi.2017.06.011

[13] Sriram R, Sundaram SS, Sophie SL, "Deep learning models for automatic de-identification of clinical text," *In: Computer, Communication, and Signal Processing. AI, Knowledge Engineering and IoT for Smart Systems. Springer Nature Switzerland*, pp. 116–127, 2023. doi.org/10.1007/978-3-031-39811-7 10
[14] Kovacevic, Aleksandar et al. "De-identification of clinical free text using natural language processing: A systematic review of current approaches," *Artificial intelligence in medicine.* Vol. 151, 2023. doi.org/10.1016/j.artmed.2024.102845.
[15] Dernoncourt, Franck et al. "De-identification of patient notes with recurrent neural networks." *Journal of the American Medical Informatics Association : JAMIA* vol. 24, no. 3, pp. 596-606, 2017. doi:10.1093/jamia/ocw156
[16] Ahmed T, Aziz MMA, Mohammed N., "De-identification of electronic health record using neural network," *Scientific Reports* vol. 10, 2020. doi.org/10.1038/s41598-020-75544-1
[17] Catelli R, Casola V, De Pietro G, et al., "Combining contextualized word representation and sub-document level analysis through bi-lstm+crf architecture for clinical de-identification." *Know-Based System*, vol. 213, 2021. doi.org/10.1016/j.knosys.2020.106649
[18] Hartman T, Howell MD, Dean J, et al., "Customization scenarios for de-identification of clinical notes," *BMC Medical Informatics and Decision Making* vol. 20(1), 2020:14. doi.org/10.1186/s12911-020-1026-2
[19] Devlin J, Chang MW, Lee K, et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv abs/1810.04805,* 2019.
[20] Johnson A, Bulgarelli L, Pollard T., "Deidentification of free-text medical records using pre-trained bidirectional transformers," *Proceedings of the ACM Conference on Health, Inference, and Learning,* vol. pp. 214–221, 2020. doi.org/10.1145/3368555.3384455
[21] Lin TY, Maire M, Belongie SJ, et al., "Microsoft coco: Common objects in context," *In: European Conference on Computer Vision,* 2014.
[22] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie., "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 936-944, 2017. doi: 10.1109/CVPR.2017.106.
[23] Wang, Chien-Yao & Yeh, I-Hau & Liao, Hong-yuan., "You Only Learn One Representation: Unified Network for Multiple Tasks," *10.48550/arXiv.2105.04206*, 2021.
[24] Beltagy, I., Lo, K., and Cohan, A. "SciBERT: A Pretrained Language Model for Scientific Text", *Conference on Empirical Methods in Natural Language Processing*., 2019. doi: 10.18653/v1/D19-1371
[25] Jinhyuk L, Wonjin Y, et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, Vol. 36, Issue. 4, pp. 1234–1240, February 2020, doi.org/10.1093/bioinformatics/btz682.
[26] Alsentzer, E., Murphy, et al., "Publicly Available Clinical BERT Embeddings". *ArXiv*, 2019, doi: https://arxiv.org/abs/1904.03323
[27] Liu, Y., Ott, M., et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*, 2019, doi: https://arxiv.org/abs/1907.11692
[28] Neamatullah, I., Douglass, M.M., Lehman, Lw.H. et al. "Automated de-identification of free-text medical records." BMC Medical Informatics Decision Making, vol. 8, 32, 2008. doi: https://doi.org/10.1186/1472-6947-8-32
[29] Cedric L, Bertrand L, et al., "Evaluating the Impact of Text De-Identification on Downstream NLP Tasks.", In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 10–16, 2023.
[30] Azzouzi, M.E., Coatrieux, G., Bellafqira, R. et al. "Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models.", BMC Medical Informatics and Decision Making, vol. 24(54), 2024. https://doi.org/10.1186/s12911-024-02422-5

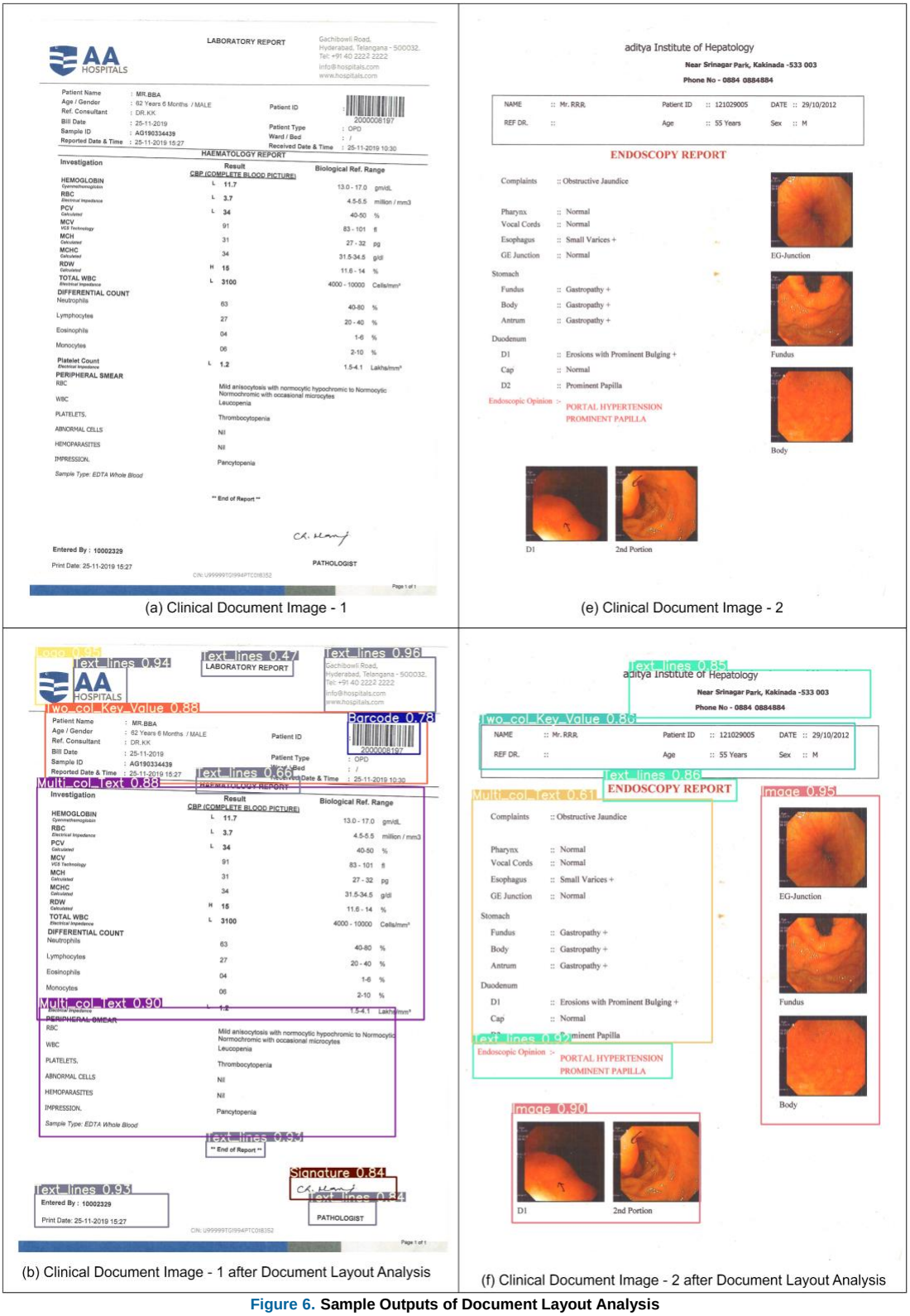
AUTHORS BIOGRAPHY

Mr. Ravichandra Sriram is presently a research scholar in the Department of Computer Science at Pondicherry University, Puducherry. He has completed his B.Tech in Information Technology from Swarnandhra College of Engineering and Technology, Narsapur in 2007 and M.Tech in Computer Science and Technology at Andhra University, Visakhapatnam in 2010. He worked as an Assistant Professor at Shri Vishnu Engineering College for Women, Bhimavaram from 2011 to 2019. His research interests include Computer Vision, Natural Language Processing, and Deep Learning. He has authored and co-authored some publications which includes journals and international conferences in the field of Computer Science. He can be contacted at email: ravichandrasriram@gmail.com

Dr. S. Siva Sathya completed her M.Tech and Ph.D. in Computer Science and Engineering from Pondicherry University. She has 25 years of teaching experience and specializes in Evolutionary and Bio-inspired Computing, Spatio-Temporal Data Mining, VANET and Natural Language Processing. She is UGC NET qualified and has published several research articles in reputed journals. She is the recipient of the Naari Shakthi Award 2017 from the President of India for her mobile Innovation "MITRA" for Women safety. She has also received the Chairman's Distinction award in the South Asia's mBillionth Mobile innovation contest. She can be contacted at email: ssivasathya@pondiuni.ac.in.

Mrs. S. Lourdumarie Sophie is presently a research scholar in the Department of Computer Science and Engineering at Pondicherry University, Puducherry. She has completed her B.Tech in Computer Science and Engineering from Manakula Vinayagar Institute of Technology, Puducherry in 2015 and M.Tech in Computer Science and Engineering at Pondicherry University, Puducherry in 2019. She worked as an Assistant System Engineer at Tata Consultancy Service, Chennai, from 2015 to 2017. She also has a year of teaching experience as a guest faculty from Pondicherry University in 2019-2020. She qualified for the UGC NET examination in 2019. Her research interests include Natural Language Processing, Machine Learning, and Deep Learning. She has authored and co-authored more than 10 publications, which include journals and international conferences in the field of Computer Science. She can be contacted at email: lourdumariesophie15@gmail.com.

APENDIX 1



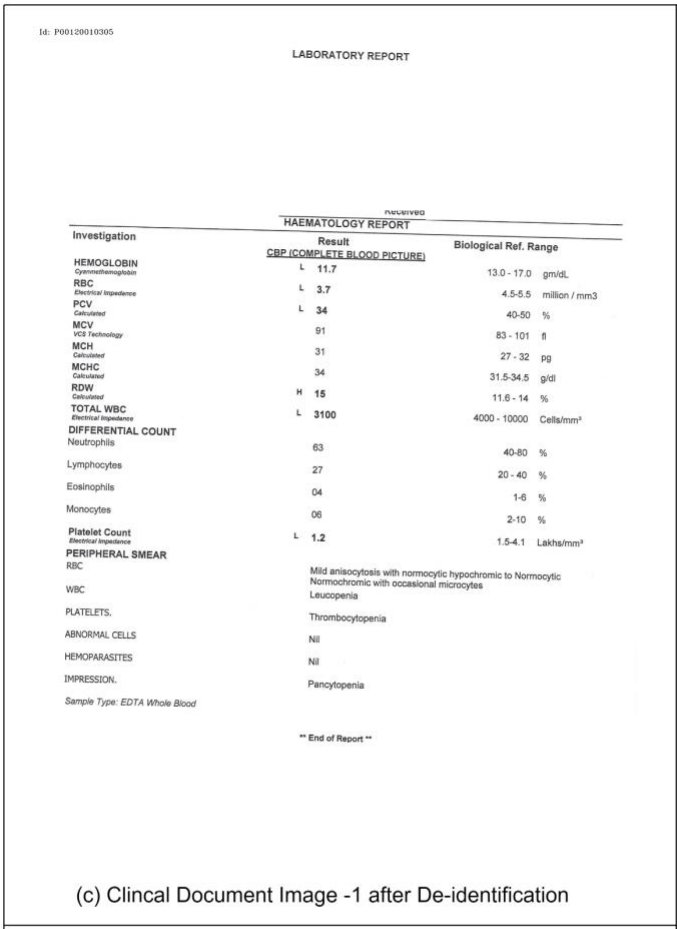
(a) Clinical Document Image - 1

(e) Clinical Document Image - 2

(b) Clinical Document Image - 1 after Document Layout Analysis

(f) Clinical Document Image - 2 after Document Layout Analysis

Figure 6. Sample Outputs of Document Layout Analysis



Id: P00120010305

LABORATORY REPORT

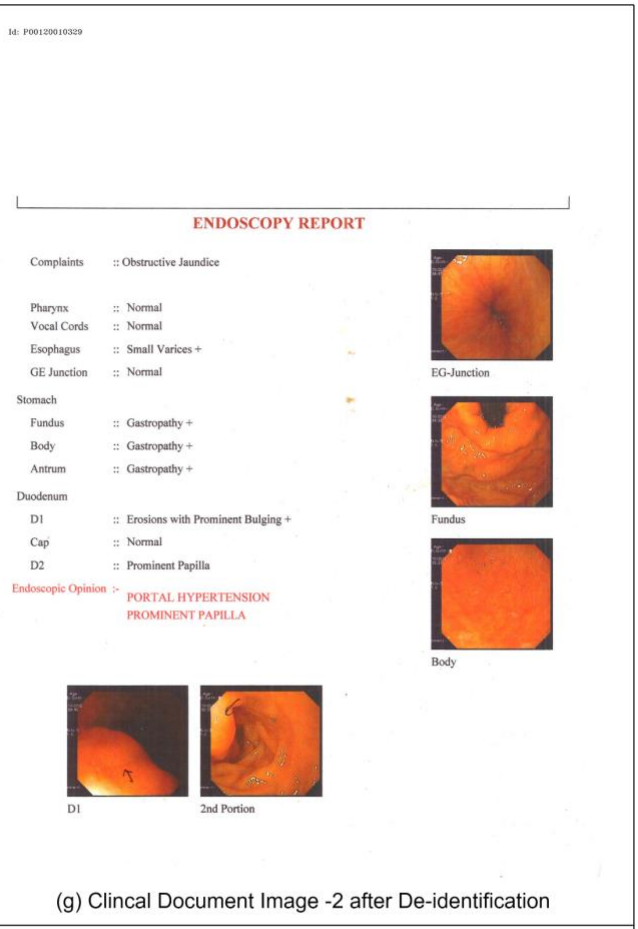
HAEMATOLOGY REPORT

Investigation	Result	Biological Ref. Range	
CBP (COMPLETE BLOOD PICTURE)			
HEMOGLOBIN (Cyanmethemoglobin)	L 11.7	13.0 - 17.0	gm/dL
RBC (Electrical Impedence)	L 3.7	4.5-5.5	million / mm3
PCV (Calculated)	L 34	40-50	%
MCV (VCS Technology)	91	83 - 101	fl
MCH (Calculated)	31	27 - 32	pg
MCHC (Calculated)	34	31.5-34.5	g/dl
RDW (Calculated)	H 15	11.6 - 14	%
TOTAL WBC (Electrical Impedence)	L 3100	4000 - 10000	Cells/mm³
DIFFERENTIAL COUNT			
Neutrophils	63	40-80	%
Lymphocytes	27	20 - 40	%
Eosinophils	04	1-6	%
Monocytes	06	2-10	%
Platelet Count (Electrical Impedence)	L 1.2	1.5-4.1	Lakhs/mm³
PERIPHERAL SMEAR			
RBC	Mild anisocytosis with normocytic hypochromic to Normocytic Normochromic with occasional microcytes		
WBC	Leucopenia		
PLATELETS.	Thrombocytopenia		
ABNORMAL CELLS	Nil		
HEMOPARASITES	Nil		
IMPRESSION.	Pancytopenia		

Sample Type: EDTA Whole Blood

** End of Report **

(c) Clincal Document Image -1 after De-identification



Id: P00120010329

ENDOSCOPY REPORT

Complaints	::	Obstructive Jaundice
Pharynx	::	Normal
Vocal Cords	::	Normal
Esophagus	::	Small Varices +
GE Junction	::	Normal
Stomach		
Fundus	::	Gastropathy +
Body	::	Gastropathy +
Antrum	::	Gastropathy +
Duodenum		
D1	::	Erosions with Prominent Bulging +
Cap	::	Normal
D2	::	Prominent Papilla

Endoscopic Opinion :- PORTAL HYPERTENSION
PROMINENT PAPILLA

EG-Junction

Fundus

Body

D1

2nd Portion

(g) Clincal Document Image -2 after De-identification

```
P00120010305.jpg
 Tel :  +91 40 2222 2222
 Entered By  :  10002329
 Patient Name  :  MR. BBA
 Age / Gender  :  62 Years 6 Months / MALE
 Patient ID AAT Ref. Consultant  :  DR. KK       2000008197
 Bill Date  :  25-11-2019
 Patient Type > :  OPD
 Sampie ID  :  AG190334439 Ward / Bed 2 of
 Reported Date Time  :  25-11-2019 15:27
 Received Date Time  :  25-11-2019 10:30
```

(d) PHI Fields extracted from Clinical Document Image -1

```
P00120010329.jpg
 NAME  :  Mr. RRR
 Patient ID 1 :  121029005
 DATE  :  29/10/2012
 Age  : 55 Years
 Sex  : M
```

(h) PHI Fields extracted from Clinical Document Image - 2

Figure 7. Sample Outputs of De-identification