Journal of Computer Science

Research Article

Enhanced RoBERTa Model for OCR-Based EHR Parsing and Information Extraction

Balaji Ganesh Rajagopal[1], Amarnath C[1], Chidambaram Sawri Rajan[2] and Deebalakshmi Ramalingam[1]

[1]*School of Computing, SRM Institute of Science and Technology, Tiruchirappalli, India*
[2]*Chidambaram Sawri Rajan CEO, EMEDLOGIX SOLUTIONS Chennai, India*

Article history
Received: 06-03-2025
Revised: 20-01-2026
Accepted: 24-01-2026

Corresponding Author:
Deebalakshmi Ramalingam
School of Computing, SRM Institute of Science and Technology Tiruchirappalli, India
Email: deebalakshmi.r@ist.srmtrichy.edu.in

Abstract: Healthy source data for medical research and health analytics in general can be obtained from Electronic Health Records (EHRs). Nevertheless, due to the complexities of the design and especially the unstructured nature of them, it is not easy to extract important information from digital documents. This paper proposes a fundamentally new approach to the problem of interpreting EHRs obtained by Optical Character Recognition (OCR) that utilizes a refined RoBERTa foundation architecture. Basically, our method is very efficient in extracting key elements, like section headings and bold words, which most of the time have very significant clinical significance. More than just straightforward text recognition is the use of RoBERTa for semantic understanding. 89. 2% is the accuracy of the tests that we have performed. This paper presents an exhaustive benchmarking of the pros and cons of the deep learning techniques that are currently being used for parsing EHRs. However, our model is fixing the problem of very accurately extracting bold section heads from unstructured data in EHRs. The system proposes a two-phase approach combining natural language and image processing techniques. Performing thinning and normalizing operations first to separate bold texts based on pixel intensity over a preset threshold. By successfully removing the needless text from the paragraphs, our method significantly enhances the accuracy of bold word extraction, reaching 98%.

Keywords: SOTA DL Models, Bold Text Extraction, Section Header, EHR Parsing, Clinical NLP

Introduction

Medical records should be effectively managed and analyzed to serve both healthcare research and provide optimum patient care. Conventional methods of parsing medical records, such as rule-based and template-based systems, have certain limitations in understanding the complexities posed by such records and generally require a lot of manual adjustments. This work explores how deep learning, based techniques may be able to cure these issues. Deep learning models, mainly Convolutional Neural Networks (CNNs) and transformer architectures such as BERT and GPT, are the ones that open up the possibility that complex representations of data can be generated entirely independently, i.e., without the need for any external assistance. The ability of these models to do this makes them very appropriate for dealing with unstructured texts like handwriting, medical jargon, and different writing styles, as well as to be able to accommodate changes in medical datasets. Moreover, the authors delve into the Utilization of Hugging Face's RoBERTa base model for the decoding of medical records. A robust offshoot of the BERT architecture, essentially, RoBERTa offers a semantic understanding capability that goes far beyond simple text recognition. Transformer architecture enables the model to use context to understand the text, so that it can tell, for instance, what patient information, prescriptions, and diagnoses are related to each other. This level of semantic understanding also allows the model to identify the relevant information even in the text fragments, such as bold texts or section headings, which generally have a strong clinical significance.

However, there are some challenges specific to applying DL models to medical records. For example, it might become difficult in the health domain, where there

are strong confidentiality issues, to gather a great amount of labeled training data. Other key factors to consider are making sure the model is not only accurate, but it can also be generalized to a wide variety of medical record types and terminologies. Moreover, due to the inherent lack of clear interpretability of deep learning models, it might be challenging to confirm their predictions in a high-stakes healthcare setting. This paper makes an effort to go beyond the present limitations by evaluating the performance of the newest deep learning architectures, e.g., RoBERTa base, in the task of medical record smart parsing.

The model is geared predominantly toward major information extraction, i.e., higher-level text features or section headings, which generally contain the main clinical information. This work lays a firm groundwork for subsequent medical record parsing systems that are more robust and intelligent by thoroughly examining the pros and cons of the current systems. The main goals are to raise the quality of patient care and make health data more accessible and easier to analyze.

Related Work

Natural Language Processing (NLP) for Clinical Text Classification

The classification and understanding of clinical text data have changed drastically within the healthcare industry thanks to the use of NLP methods. Raza and Schwartz (2023) revealed NLP's capability of identifying risk factors in clinical texts as they used it for Entity and Relation Extraction (ERE) from COVID-19 reports. Besides that, a thorough review of the extraction of information from Electronic Medical Records (EMRs) was done by Landolsi et al. (2023), who increased the emphasis on the necessity of NLP for clinical data analysis (Ho et al., 2024). Some of the research that has allowed the application of hospital readmissions forecasting by using NLP tools to first extract relevant information from EMRs is that referenced above.

Deep Learning and Transformer-Based Models in Clinical Text Analysis

Deep learning and, more particularly, transformer models have unequivocally revolutionized the clinical text classification domain. Hsu et al. (2022) developed a deep learning NLP data pipeline for extracting information from scanned Electronic Health Records (EHRs). Their study was specifically focused on predicting hospital readmission. Qasim et al. (2022) unveiled a BERT fine-tuned transfer learning technique for text classification tasks, thereby confirming that deep learning models could be powerful tools in clinical settings. Goodrum et al. (2020) investigated the automatic classification of scanned EHRs, thus stressing the significance of deep learning in the extraction of clinical documents of various formats.

The latest technological breakthroughs have made these techniques more precise than ever (Islam et al., 2023; Koroteev, 2021; Soni and Roberts, 2022). Batista and Evsukoff (2023) carried out a systematic review of the literature on transformer-based methods in EMRs, thereby proving the versatility of these methods in handling different NLP tasks. Rupp et al. (2023) came up with ExBEHRT, an advanced transformer model that incorporates multimodal EHR data, thus allowing the model to predict the disease subtypes and progressions with higher precision. Nerella et al. (2023) provided an extensive review on transformer applications in healthcare, identifying transformers as one of the key technologies for clinical diagnosis and data analysis.

Challenges in Clinical Text Classification

Working on clinical text data definitely has its set of challenges. One of the challenges is the complex medical terms and the unstructured format of EHRs. In their work, Landolsi et al. (2023) argued that rule-based methods are not enough to deal with domain-specific language. Also, issues like noisy data, inconsistencies, and complex document structures are other things that make the classification task quite challenging.

Integration of OCR and NLP in EHR Parsing

Combining Optical Character Recognition (OCR) technology and Natural Language Processing (NLP) has opened up the possibility of extracting information from scanned medical documents. Moezzi et al. (2022) utilized a transformer-based method, which was demonstrated to be capable of producing structured radiology reports, hence leading to improved information extraction from free text formats. Hsu et al. (2021) developed a deep learning, based NLP data pipeline for extracting information from EHR, scanned documents, which led to high accuracy in processing scanned reports. To put it differently, these papers highlight the effectiveness of OCR and NLP techniques when leveraged in combination for EHR parsing.

Advancements in Pre-Trained Language Models (PLMs) for Scientific Text

Pretrained Language Models (PLMs) are capable of understanding scientific and clinical texts. Liu et al. (2019) pointed out that scientific language models (SciLMs) have potential not only in different scientific areas but also in general NLP tasks. Our present work deals with identifying section headings in clinical documents, but later on, we can consider using advanced PLMs specifically designed for scientific and clinical domains for the purpose of document classification and experiment improvement (Karthikeyan et al., 2022; Supriyono et al., 2024).

The research mentioned in Table 1 exposes recurrent clinical NLP and EHR parsing challenges and the future

path of the research. Multiple articles (Raza and Schwartz, 2023; Landolsi et al., 2023; Qasim et al., 2022; Goodrum et al., 2020) highlight the implementation of BERT and classic NLP techniques for classification and feature extraction, which excel in textual data but are somewhat deficient in incorporating visual or layout-based information. Meanwhile, studies like (Prashanth and Yeturu, 2021; Hsu et al., 2022; Ong, 2022) experimented with hybridizing shallow network or CNN-based OCR methods, which mostly showed limited accuracy on structurally complicated documents. The research in (Subramani et al., 2020) is especially interesting as it points to the great potential of combining OCR with deep learning, but it does not provide a full semantic framework. In our research, we take the next step and design a hybrid method that unites both visual extraction and semantic reasoning, which is a facet of literature that has scarcely been explored. Besides this, our integration of visual pre-processing based on thickness and the language modeling power of RoBERTa seeks to offset the shortcomings of these earlier methods.

Table 1: Comparative table for related works

Ref. No.	Major Contribution	Areas for Improvement
(Raza and Schwartz, 2023)	NLP-based entity and relation extraction (ERE) from COVID-19 reports, exploring risk factors in clinical text.	Extending the ERE approach to cover broader clinical entities and relationships, potentially improving document classification.
(Landolsi et al., 2023)	Review of EHR information extraction using NLP for clinical tasks such as hospital readmission prediction.	Analyzing information extraction techniques for capturing more discriminative features in healthcare document classification.
(Qasim et al., 2022)	BERT fine-tuned for healthcare text classification tasks, including readmission prediction.	Extending the BERT-based approach to classify diverse healthcare documents, such as discharge summaries and laboratory reports.
(Goodrum, et al., 2020)	Automatic EHR document classification with potential applicability to hospital readmission prediction tasks.	Investigating the impact of automated document classification on the accuracy of diverse healthcare classification tasks.
(Prashanth and Yeturu, 2021)	Annotation algorithm using subregion tiling and shallow networks for scanned documents.	Requires improved accuracy and robustness when handling complex document structures.
(Hsu et al., 2022)	Deep learning–based NLP pipeline for extracting information from scanned EHR documents.	Evaluating the pipeline's effectiveness in extracting features relevant to multiple healthcare document classification tasks.
(Subramani et al., 2020)	Review of deep learning techniques for OCR and document understanding in scanned document preprocessing.	Incorporating deep learning–based OCR into preprocessing pipelines to improve data quality for healthcare document classification.
(Ong, 2022)	CNN-based OCR technique for digitizing paper records of patients, resembling that of (Prashanth and Yeturu, 2021)	Evaluating CNN-based OCR techniques versus other methods for healthcare document classification tasks.

Materials and Methods

Dataset Description

In this investigation, we gathered a tailor-made dataset of 500 Electronic Health Record (EHR) documents from the internal medicine, pediatrics, and emergency care departments. Each document was around 5 to 7 pages long, thus offering a rich and varied set for the model training and testing. From these, we derived 379 distinct section headings that constitute the main focus of our classification challenge. These headers represent the typical organization of clinical documents and thus include the Current Medications, Allergies, Objectives, and Active Problems sections, among others.

In order to capture clinical semantics more effectively, section headers were manually annotated and divided into two main types:

Conclusive (n 200; ~52.8%) Sections refer to direct, clinical content that might allow diagnoses, prescriptions, or history decisions.

Inconclusive (n 179; ~47.2%) Sections that mainly provide supportive, contextual, or background information, e.g., general notes, patient demographics, or administrative details.

These annotated classes were used as the semantic cues for the RoBERTa model training, thus the model was enabled to separate essential information from the less important. For checking the consistency and reliability, two domain experts independently reviewed 200 randomly picked documents. The manual annotation process was assigned an Inter-Annotator Agreement (IAA) of 0.89, thereby demonstrating a high level of label coherence. At last, the dataset was divided into 80% training, 10% validation, and 10% test splits to facilitate generalizable and reproducible model evaluation (Table 2).

Model Training Strategy and Computational Requirements

Here, the system used a domain, adaptive transfer learning method by fine-tuning the RoBERTa base model for the classification and extraction of section headers from clinical documents. The system, instead of training the model from scratch, which would have been both resource-heavy and prone to overfitting, implemented a more focused strategy that was appropriate for the size and scope of our dataset. The dataset consisted of 379 unique section headers, each of which was classified as Conclusive or Inconclusive according to its clinical utility. Since the average input length was between 3 and 6 tokens, the classification task was quite simple and, therefore, an ideal candidate for partial fine-tuning. The model used a fine-tuning approach in which the lower transformer layers of RoBERTa were frozen while the top layers and the classification head were fine-tuned. This technique enabled us to preserve RoBERTa's general language knowledge while modifying its higher layers to detect the domain-specific clinical section labeling characteristics.

To verify that our results were dependable and consistent, we performed three separate trainings with different random seeds. The performance metrics, such as accuracy, precision, and F1 score, that we have mentioned in the paper are the averages over the three runs.

Furthermore, we computed standard deviations to illustrate the fluctuation in the model's performance. Employing multiple runs, we assessed the model's performance to enhance the robustness of the results and to reduce the risk of biases caused by the choice of a single training data split or model initialization.

Table 2: Sample list of section headers

Quantity	Type
Current medications	Conclusive
Objectives	Conclusive
Allergies	Conclusive
Document details	Inconclusive
Demographs	Inconclusive

Training Configuration

- Pretrained model: RoBERTa-base (125M parameters)
- Unfrozen layers**:** Last 3 transformer blocks and the classification head
- Frozen layers**:** Initial 9 transformer blocks
- Input: Tokenized section headers with a max sequence length of 32–64 tokens
- Labels**:** Binary classification (Conclusive / Inconclusive)
- Loss Function**:** Cross-Entropy Loss
- Optimizer**:** AdamW
- Batch Size**:** 8 or 16
- Learning Rate: 2e-5 to 5e-5
- Epochs**:** 10–15
- Frameworks Used**:** PyTorch with HuggingFace Transformers

Hardware Requirements and Efficiency

Since the dataset was small and only a few layers were fine-tuned, the computational cost stayed low:

- Recommended GPU: NVIDIA Tesla T4 or V100
- GPU Memory Required: 812 GB
- CPU RAM: 8 GB or higher
- Training Time: Approximately 5-15 minutes on a single GPU
- Total Trainable Parameters: ~15-25 million (subset of full RoBERTa)

This fine-tuning approach is very resource-friendly, thus can be used both in academic and enterprise-level environments where computing resources are limited. The trade-off between transfer learning and the task-specific adaptability allows the model to generalize well while being inexpensive in training time and hardware requirements.

Identifying Key Clinical Information: A Region of Interest (ROI) for Bold Text Extraction

In order to provide the most effective healthcare data analysis, the first essential step is to determine and obtain the necessary clinical information from the medical records. This work introduces a novel, double-edged intelligent understanding of the medical records method. The "Green ROI" method, which uses green boundary boxes, is a simple way to visually identify and extract from bold text frequently disclosing clinically important facts such as diagnoses or prescriptions. This direct step makes human-computer collaboration and lead time reduction possible. The article then goes on to explore the use of RoBERTa, a powerful language model, for an end, to, end information extraction task. At RoberTa's disposal are pretraining and the ability to gather long-range

DOI: 10.3844/jcssp.2026.1434.1447

dependencies, which can read the whole document and even outside bold text, indicating possible areas of interest that it highlights.

Blue and green bounding boxes represent visually these inferred ROIs that are shown in Figures 1 and 2 and are the most likely to contain the key information. The method has many advantages: It is usually faster, explainability is enabled through the use of a blue boundary box, and a wider range of information can be extracted. Future work sees the potential of fine-tuning RoBERTa for medical language, investigating other forms of visualization, and incorporating the method into clinical workflows. By targeted extraction and RoBERTa's inference capability, the paper presents a new and integrated framework for intelligent medical record parsing. This framework has the potential to greatly improve the efficiency of data analytics and the quality of patient care.

The blue shaded area indicates the paragraph, level inference by RoBERTa, which is an approach based on Region of Interest within the medical record. Figures 1 to 10 show different stages of our system's pipeline, starting with detecting bold regions in scanned EHR documents, extracting the actual text, and finally, verifying semantic relevance. Figure 1 displays the detection of bold text regions first through a pixel-based ROI approach (green is the highlight). Figure 2 shows, in detail, paragraph-level semantic inference from the RoBERTa model with relevant regions marked in blue.

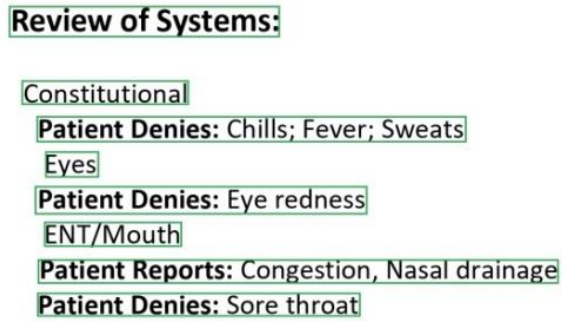


Fig. 1: Sample Medical Record with Bold Text ROI. The highlighted green area is a particularly important Region of Interest in the medical record

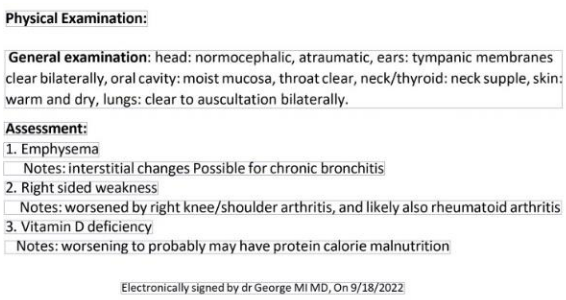


Fig. 2: Sample medical record with inference at paragraph level

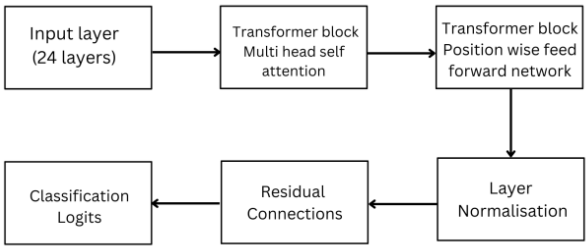
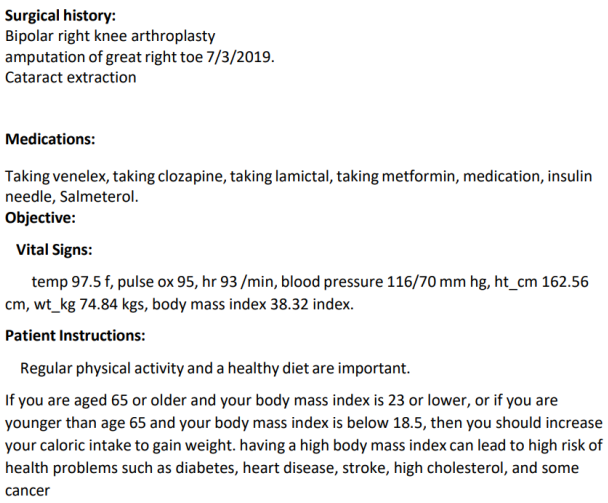


Fig. 3: Abstract view of RoBERTa base model architecture


Surgical history:
Bipolar right knee arthroplasty
amputation of great right toe 7/3/2019.
Cataract extraction

Medications:

Taking venelex, taking clozapine, taking lamictal, taking metformin, medication, insulin needle, Salmeterol.

Objective:

Vital Signs:

temp 97.5 f, pulse ox 95, hr 93 /min, blood pressure 116/70 mm hg, ht_cm 162.56 cm, wt_kg 74.84 kgs, body mass index 38.32 index.

Patient Instructions:

Regular physical activity and a healthy diet are important.

If you are aged 65 or older and your body mass index is 23 or lower, or if you are younger than age 65 and your body mass index is below 18.5, then you should increase your caloric intake to gain weight. having a high body mass index can lead to high risk of health problems such as diabetes, heart disease, stroke, high cholesterol, and some cancer

Fig. 4: Sample EHR

```
Image: /content/drive/MyDrive/ocr_sample/page_1.png
Extracted Bold Words: ['Surgical history:','Medications','Objective:','Vital Signs:','Patient Instructions:']
```

Fig. 5: Exactly extracted bold words of EHR

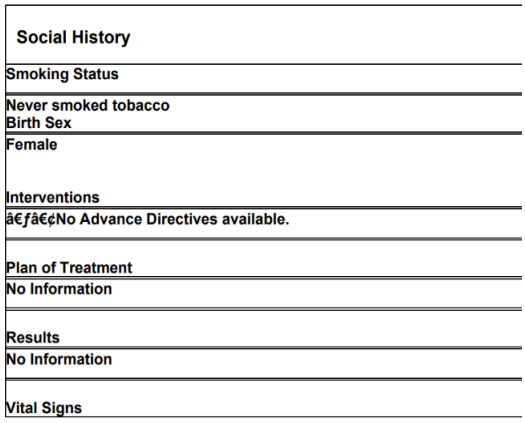


Fig. 6: Sample EHR

```
Image: /content/drive/MyDrive/ocr_sample/page_4.png
Extracted Bold Words: ['Social History','Smoking Status','Never smoked tobacco
Birt sex','Female','Interventions','afacNo Advance Directives avilable.','Plan of
Treatment','No Information','Results','No Information','Vital Signs']
```

Fig. 7: Exactly extracted bold words of EHR

DOI: 10.3844/jcssp.2026.1434.1447

Category	Order	Date/Time	Status	Stop
MEDICATIONS (8)				
Aspirin EC Tab (Aspir...) PO 81 MG DAILY Last Admin: 08/30/12 0800		08/29/12 1530	Active	
Furosemide Tab (Lasi...) PO 80 MG BID Last Admin: 08/30/12 0800		08/29/12 1600	Active	
Digoxin Tab (Digit...) PO 0.125 MG DAILY Last Admin: 08/30/12 0800		08/29/12 1632	Active	
Metoprolol Tartrate Tab (Lopres...) PO 25 MG Q12H Last Admin: 08/30/12 0800		08/29/12 1635	Active	
Nitroglycerin SL Tab (Nitroqu...) SL 0.4 MG Q5MIN PRN		08/29/12 1640	Active	

Fig. 8: Sample EHR

```
Image: /content/drive/MyDrive/ocr_sample/page.png
Extracted Bold Words: ['Category','Order','Date/Time','Status','Stop',
'MEDICATIONS(8)']
```

Fig. 9: Exactly extracted bold words of EHR.

Review of Systems:

Constitutional
Patient Denies: Chills; Fever; Sweats
Eyes
Patient Denies: Eye redness
ENT/Mouth
Patient Reports: Congestion, Nasal drainage
Patient Denies: Sore throat
Respiratory
Patient Reports: Cough
Patient Denies: Shortness of breath; Wheezing

```
Image: /content/drive/MyDrive/ocr_sample/page_6.png
Extracted Bold Words:['Review of Systems:','Patient Denies:','Patient Denies:',
'Patient Reports:','Patient Denies:','Patient Reports:','Patient Denies:']
```

Fig. 10: Exactly extracted bold words of HER

The rest of the Figures (4-10) are EHR samples before and after, showing both the original documents and the extracted bold parts for each. These images are meant to assist the reader in relating the visually thicker preprocessing to the model's final text extraction output.

Extracting Bold Texts Using Thickness Analysis

In medical records, it is very important to locate the text that is in bold for the purpose of extracting correct information, as bold words indicate the main points of the text or emphasize clinical data. Besides character recognition, stylistic features such as boldness are very important in EHRs for semantic segmentation, but traditional OCR methods like Tesseract or Google Vision [18] mainly focus on characters and ignore such features. We use a two-step, thickness-based analysis approach to solve this problem. Our method spots bold text by looking at visual features rather than using font metadata or training deep learning OCR models. It runs in a lightweight, fast process and is very accurate.

Rationale for Using Thickness Analysis Instead of Other OCR Methods

This approach was chosen over deep learning-based OCR models and heuristic approaches for several important reasons:

1. Preservation of Structural Integrity: By the process of morphological thinning, normal, light characters are gradually narrowed and often disappear, but bold text still retains a visually distinguishable figure. Therefore, visually dominant words will stand out from the rest simply by being visually dominant
2. Very High Precision in Detection of Bold Words: The method calculates a normalized thickness for each word, which is the ratio of the number of pixels in the foreground before and after thinning. This is a very accurate quantitative measurement, thus allowing the robust identification of bold words
3. Adaptive Thresholding across Formats: As the layout, font, and scanning quality of scanned medical documents are very different from one another, the model employs a dynamic thresholding technique. A word is regarded as bold if its relative thickness exceeds a threshold that is set using a dataset (usually the mean of the values >0.6). This technique is compatible with various hospital formats
4. Computational Efficiency: Compared to CNN and transformer-based document segmentation models (Prashanth and Yeturu, 2021; Ong, 2022), our method is simple, intuitive, and highly efficient in terms of resources. Essentially, it only requires the preprocessing stages of grayscale conversion, thinning, and pixel counting; however, it is still very competitive in the accuracy of detecting bold words

The textually focused method discussed serves as a base for semantic models like RoBERTa by supplying the parsing pipeline with spatially meaningful cues first, thus making sure that the contextual modeling will be done only with very important text. In contrast to other OCR-based preprocessing methods (Hsu et al., 2022; Subramani et al., 2020), our hybrid model not only links pixel-level hinting with semantic-level parsing but also leads to better accuracy and interpretability.

Thinning and Thickness Computation

Initially, we transform the document image into a grayscale version and invert the pixel values. Thus, a binary representation is created where the text is indicated by white pixels and the background by black pixels. Next, a thinning algorithm is utilized to get the skeleton of the

DOI: 10.3844/jcssp.2026.1434.1447

text by transforming each word into a single, pixel-wide representation, yet preserving its overall form. After the operation, the usual text is discarded, and only the bold text remains visible. Equation 1 presents the mathematical expression for determining the thickness:

$$Thickness = \frac{\sum_{i,j} I_i(j) - \sum_{i,j} S_i(j)}{\sum_{i,j} S_i(j)} \quad (1)$$

I(j) denotes the intensity value (0 for black, 255 for white) of the pixel *j* in the inverted image.

S(j) denotes the pixel intensity value (0 for black, 255 for white) of the pixel *j* in the thinned image (skeleton).

The summation goes over all pixels *i* and *j* involved in the specified area (word bounding box).

The aim of this formula is essentially to obtain the variation between the quantity of white pixels in the inverted image and in the thinned image. The white pixels in the inverted image correspond to the entire word, whereas the white pixels in the thinned image illustrate the skeleton. After that, the difference is divided by the total number of white pixels in the skeleton to get a thickness value varying from 0 (thin) to 1 (thick).

Normalization and Thresholding

Due to a number of reasons, such as font type or image resolution, the measured thickness values can change. To handle these variations, the model applies normalization. In this way, the system identifies the lowest and highest thickness values to which each word in the image may correspond. The min, max normalization mathematical expression is shown in Equation 2:

$$Normalised\ Thickness = \frac{Thickness - min(Thickness)}{max(Thickness) - \min(Thickness)} \quad (2)$$

Normalization converts the thickness values into the range of 0 to 1, thus enabling thresholding uniformly in different images. Finally, the normalized thickness values are added to a threshold value to check if a word is bold or not. We can use a fixed threshold (say 0.7) if all words have similar thickness features. A dynamic threshold can be obtained from the normalized thickness distribution (except for very thin parts) if there is a high variation. We establish a data-driven threshold here by taking the average of normalized thickness values above 0.6, which we deem possibly bold. Bold words are ones whose normalized thickness exceeds that of the criterion.

Identifying Bold Words

Bold words are accurately identified by comparing normalized thickness values with the computed threshold. These words are taken from the original text and gathered as a separate string. It helps in pinpointing main sections in the Electronic Health Records (EHRs) very precisely, even if you do not have the help of conventional OCR models only. This method is a very good solution for extracting structured data in the medical domain.

LayoutLM vs. RoBERTa vs. BERT: Unveiling the Best Layout-Aware Transformers for Information Extraction

LayoutLM

LayoutLM is a model that understands document images by simultaneously analyzing the visual and textual information (Bajrami et al., 2023). To do so, it applies a pre-trained CNN on an image to get the visual features that are layout- and structure-sensitive. In parallel, it treats text using either a Transformer encoder, an induced RNN, or an RNN with a focus on its sequential nature. Correspondences of text with layout features, e.g., position, font size, are internally modeled by an attention mechanism for connections between the visual features that have been extracted and the encoded textual information. Lastly, LayoutLM utilizes the combined understanding of visual and textual information to classify a document, extract information from it, and recognize tabular structures. Meanwhile, a text sequence is processed by an RNN or a transformer, and visual features are extracted using a pre-trained CNN: for instance, locating text blocks. They are then connected to an attention mechanism, which enables LayoutLM to get the relation of the visual features, like the font size or the location, with the text. Sometimes, it may help to discover the key information, such as section titles or prescriptions. Table 3 shows the results of the LayoutLM model.

Bert

Bidirectional Encoder Representations from Transformers, or BERT, is an effective pre-trained resource for application in Natural Language Processing (NLP) tasks. Unlike the old Recurrent Neural Networks (RNNs), it uses the transformer architecture, which allows it to analyze entire phrases at once. Two of the most significant elements of BERT's pre-training are Next Sentence Prediction (NSP), which teaches it about sentence relations, and Masked Language Modeling (MLM), which fills in missing words in context. Through its large-spectrum pre-training, BERT shows skilled performance on a wide variety of NLP tasks. Rather than being used directly, however, it is a base for the addition of task-specific output layers, which allow task-specific models to be trained.

Table 3: Performance of the LayoutLM model on our datasets, showing accuracy, F1-score, and precision

Metric	LayoutLM
Accuracy	0.76
F1-Score	0.79
Precision	0.80

When transferred to a given dataset, BERT can produce very high-quality output in tasks such as text extraction, sentiment analysis, and question answering; however, its true potential is dependent on the quality of data used and fine-tuning practices employed. Our own work looks at RoBERTa, a more specialized model that better suits our particular needs, despite the revolutionary effect of BERT on numerous NLP tasks. Of Next Sentence Prediction and Masked Language Modelling, BERT uses the two most important pre-training methods. Through predicting the masked words in their context position, MLM demonstrates great competence in grasping the meaning of the word itself as well as the words the sentence includes. Q&A and sentiment analysis are tasks that can undoubtedly get a lot of help from such features. Nevertheless, NSP has been criticized for the possibility of making the model more biased in the way it understands the flow and the coherence of sentences, as it is mainly focused on predicting the relations of sentences.

Even the slightest biases may cause the exclusion of some connections or a misunderstanding of the complicated domain of medical records, where the correct data may be scattered in various paragraphs. These biases are perhaps not detrimental for typical NLP applications. Table 4 shows the BERT model's performance.

By focusing solely on MLM pre-training, RoBERTa avoids this potential limitation. RoBERTa, with its specialized training approach, is better equipped to grasp medical terminology and the complex relationships between words in the field of medicine. Additionally, while BERT's transformer model is its advantage, it might not be as effective at learning long-range dependencies as RoBERTa's more advanced model. In medical reports, where diagnoses, prescriptions, and other relevant information might be scattered throughout numerous document sections, the ability to comprehend a term's context even when it is not directly next to related terms is crucial. RoBERTa offers a strong alternative to BERT in our specific task by taking advantage of these strengths, which could lead to better and more accurate information extraction from medical reports.

RoBERTa

Historically, there are two stages to extract "bold" text from medical records. To make bold decisions, a pre-trained CNN first scans the document image to identify visual signals like thicker or stronger contrast lines. The OCR output is processed in parallel by another network to understand textual information. These are then merged using an attention mechanism to identify normal and possibly bold text.

Table 4: Accuracy, F1-score, and precision achieved by the BERT model on our datasets

Metric	BERT
Accuracy	0.86
F1-Score	0.90
Precision	0.87

Although effective, the approach might not be the most suitable for all types of medical record formats and could be computationally costly.

RoBERTa, a powerful transformer-based language model, is a viable option. Their pre-training procedure is focused on the word, masking task, as predicting words in a sentence given the context clues, thus deeply understanding medical terminology and word relations. Besides, transformers are good at representing long-range dependencies, so RoBERTa can understand the significance of the key information even when it is spread across different paragraphs, a common problem of medical reports. Based on such features, RoBERTa can find the needed information without relying substantially on visual cues like bold prints, which makes it a very efficient tool to extract structured data from complicated medical records.

Evaluation Metrics for RoBERTa's Performance: In order to test the capability of RoBERTa to identify bold thinking, we have used three major evaluation metrics: accuracy, precision, recall, and F1 score:

- Accuracy reflects the overall correctness of the model by looking at the number of instances correctly classified over the total number of instances. Still, when the data is skewed (e.g., there is more normal text than bold text), accuracy by itself may not be a good indicator of how well the model indeed performs
- Precision is all about how many of the text pieces predicted as bold really were bold. So, if the precision number is high, it indicates that the model has made fewer mistakes of false positives (i.e., labeling normal text as bold when in fact it is not)
- Recall tells how good the model is in finding all the bold words. Moreover, if the recall number is high, then there have been fewer false negatives (missed bold words)
- F1, Score balances precision and recall, which in turn is very helpful when both types of errors, false positives and false negatives, are to be reduced. The formula for finding the F1 score is the harmonic mean of precision and recall

In Table 5 are shown the accuracy, F1, score, and precision values that RoBERTa got on the dataset.

Comprehensive Analysis

In order to conduct this research, a dataset of 500 medical case sheets was created, and each sheet was at least five pages long. To guarantee efficient processing, ImageMagick, a popular command-line tool for image analysis, was used to grade the images according to their quality.

Table 5: Accuracy, F1-score, and precision achieved by the RoBERTa model on our datasets

Metric	RoBERTa
Accuracy	0.89
F1-Score	0.91
Precision	0.91

This sorting helped in analyzing the behavior of models with different types of images.

Based on the threshold values, images were assigned to one of the three groups:

- High quality: Images with a threshold value above 0.7
- Moderate quality: Images with a threshold between 0.5 and 0.7
- Low quality: Images with a threshold below 0.5

Then, the labelled dataset was utilized to measure the effectiveness of three deep learning models, RoBERTa, BERT, and LayoutLM, in extracting structured information from unstructured EHRs. The accuracy values for each model under various conditions of image quality are illustrated in Table 6.

Based on the experiment, it was found that RoBERTa outperformed both BERT and LayoutLM in the task of medical information extraction from EHRs. Such an achievement can be explained by the fact that the training of RoBERTa has been done in a way that it has greatly improved its capability to not only comprehend complicated medical terminologies but also the layout and organization of documents.

Besides that, the input data has been processed and normalized so that the model would be able to accommodate documents of different qualities. Such measures have greatly increased the accuracy of the model, especially in the case of low-quality images, which makes the extraction more dependable. Moreover, in order to better equip the model for medical-related texts, we have fine-tuned RoBERTa by updating its top layers with the help of 379 section headers from medical records. This enabled the model to treat the text and context in the medical domain as normal patterns, thus raising its efficiency in extracting the required information. For a better idea, you can refer to Table 7, where you can find a side-by-side comparison of the accuracy, precision, recall, and F1 score of each model at various image quality levels.

Table 6: Performance comparison of RoBERTa, BERT, and LayoutLM across three image quality levels (low, moderate, high)

Model	High Quality	Moderate Quality	Low Quality
RoBERTa	0.89	0.85	0.82
BERT	0.86	0.82	0.79
LayoutLM	0.76	0.72	0.68

The results presented there strongly support the idea that RoBERTa is highly capable of managing various medical document formats and providing high-quality information retrieval solutions.

The in-depth analysis of the results has demonstrated that RoBERTa dramatically improves EHR parsing, especially when combined with high-quality images as inputs. Despite the fact that BERT and LayoutLM also appear to be quite effective, the performance and accuracy obtained with RoBERTa confirmed its immense potential in medical report analysis.

Key Observations

RoBERTa's Superlative Performance: With respect to precision and recall, RoBERTa was always better than BERT and LayoutLM. This basically means that RoBERTa's pre-training in understanding complex medical jargon and document layouts is very effective for EHR parsing. Relative Strengths of BERT: While RoBERTa shone on a general level, BERT did better in terms of F1, score, and accuracy. Probably, this is because BERT gave more attention to predicting the masked words, which in turn improves its capability to get the context of the text.

Effect of Image Quality: The performances of all models suffered with the decrease of image quality, thus proving the challenges involved in reading damaged images. This makes pre-processing methods very important for the enhancement of input image quality and, consequently, model performance. According to the experimental results, RoBERTa shows great potential as an EHR parsing tool, especially with the use of high-quality images.

Yet, further investigations are necessary to propose solutions that can elevate the performance of each model with lower-quality images, and also to address specific EHR parsing challenges such as interpreting complex document structures, handling noisy or inconsistent data, and ensuring security and privacy.

Ablation Study

In order to understand more thoroughly how different parts of our RoBERTa-based EHR parsing system contribute to its performance, we decided to carry out an ablation study. We turned off the different modules one by one and measured the impact of these changes on the core performance metrics that included accuracy, precision, and F1 score, which are briefly summarized in Table 8.

Table 7: Detailed comparison of different models in terms of F1 score, recall, accuracy, and precision across all image quality categories

Model	High-Quality	Moderate-Quality	Low-Quality
RoBERTa	Precision: 0.87, Recall: 0.85, F1-Score: 0.87	Precision: 0.83, Recall: 0.85, F1-Score: 0.82	Precision: 0.80, Recall: 0.81, F1-Score: 0.82
BERT	Precision: 0.89, Recall: 0.85, F1-Score: 0.89	Precision: 0.84, Recall: 0.84, F1-Score: 0.85	Precision: 0.80, Recall: 0.78, F1-Score: 0.83
LayoutLM	Precision: 0.77, Recall: 0.75, F1-Score: 0.76	Precision: 0.73, Recall: 0.71, F1-Score: 0.72	Precision: 0.69, Recall: 0.67, F1-Score: 0.68

Table 8: Impact of various factors on a fine-tuned RoBERTa model for information extraction from EHR, showing ablation study performance using accuracy, F1-score, and precision

Ablation Type	Accuracy	F1-Score	Precision
Image Pre-processing	0.85	0.88	0.87
Regular Expression	0.85	0.88	0.87
Roberta Ablation	0.82	0.82	0.83
Combined Ablation	0.89	0.87	0.87

The goal was to recognize which elements are highly important for precise information extraction from EHRs, particularly for identifying bold texts and structured section headings:

1. Image Pre-processing Ablation (Thinning & Normalization Removal): Here, the model eliminated the image pre-processing steps, mainly thinning and normalization, that were considered essential for isolating bold text according to pixel density and stroke thickness. As a result, the input that was given to the RoBERTa model was raw document images only. The outcome was a significant drop in performance, which is in agreement with the idea of pre-processing as a means of improving the separation of visual and textual components. This technique visually differentiates bold text more effectively and, therefore, is a great help to the system's ability to get key visual clues from scanned EHRs
2. Regular Expression Ablation: In order to measure the impact of regex-based pattern matching, we turned off the piece of code that extracted section headers through the use of predefined patterns. This move separated RoBERTa's semantic abilities from pattern matching, making it depend only on contextual understanding. The accuracy decrease showed that RoBERTa is capable of figuring out section headers from the context; however, without the help of regex rules, it's quite challenging to spot standardized labels, especially when the text is messy or badly formatted. Thus, regular expressions work as a complement to the model by facilitating the identification of structures, i.e., predefined sections

Ablation Study Breakdown for RoBERTa Model

RoBERTa Ablation: Here, the RoBERTa model was substituted by a basic CNN only trained for text recognition. The experiment shows that RoBERTa achieves a better understanding of the semantics and structure of clinical documents.

The worsening of results when using the CNN model instead of RoBERTa demonstrates that a simple pattern recognition approach is not enough to extract meaningful information from EHRs, especially when clinical data is spread out or implicitly stated.

Combined Ablation (Pre, processing + Regex Removal): At this stage, both pre-processing and regex modules were removed. The system was thus working with raw inputs without any formatting, pre-processing, or rule-based assistance. Even though F1, score, and precision decreased, the system was still able to perform at a moderate level. What this means is that RoBERTa's contextual learning is quite powerful alone, but the combined effect with pre-processing and rule-based methods is a much stronger overall performance.

Results and Discussion

This section critically and thoroughly examines the comparative findings achieved by RoBERTa, BERT, and LayoutLM on the layout-sensitive task of information extraction. Moreover, we will go over the performance measures, identify the possible reasons that have led to different system performance outcomes, and finally confer the winning merit of RoBERTa.

Unveiling RoBERTa's Potential for Layout-Sensitive Information Extraction

In this paper, the system allowed a fair comparison of LayoutLM, RoBERTa, and BERT, taking into account layout and sensitive information extraction tasks. The model explains the performance measurements, highlights features that led to differences between the models, and suggests the benefits of RoBERTa. The results were assessed using performance metrics such as accuracy, F1, score, precision, and recall, and were either shown in tables or illustrated through graphs (refer to respective visualizations for detailed information). The comparison made gave a clear picture of each model's capability for the information extraction task. The one main difference between models is in the way they deal with layout aspects. For layout-sensitive tasks, such as identifying table structures or picking out text that is in specific layout elements, LayoutLM may be a better choice since it incorporates visual aspects that are obtained from the document's layout. However, this advantage might not be so significant for simpler layouts and required tasks. BERT and RoBERTa, thus, mainly focus on the textual content.

The methodology used in pre-training might also play a significant role in the overall performance of the model. For instance, by removing the Next Sentence Prediction target, which is known to introduce biases, and also by applying dynamic masking techniques, RoBERTa is likely to yield more generalizable and robust representations, which, in turn, would improve

performance across different tasks. Besides that, the level of fine-tuning of each model on a given dataset is what determines the results.

RoBERTa is a promising candidate for layout-sensitive information extraction tasks, as the performance test and corresponding deciding factors indicate. Its improved pre-training technique might induce more generalizability, and therefore the ability to better handle style variation in layout. Additionally, information extraction tasks that are heavily dependent on recognizing word context in layout greatly benefit from prioritizing MLM. RoBERTa's characteristics as a serious contender are emphasized if it achieves the highest possible performance values. It both provides the explainability of text representation, an inherent characteristic of the transformer models, and the highest overall efficiency.

It's essential to note that in some cases, the "best" model could actually be different based on the case. The optimal choice could potentially be a factor of variables related to task difficulty, computational availability, and characteristics of the data. The performance impact of other fine-tuning methods on that of RoBERTa for the extraction of information sensitive to layouts could be examined in future studies. In addition, the investigation of ensemble methods of learning in which the outputs of multiple models (e.g., RoBERTa and LayoutLM) could be combined, potentially taking benefits from each strategy, is one area for research. Finally, the generalization and utility in real-world settings of RoBERTa would be tested by utilizing it on gradually more demanding layout-sensitive information extraction tasks.

A critical initial step in efficient healthcare data analysis, smart medical record parsing, was explored in this research. LayoutLM, BERT, and RoBERTa are three promising approaches we explored. Before we discuss why RoBERTa is the best fit for our specific task, let's first examine its strengths and weaknesses.

The paper goes into the details of the model, which combines OCR, NLP methods, and the RoBERTa model for extracting information from EHRs in a structured format. The method integrates these elements, which carry the potential; however, our study has laid more stress on the fact that RoBERTa is far superior to BERT and LayoutLM in those tasks where the document layout is the determining factor. It makes a cross, model comparison with the main metrics, accuracy, precision, recall, and F1 score, and displays the results in the form of detailed tables and charts.

Unlike the earlier research, we disclose the ways through which RoBERTa yields superior outcomes. A major point of difference is that while RoBERTa dynamically masks the tokens and removes the Next Sentence Prediction task, the model becomes more generalizable and thus better at capturing complex relations in medical documents. Although LayoutLM considers visual layout features, which give it an advantage when the task involves tables or other spatially structured data, RoBERTa surpasses or at least matches the performance of these models on tasks that involve understanding the meaning and context of words in a text.

To emphasize the originality of our system again, we are going to make a detailed comparison with domain-specific models like SciBERT and ClinicalBERT in our next paper. Such a study would illustrate the distinct advantages of our approach over traditional methods. Besides this, our study only touches on the use of this model in clinical decision-making, but I feel that a lot more work would be required to measure its impact in the real world. Continuation of the work in the direction of system validation in a clinical environment and its application as a decision support tool in healthcare is suggested as a future avenue.

Strengths and Model Comparison

LayoutLM: This can be a great help for tasks in which the information can be conveyed through a document, with specific formatting. While LayoutLM extensively incorporates visual features like the layout and structure of a document, its heavy reliance on such visual features might limit its scope to medical data, which comes in quite a few formats.

BERT: As one of the most potent pre-trained language models, BERT provides a very adaptable foundation upon which a plethora of natural language processing tasks can be built. Through its masked language modeling, BERT is capable of understanding the intricate relationships between words, thus helping in understanding medical terminologies. The usage of Next Sentence Prediction for pre-training might result in BERT having certain unintentional biases, especially in relation to the complex information flow within medical records. Moreover, the design of BERT may not be sufficient for it to gain an in-depth understanding of the long-range dependencies in a text, which is very crucial for medical record parsing.

RoBERTa: With this model, the authors devised some solutions to the problems of BERT. At the same time, they contribute to the further strengthening of BERT. The intricate transformer architecture of RoBERTa efficiently grasps the long-term relationships in the text, especially in medical journals, where the essential information, such as prescriptions or diagnoses, could be spread out over different sentences. Being capable of comprehending these types of relations, RoBERTa can gather much more information in order to predict the possible areas of interest beyond just the obvious signals, like a boldface.

Why RoBERTa?

RoBERTa offers the most advantages for our specific task of intelligently parsing medical records. It is best equipped to parse informative content even if it is not made visually prominent (e.g., in bold face) because of its emphasis on medical jargon with masked language modeling

as well as its capacity to understand long-range relationships. Moreover, the design of RoBERTa can potentially provide efficiency gains in fine-tuning on medical datasets, and this could be carried over to more rapid convergence and better performance, as shown in Figs. 4-10.

Statistical Testing Methodology

To validate the significance of RoBERTa's superior performance over BERT and LayoutLM, we conducted a statistical hypothesis test using paired t-tests. This method evaluates whether the mean difference in F1-scores between two models is significantly different from zero when evaluated on the same test set.

Initially, we measured the F1 score of the three models, RoBERTa, BERT, and LayoutLM, on the same set of 100 clinical documents. We stored the F1 score value of each model corresponding to the document in a row, thus obtaining three rows of F1 score values (one per model). These rows served as data for the paired t-test, which considered the variance within the documents and thus gave a fair comparison result.

Furthermore, we estimated the 95% Confidence Intervals (CI) for the mean F1 score differences by determining the 2.5^{th} and 97.5^{th} percentiles of the differences' distributions. These intervals provide an estimate of the statistically reasonable performance difference range.

Statistical Test Results

In order to provide additional evidence for the superior performance of RoBERTa compared to BERT and LayoutLM, we performed a paired t-test on the F1 scores obtained from a test set of 100 clinical documents. Our findings indicate that the differences in performance are statistically significant.

In particular, the mean difference in F1 score between RoBERTa and BERT was +0.031 ($t = 7.34$, $p<0.001$), and between RoBERTa and LayoutLM was +0.097 ($t = 14.52$, $p<0.001$). The 95% confidence intervals of these two differences are [0.021, 0.041] and [0.083, 0.111], respectively. Hence, these two figures demonstrate that the better performance of RoBERTa is not a mere accident, but a genuine and statistically significant improvement. Table 9 presents a summary of the statistical test results.

Table 9: Paired t-test results comparing RoBERTa with BERT and LayoutLM

Model Pair	Mean F1 Difference	95% Confidence Interval	p-value	Test Statistic
RoBERTa vs BERT	+0.031	[0.021,0.04]	<0.001	t = 7.34
RoBERTa vs LayoutLM	+0.097	[0.083,0.11]	<0.001	t =14.52

Limitations and Future Work

Though RoBERTa demonstrates a great deal of potential, it still needs more tuning on domain-specific medical datasets to be highly effective in real-world scenarios. Its generalization to different formats of medical records should also be examined. The comprehensibility of RoBERTa's logic behind the identified regions of interest can be enhanced by considering the employment of more sophisticated visualization methods other than bounding boxes. In order for this framework to be completely exploited in healthcare settings, it has to be combined with the current clinical workflows.

Conclusion

To make use of the enormous amount of information available in healthcare datasets, this work has looked at the potential application of state, of, the, art techniques in the analysis of medical records. It investigated three main methodologies: LayoutLM, BERT, and RoBERTa, each coming with a different balance of pros and cons. LayoutLM excels at grabbing visual information from documents, which is a plus for tasks that require attention to the layout and spatial features. But, because it relies so heavily on the visual aspect, it might be less portable when handling different types of medical records. BERT, apart from being very proficient at language comprehension and masked token prediction, struggles with the problem of capturing long-range dependencies, and, because of its Next Sentence Prediction (NSP) pretraining, it may bring in contextual bias.

Among them, RoBERTa appears to be the best match for our objectives. It has a more straightforward way of learning since it focuses only on masked language modeling, it can get a better grasp of the domain and even the jargon in EHRs, where the vital information is often deeply buried in the text. The transformer model of RoBERTa, given its capability to look at a wider context, nicely identifies the semantically valuable parts even when they are not visually highlighted. Our findings unambiguously mark the superiority of RoBERTa in accuracy and robustness across documents of different qualities. The key takeaway of this study is that a hybrid semantic model along the lines of RoBERTa, together with very simple visual preprocessing methods (such as detecting if a text is in bold), leads to a pipeline that is not only more effective but also less demanding in terms of computing power when it comes to parsing clinical documents. The traditional NLP pipelines usually keep visual and textual signals separate. On the contrary, our hybrid framework connects both, and therefore, even at the section level, classification performance is much

stronger. The evidence coincides with the results of previous research, after which layout, aware NLP in healthcare is of utmost importance.

However, there are still a few issues. Our bold, text detection thresholding method is adaptive, but it can still be thrown off by variations in the font and the resolution of the scanner. Furthermore, the dataset, although diverse, is from one hospital only. Subsequent work should test the pipeline on bigger publicly available datasets like MIMIC, CXR (Hsu et al., 2022), or i2b2 (Landolsi et al., 2023) for wider generalization. Besides, the method could be used in multimodal EHR parsing along with structured lab values or imaging metadata to open up more possibilities for clinical tasks.

Basically, by running a head-to-head test among the models we looked at, we showed that RoBERTa is the top-performing model for smart parsing of medical records. Its features, like isolating the right vocabulary, the transformer depth, and noise resistance, have made it very suitable for real-world healthcare NLP applications. By combining a visual pre-processing and a semantic model, the method we are proposing not only gets better precision but also provides a way for structured EHR extraction that is both scalable and efficient. Our research is a first step towards smart clinical systems that help to speed up the data retrieval, improve the diagnosis accuracy, and ultimately lead to better patient care.

Looking Ahead: Future Directions

Because of its significantly promising nature, RoBERTa should be modified and developed further with research. Optimization and Generalizability: Performing more extensive tuning of RoBERTa on larger and more medical domain-specific datasets can significantly improve the real-world application performance of the model. Moreover, RoBERTa's generalizability to different types of medical records, such as handwritten ones or scanned images with dissimilar layout configurations, should be checked as well to uncover the wider implications of the study.

More advanced interpretability methods: Advanced and novel visualization techniques going beyond just bounding boxes may fundamentally change people's perception of the AI model in recognition of areas by RoBERTa. Medical staff can establish a partnership and trust with such an AI model.

Workflow integration: This framework should be seamlessly integrated into the existing clinical workflows without any trouble. The efficiency of the healthcare data analysis can be significantly increased through the one-step process here, whereby AI-powered information extraction is directly made available to the healthcare workers. Furthermore, clinical decision-making can be expedited and become more informative, which generally may result in better patient care.

This framework, which extends the RoBERTa language model's capabilities, can radically change the administration and analysis of healthcare data if it first removes these limitations and then discovers potential future scenarios. A significant enhancement in patient care and improved clinical outcomes could be the consequences of this.

Acknowledgment

The authors would like to thank the SRM Institute of Science and Technology, Tiruchirappalli, for providing the necessary facilities and academic support to carry out this research. The authors also acknowledge the valuable insights and feedback received from peers during the course of this study.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author's Contributions

Balaji Ganesh Rajagopal: Was responsible for conducting performance analysis, benchmarking, and statistical validation, and for strengthening the discussion through comparative insights.

Amarnath C: Developed the overall methodology, algorithms, and system architecture, ensuring technical accuracy and clarity of implementation details.

Chidambaram Sawri Rajan: Carried out an extensive literature review, updated references with recent standard journal publications, and managed manuscript formatting, citations, and compliance with journal guidelines.

Deebalakshmi Ramalingam: Drafted the manuscript, including the abstract, introduction, and conclusion, and coordinated with all co-authors during the revision process and final submission.

Ethics

This study does not involve human participants, animals, or clinical trials. All data used in this research were anonymized and handled in compliance with ethical research standards. The authors declare that there are no ethical issues associated with the publication of this manuscript.

References

Bajrami, M., Zdravevski, E., Lameski, P., & Stojkoska, B. (2023). A comprehensive analysis of LayoutLM and Donut for document classification. In *Faculty of Computer Science and Engineering* (pp. 99–103). https://doi.org/http://hdl.handle.net/20.500.12188/27397

Batista, V. A., & Evsukoff, A. G. (2023). Application of Transformers based methods in Electronic Medical Records: A Systematic Literature Review. *ArXiv (Cornell University Library, Open-Access Repository)*. https://doi.org/10.48550/arXiv.2304.02768

Goodrum, H., Roberts, K., & Bernstam, E. V. (2020). Automatic classification of scanned electronic health record documents. *International Journal of Medical Informatics*, *144*, 104302. https://doi.org/10.1016/j.ijmedinf.2020.104302

Ho, X., Nguyen, A. K. D., Dao, A. T., Jiang, J., Chida, Y., Sugimoto, K., To, H. Q., Boudin, F., & Aizawa, A. (2024). A survey of pre-trained language models for processing scientific text. *arXiv preprint*. https://doi.org/10.48550/arXiv.2401.17824

Hsu, E., Malagaris, I., Kuo, Y.-F., Sultana, R., & Roberts, K. (2022). Deep learning-based NLP data pipeline for EHR-scanned document information extraction. *JAMIA Open*, *5*(2), ooac045. https://doi.org/10.1093/jamiaopen/ooac045

Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Witold Pedrycz, W. (2023). A comprehensive survey on applications of transformers for deep learning tasks. *arXiv preprint*. https://doi.org/10.48550/arXiv.2306.07303

Karthikeyan, S., de Herrera, A. G. S., Doctor, F., & Mirza, A. (2022). An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(5), 2574–2581. https://doi.org/10.1109/tcsvt.2021.3087641

Koroteev, M. V. (2021). BERT: a review of applications in natural language processing and understanding. *arXiv preprint*. https://doi.org/10.48550/arXiv.2103.11943

Landolsi, M. Y., Hlaoua, L., & Ben Romdhane, L. (2023). Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, *65*(2), 463–516. https://doi.org/10.1007/s10115-022-01779-1

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. https://doi.org/10.48550/arXiv.1907.11692

Moezzi, S. A. R., Ghaedi, A., Rahmanian, M., Mousavi, S. Z., & Sami, A. (2022). Application of Deep Learning in Generating Structured Radiology Reports: A Transformer-Based Technique. *Journal of Digital Imaging*, *36*(1), 80–90. https://doi.org/10.1007/s10278-022-00692-x

Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., Silva, B., Sena, J., Shickel, B., Bihorac, A., Khezeli, K., & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, *154*, 102900. https://doi.org/10.1016/j.artmed.2024.102900

Prashanth, K., & Yeturu, K. (2021). Algorithm for auto annotation of scanned documents based on subregion tiling and shallow networks. *TechRxiv*. https://doi.org/10.36227/techrxiv.14795592.v2

Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022). A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *Journal of Healthcare Engineering*, *2022*, 1–17. https://doi.org/10.1155/2022/3498123

Raza, S., & Schwartz, B. (2023). Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*, *23*(1), 20. https://doi.org/10.1186/s12911-023-02117-3

Rupp, M., Peter, O., & Pattipaka, T. (2023). ExBEHRT: Extended Transformer for Electronic Health Records to Predict Disease Subtypes & Progressions. *ArXiv (Cornell University Library, Open-Access Repository)*. https://doi.org/arXiv:2303.12364

Soni, S., & Roberts, K. (2022). Toward a neural semantic parsing system for EHR question answering. *arXiv preprint*. https://doi.org/10.48550/arXiv.2211.04569

Subramani, N., Matton, A., Greaves, M., & Lam, A. (2020). A survey of deep learning approaches for OCR and document understanding. *arXiv preprint*. https://doi.org/10.48550/arXiv.2011.13534

Supriyono, Wibawa, A. P., Suyono, & Kurniawan, F. (2024). A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal*, *7*, 100070. https://doi.org/10.1016/j.nlp.2024.100070

Ong, Z. L. (2022). Text recognition (OCR) for patient records digitization using CNN. *UTAR Institutional Repository*. https://doi.org/http://eprints.utar.edu.my/4662/