Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023

IS in Healthcare Addressing the needs of post-pandemic digital healthcare

Dec 11th, 12:00 AM

A Vietnamese Handwritten Text Recognition Pipeline for Tetanus Medical Records

Minh N. Dinh
RMIT University, minh.dinh4@rmit.edu.vn

Mau Toan Le
Hospital for Tropical Diseases, drmautoan@gmail.com

Triet Bui
RMIT University, s3694551@rmit.edu.vn

Minh Mai
RMIT University, s3681447@rmit.edu.vn

Long Tran
RMIT University, s3755614@rmit.edu.vn

See next page for additional authors

Recommended Citation

Dinh, Minh N.; Le, Mau Toan; Bui, Triet; Mai, Minh; Tran, Long; Nguyen, Nhan; and Vo, Tan Hoang, "A Vietnamese Handwritten Text Recognition Pipeline for Tetanus Medical Records" (2023). *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023*. 2.
https://aisel.aisnet.org/icis2023/ishealthcare/ishealthcare/2

Presenter Information

Minh N. Dinh, Mau Toan Le, Triet Bui, Minh Mai, Long Tran, Nhan Nguyen, and Tan Hoang Vo

A Vietnamese Handwritten Text Recognition Pipeline for Tetanus Medical Records

Completed Research Paper

Minh Ngoc Dinh
RMIT University
Ho Chi Minh City, Vietnam
minh.dinh4@rmit.edu.vn

Mau Toan Le
Hospital for Tropical Diseases
Ho Chi Minh City, Vietnam
drmautoan@gmail.com

Triet Bui
RMIT University
Ho Chi Minh City, Vietnam
s3694551@rmit.edu.vn

Minh Duc Mai
RMIT University
Ho Chi Minh City, Vietnam
s3681447@rmit.edu.vn

Long Tran
RMIT University
Ho Chi Minh City, Vietnam
s3755614@rmit.edu.vn

Nhan Nguyen
RMIT University
Ho Chi Minh City, Vietnam
s3687637@rmit.edu.vn

Tan Hoang Vo
Oxford University Clinical Research Unit
Ho Chi Minh City, Vietnam
hoangvt@oucru.org

Louise Thwaites
Oxford University Clinical Research Unit
Ho Chi Minh City, Vietnam
lthwaites@oucru.org

Hai Ho Bich
Oxford University Clinical Research Unit
Ho Chi Minh City, Vietnam
haihb@oucru.org

Abstract

Machine learning techniques are successful for optical character recognition tasks, especially in recognizing handwriting. However, recognizing Vietnamese handwriting is challenging with the presence of extra six distinctive tonal symbols and vowels. Such a challenge is amplified given the handwriting of health workers in an emergency care setting, where staff is under constant pressure to record the well-being of patients. In this study, we aim to digitize the handwriting of Vietnamese health workers. We develop a complete handwritten text recognition pipeline that receives scanned documents, detects, and enhances the handwriting text areas of interest, transcribes the images into computer text, and finally auto-corrects invalid words and terms to achieve high accuracy. From experiments with medical documents written by 30 doctors and nurses from the Tetanus Emergency Care unit at the Hospital for Tropical Diseases, we obtain promising results of 2% and 12% for Character Error Rate and Word Error Rate, respectively.

Keywords: Doctor handwriting, Hand-written recognition, Deep learning, Image processing

Introduction

As a developing country with a population of approximately 100 million, Vietnam has only focused on digital transformation in recent years. As a result, Vietnam has a significant number of physical documents, that need to be digitized. Furthermore, the Vietnamese government has authorized The National Digital Transformation Programme by 2025 to aid in accelerating the digitization of organizations, firms, and households in order to enhance their commercial goods, competitiveness, and efficiency (Samuel, 2021). In the context of public healthcare, Circular 46/2018/TT-BYT (Thu Vien Phap Luat, 2018) clearly states that by the end of 2023, first-class medical examination and treatment facilities nationwide will complete the application of electronic medical records (EMR). However, due to the epidemic and many other difficulties, by 2023, only 37/135 1st grade hospitals will complete the above target, while 1,400 other hospitals have only applied low-level informatics.

Our study shows that hospital digitalization is a complex process. Studies in this area have suggested that the process of building an Information System (IS) for medical records needs to go through 5 levels.

1. Automated Medical Record (AMR): Still employing traditional paper-based medical records with some information managed by computer systems.
2. Computerized Medical Record (CMR): Most data, information, and records are stored on computers.
3. Electronic Medical Record (EMR): Migrating the paper medical records to suitable storage and enabling data and information lookup by computers.
4. Electronic Patient Record (EPR): Highly manageable compared to EMR, which supports cross-referencing of patient records between hospitals.
5. Electronic Health Record (EHR): An individual's electronic health record from birth to death, with links to all relevant personal information.

Delivering a Level 3 IS for healthcare in Vietnam is challenging because (1) healthcare providers often rely on handwritten notes to document patient information, and (2) the diversity in medical contexts, even just within a specialized hospital such as the Hospital for Tropical Diseases (HTD - https://www.bvbnd.vn/). Over 200 cases of tetanus were reported annually by the HTD in the Southern Regions of Vietnam since 1997. While current preventative measures are effective, therapy still needs constant observation for up to four weeks to guarantee the complete recovery of patients. The time-consuming documentation process, besides the medical examination and treatment, results in doctors and nurses having to hand-write medical records, sheets, and books. To make the matter worse, doctor’s handwritings, not just in Vietnam, are quite unique due to the specific features of the technical language used in healthcare system.

In this work, we develop an image processing pipeline, namely DOCR, which takes an image of a hard-copy medical file, especially for Tetanus cases, and generates an electronic document. On a theoretical level, this work draws advances from the fields of image processing, pattern recognition, and machine learning. The challenges of deciphering varying handwriting styles, coping with noisy and degraded inputs, and handling contextual understanding, emphasize the theoretical underpinnings of feature extraction, sequence modeling, and linguistic analysis. The practicality of this work lies at the automation of digitizing the doctor handwritten documents without intruding the conventional medical recording process (e.g., handwriting the medical details during routine patient examinations) at Vietnam hospitals. Furthermore, the technology can reduce the risk of losing important medical information, while helping doctors to extract valuable data from handwritten medical records, which can be used to identify trends, inform treatment protocols, and advance clinical process.

The complete end-to-end pipeline employs:

- A computer-vision based document-layout analysis component to identify the text area in a scanned document and partitions each image into distinct text regions of interest.
- An adaptive-preprocessing component to address imaging issues such as blurry handwriting strokes, and noises generated as part of the scanning step or occurred because of paper degradation.
- A handwritten text recognition (HTR) module to detect and transcribe all handwriting text presented in ROIs. Specifically, this module performs segmentation to the word- and the character-level using a hybrid VGG19–Transformer model to convert the image into electronic text.

From the experiments with the records written by 30 doctors, we have obtained encouraging results of 2% and 12% of Character Error Rate (CER) and Word Error Rate (WER), respectively.

The rest of the paper is structured as follows. Section 2 presents the challenges in digitizing medical handwritten documents and compares related works. Section 3 presents our DOCR pipeline consisting of text and image processing components. We describe the experimental implementation and results in section 4. Section 5 concludes the paper with directions for future work.

Background

Optical Character Recognition (OCR) is a technique used to obtain textual information from scanned documents directly without any manual human intervention. Image processing solutions such as contour detection (Gong et al., 2018) and image classification (Lu & Weng, 2007) can be used effectively for documents that were scanned with good image quality, and that have comparable text size and font. However, when it comes to handwriting text, such methods are not effective. Therefore, a separate stream of research focuses on the automatic transcription of handwritten documents, referred to as Handwritten Text Recognition (HTR). Recent studies show that supervised learning methods are widely used for OCR and HTR (Agarwal et al., 2013; Pu, 2021). While these approaches reach an accuracy of close to 99% on popular datasets, such as the MNIST (LeCun et al., 1998), they also require huge amounts of labeled data for training the models due to their complexity. Furthermore, the annotation process is usually manual, resulting in a time-consuming, and expensive process. This process is particularly laborious for handwritten text recognition because the same text could be written in various forms and styles, and by different people.

In the subsections below, we review recent works in various subtasks of an OCR solution, which reflects our proposed solution in supporting HTR for Vietnamese medical documents.

Deep-Learning Based Optical Character Recognition

Text Detection

Text, including characters and punctuations, can be treated as objects on an image and therefore, can be detected using *object detection* techniques. Before the emergence of the deep learning paradigm, text detection was mainly handled using hand-crafting features to detect characters, as individual objects, in an image (Matas et al., 2004). With deep learning, especially in using Convolutional Neural Network (CNN) method for object detection and semantic segmentation, advanced object detection architectures such as the Single-Shot MultiBox Detector (SSD) and Faster R-CNN models (Liu et al., 2016; Ren et al., 2017) are shown to be efficient text detectors. Liao et al. develop a regression-based text detector called TextBoxes (Liao et al., 2016). Subsequent works, including Deep Matching Prior Network (DMPNet) and the Rotation-Sensitive Regression Detector (RRD) (Liao et al., 2018; Liu & Jin, 2017), follow this approach to develop regression-based models, which are adaptive to orientations. Tian et al. propose Connectionist Text Proposal Network, which combines CNN with recurrent networks using a vertical anchor mechanism to improve accuracy for detecting horizontal text (Tian et al., 2016).

A character in an image is also made of a collection of pixels. As a result, *instance segmentation*, which is the task of classifying pixels in an image as pre-defined classes, is a useful approach to detecting characters. The per-pixel classification process can be quite accurate by estimating the probabilities of characters, and their relationships among adjacent characters to form text units. Compared to the object detection approach, popular segmentation methods such as Fully Convolutional Networks (FCN) are used successfully to improve text detection, especially when text is misaligned or distorted (Long et al., 2015). Related work including (Deng et al., 2018; He et al., 2017; Yao et al., 2016) uses segmentation methods to generate word-bounding boxes by extracting bounding areas directly from the segmentation output.

In this work, we perform instance segmentation for both the text area detection and the text detection tasks. Especially, we combine deep learning models such as Detectron2 with either Mask R-CNN or Faster R-CNN to identify the handwriting ROIs. We discuss and evaluate this technical stack in sections 3 & 4, respectively.

Text Recognition

Visual Geometry Group (VGG), originally designed for object recognition tasks as the successor of AlexNet, demonstrates its competency for OCR problems (Ahmad & Farooqui, 2021). In the study, the proposed method is a transfer learning-based approach with a VGG19 already trained on millions of images derived from the ImageNet dataset. The model achieves 88.65% accuracy in the Validation set (766 images), 87.56% in Test Set (1543 images), and 93.24% in Train Set (5403 images).

A transformer-based approach, TrOCR, deviates from the popular CNN approach to extract character features. Instead, the input images are resized and split into sequences of 16 x 16 patches. The standard Transformer architecture which consists of Encoder and Decoder blocks, combined with the Self-attention mechanism, is the generator of word-level prediction. According to the authors, both Encoder and Decoder could be initialized with pre-trained ViT-style models and BERT-style models (Li et al., 2021).

Efforts in Vietnamese Handwritten Text Recognition

Recent advances base on deep-learning (DL) framework are practical, however, it is still challenging in recognizing Vietnamese handwriting because of the presence of extra six distinctive tonal symbols and extra vowels. Google Vision API (Google Inc, 2022) and Microsoft Computer Vision API (Microsoft Inc, 2022) could be considered the most popular OCR solutions for NLP tasks such as handwriting detection, with reliable accuracy. However, these commercial solutions do not work well for Vietnamese handwriting. Nguyen et al. used BLSTM to handle delayed stroke problem in Vietnamese handwriting. The authors achieve this milestone through BLSTM learning capability that bypass vanishing gradient (Nguyen et al., 2018). Phung et al. also proposed a solution which takes an image of a medical file and transcribes it into an electronic document (Phung et al., 2020). The approach performs segmentation at the word-level and uses an DL architecture consisting of ResNet – BiLSTM - CTC to recognize and convert the text. The output sequence is then tied to a lexicon to further boost the accuracy. While the work returns encouraging accuracy results, their text recognition model is expensive to train. Furthermore, their pipeline fails to detect the text located outside of the preset writing area, and thus key information could be omitted. The pipeline proposed here addresses this limitation by adding a *Text Area Detection* component in the solution.

Efforts in Medical Handwritten Text Recognition

To detect multilingual texts (Chinese and Latin characters) in medical laboratory reports, Xue et al. built a model based on the Faster RCNN architecture with VGG16 and a regional proposal network (RPN) (Xue et al., 2020). The model output loss of text/non-test classification and bounding box regression, crucial components for the subsequent training procedure. A patch-based strategy is utilized for training the model on high-resolution images and results in positive Recall (99.5%), F1-Measure (99.1%) and Average Precision (90.9%) when experiment on 18402 textual instances with 351 different characters. Also, Wu et al. experiment with a multicomponent Text Detection pipeline consists of 4 parts: feature extractors (CSPDarknet53 and SPP structure), feature fusion (PANet structure), fine-grained text prediction (RPM structure) and text line construction (Connectionist Text Proposal Network). The authors used the model to detect text region of instructions in Chinese medicine package (Wu et al., 2021). Recently, Tabassum et al. present an online handwritten medical words recognition system to support busy doctors in Bangladesh (Tabassum et al., 2022). The work bases on a handwritten medical-term dataset, with containing 17431 samples of 480 medical terms, and performs the text recognition using Bidirectional LSTM and RSS data augmentation.

The Handwritten Text Recognition Pipeline

In this section, we describe a solution for the handwritten text recognition tasks using a Vietnamese doctor's handwriting dataset. The advance of this pipeline is the combination of image processing tasks including deblurring and denoising tasks, text detection task, text recognition task, and finally text correction task. Figure 1 below illustrates the key text and image processing components.

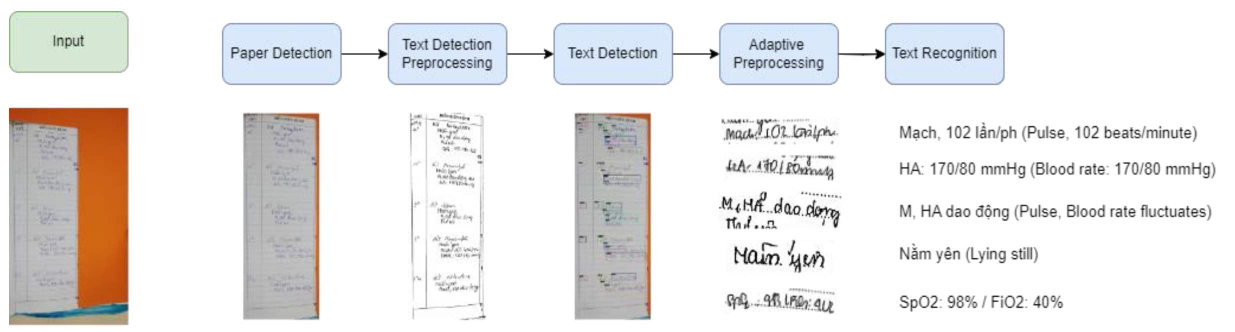

Input
Paper Detection
Text Detection Preprocessing
Text Detection
Adaptive Preprocessing
Text Recognition
Mạch, 102 lần/ph (Pulse, 102 beats/minute)
HA: 170/80 mmHg (Blood rate: 170/80 mmHg)
M, HA dao động (Pulse, Blood rate fluctuates)
Nằm yên (Lying still)
SpO2: 98% / FiO2: 40%

Figure 1. The complete HTR pipeline

Vietnamese Handwriting Characteristics

Vietnamese text uses the Roman alphabet with six based tonal symbols / diacritical marks (DMs) positioned either above or below the characters: flat – (no diacritical marks), grave – (`), acute – (´), hook – (ˀ), tilde – (~), under dot – (.) (Nguyen et al., 2018). The language also contains seven additional characters with diacritics for vowels (ă, â, ê, ô, ơ, ư) and one consonant (đ). Combining the DMs with all based characters, 67 different derivative characters can be created that do not exist in the Latin alphabet. Such extensive combinations pose a challenge for Vietnamese handwriting recognition. For example, from the base word such as "bao" we can have: "bao" (bag), "báo" (inform), "bào" (shred), "bảo" (told), "bão" (storm). Considering a short sentence such as "báo bác sĩ" (inform the doctor) in Figure 2.

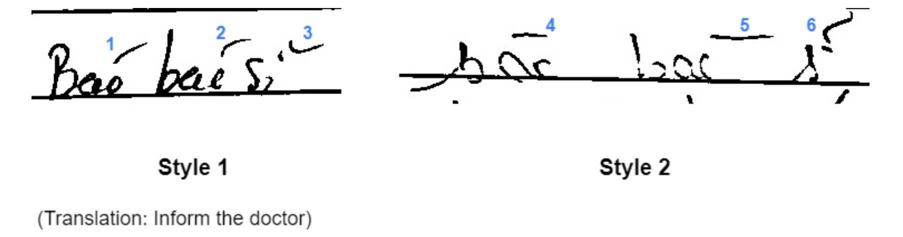


Figure 2. Base diacritical marks example

Figure 2's Style 1 has a more conventional approach in illustrating the acute marks with marks 1 and 2 having a prominent upward direction from left to right. While Figure 2's Style 2 has marks 4 and 5, also acute marks, in a much flatter manner, almost horizontally even. With the context of the sentence, a reader could infer that these marks are indeed acute marks. But if displayed solely, the style could create the false impression of marks 4 and 5 being a grave mark instead.

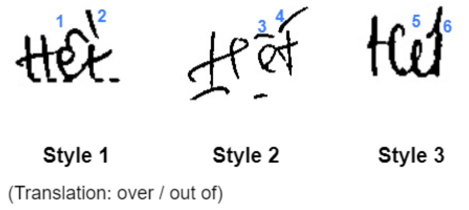


Figure 3. Combination diacritics example

Combinations of diacritical marks could also lead to diverse scenarios. With "hết" (over / out of), in Figure 3's Style 1 and 2, the diacritics have a clear separation between each pair. On the other hand, with Figure 3's Style 3, marks 5 and 6 seemingly "merged" and become a single stroke that acquires characteristics of the circumflex in "ê" and the acute mark. Such characteristics created ambiguous scenarios that affect ground-truth preparations. The research team, with the help of OUCRU medical specialists, must navigate carefully through the acquired text instances when providing annotations.

The Dataset and Data Preprocessing Tasks

For this study, records from patients with tetanus were selected from the hospital records department. Tetanus was selected because (1) HTD is a centre for tetanus referral and research into tetanus at the site has had informed global management for the disease and digitization of records could benefit future research, and (2) the disease management is standard, meaning that words and phrases are repeated though out the document. Our sample of records included 8000 scans from 118 different medical records of patients admitted between 2020 and 2021, representing approximately 30 doctors of varying degrees of seniority. Importantly, to process this data for training the HTR pipeline, we take measure to ensure patient privacy, data security, and confidentiality in this process. First, only anonymized data were captured as physical documents were de-identified by covering the text areas that disclose personal information (Figure 4). Second, the data collection and annotation process were conducted as approved by the Ethical and Scientific Committee of the Hospital for Tropical Diseases and Oxford Tropical Research Ethics Committee.

Each scanned document contains lines of the doctor's handwriting capturing the treatment and the health progress of a patient, which are the regions of interest. In such a form, the data poses several image-processing challenges.

1. Only part of the image is the region of interest (ROI).
2. The forms contain either dash line or horizontal straight line underneath each handwriting section.
3. Inconsistent clarity in each document. Some areas are blurrier than others.

These obstacles call for data preparation tasks such as text-area detection, image quality improvement, and handwriting text detection before the handwriting recognition tasks can be executed effectively.

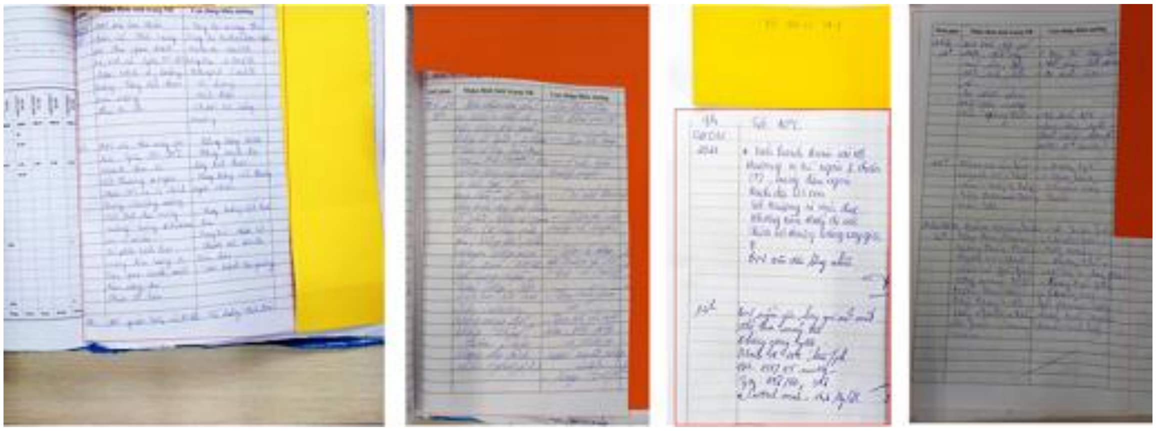


Figure 4. Examples of text area detection (the image is blurred intentionally)

Text Area Detection

We apply instance segmentation to detect the pixels that belong to the document layout, then remove the irrelevant surroundings including coverings, surface, stacked documents behind the papers. We develop the Detectron2 (Wu et al., 2019) combined with Mask R-CNN-based model. Mask R-CNN extends Faster R-CNN with additional branches, made of small Fully Convolutional Networks (FCN), applied for each ROI (He et al., 2017). These FCN branches produce pixel-to-pixel segmentation masks for each ROI while only incurring a small computational overhead. Thus, Mask R-CNN enables efficient and fast computation.

We received positive results. For instance, the text-area segmentation task gets 92% average precision (AP) on a training set of 600 images. Furthermore, our model was able to perform prediction to distinguish the document layout from the surrounding covers, background surface, or additional page in the case of the first image in Figure 4. We also customized the code to generate outputs in JSON format which can be used to create more annotated data on demand. This step is important to filter out irrelevant surroundings from our image and reshape the document into a readable format for the later image-processing steps. Note that the Text Detection task (discussed below) also uses the Detectron2 library to capture text layouts (i.e., lines of handwriting text) on the segmented document.

Image Adaptive Preprocessing

Blurry handwriting strokes and/or unexpected vertical and horizontal lines in the ROIs are inevitable, hence, an image preprocessing phase is needed to enhance the clarity of the text, as shown in Figure 5. To address blurry or thin handwriting strokes, we implement the *Contrast Limited Adaptive Histogram Equalization* (CLAHE) technique (Toet & Wu, 2014). CLAHE involves two phases of processing.

- *Adaptive Histogram Equalization* (AHE) helps divide the image into multiple small regions with a calculated histogram. The output is then used for the distribution of local grayscale values and returns a result with enhanced contrast value.
- *Contrast Limited* (CL) mechanism prevents overamplification by putting a limitation on the value of contrast enhancement to the output.

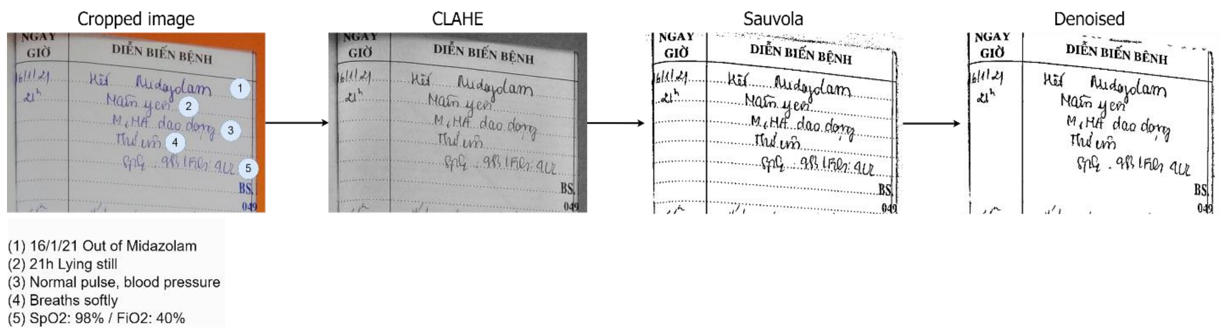

Cropped image
CLAHE
Sauvola
Denoised
(1) 16/1/21 Out of Midazolam
(2) 21h Lying still
(3) Normal pulse, blood pressure
(4) Breaths softly
(5) SpO2: 98% / FiO2: 40%

Figure 5. Complete image pre-processing steps

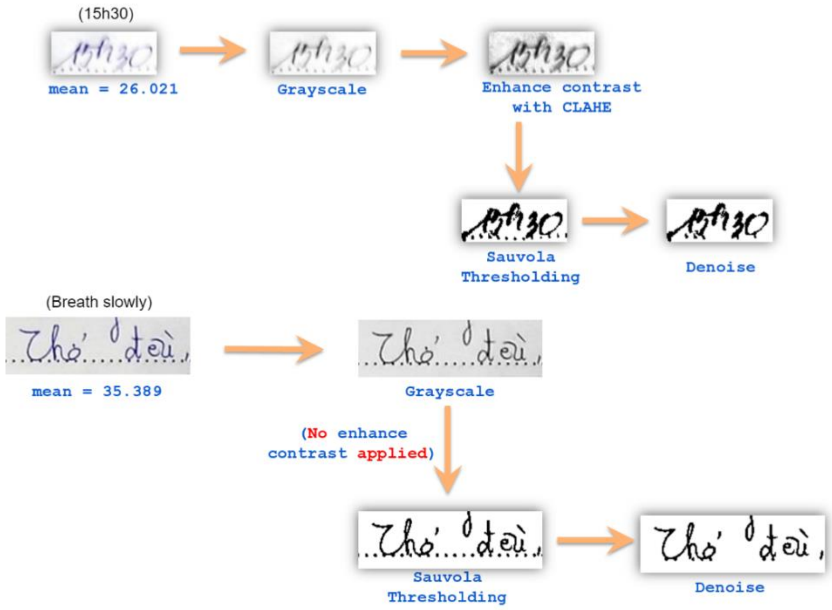


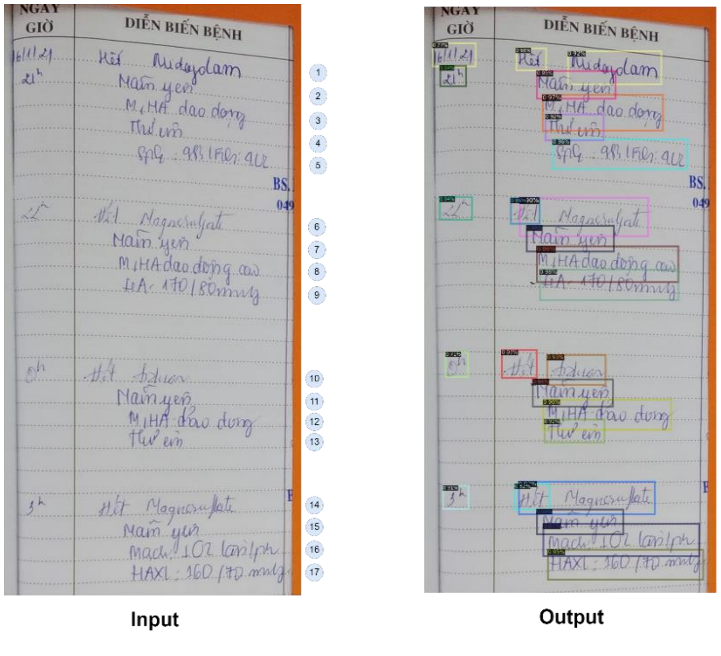
Figure 6. Examples of adaptive pre-processing mechanism

Nevertheless, the CLAHE process can increase the risk of masking away blurry yet relevant handwriting strokes. Hence, we implement a blur metric using *Fast Fourier Transform* to assess the level of blurriness in every handwriting text unit and assign appropriate CLAHE parameters. While CLAHE improves the contrast, we implement the adaptive document image binarization process proposed by Sauvola et al. (Sauvola & Pietikäinen, 2000) followed by either a *Soft Decision Method* (SDM) or a *Specialized Text Binarization* method (TBM) (Sauvola & Pietikäinen, 2000). Finally, we denoise the input by removing dots and dot lines below the text input. Figure 6 above illustrates the process. Importantly, the steps and the corresponding parameters are not fixed but adapt to the blurriness for the best image enhancement.

Handwriting Text Detection

Handwriting text detection is the process of creating bounding boxes on the scanned document to identify the text units that will be transcribed by the text recognition model. Examples of our effort are shown in Figure 7 below. Overall, handwriting text detection is a challenging problem because handwriting text diversifies in shapes, sizes, orientations, and sometimes can be distorted. Furthermore, there are computer-based text and other non-handwriting items such as seal marks, QR codes, and signatures that should be kept intact and should not be transcribed. The handwriting text detection component described here needs to ignore those units and only detects and highlights pieces of handwriting text, and we call those *detection units*. Each detection unit will consist not just one word but a complete line in the text area. This is important because we aim to explore the contextual and positional relationship of each word, thus allowing us to perform word correction at the end of the recognition process.

Similar to the Text Area Detection module, we combine Detectron2 and Faster R-CNN to identify the handwriting lines. After training with 3,303 text instances, the final model delivers an acceptable accuracy mean Average Precision (mAP) of 58%, and all the detection confidence is over 90%.



(1) 16/1/21 Out of Midazolam
(2) 21h Lying still
(3) Pulse, blood pressure fluctuates
(4) Breaths softly
(5) SpO2: 98% / FiO2: 40%
(6) 22h Out of Magnesulfate
(7) Lying still
(8) High pulse, blood pressure fluctuation
(9) HA: 170/80 mmHg
(10) 0h Out of Arduan
(11) Lying still
(12) Normal pulse, blood pressure
(13) Breaths softly
(14) Out of Magnesulfate
(15) Lying still
(16) Pulse: 102 beats/minute
(17) Invasive Blood Pressure: 160/70 mmHg

Figure 7. Handwriting text detection output

Handwritten Text Recognition with a VGG19–Transformer Model

With the cropped images of lines of handwriting text from the Image Adaptive Preprocessing step, we now extract the multifarious features of the handwriting, then predict a context behind the sentences based on the respective position of each word and deliver a full translation into machine text form. We develop two components for this task.

Feature extraction with VGG19

After experimenting with multiple handwriting recognition solutions, we utilize a variant of the VGG19-Transformer model pretrained with general Vietnamese handwriting. The VGG19 model composes of 16 convolution layers, 3 Fully connected layers, 5 MaxPool layers and 1 SoftMax layer. The full architecture of the VGG19-Transformer can be found in Figure 8 below.

With more convolutional layers, the VGG19 model can fit complex features while still at the right depth to address issues like Vanishing Gradients. The model first converts an RGB image to a matrix of shape (224,224,3). The Convolution Layers then create a feature map with kernels (3x3). The convolution stride is set to 1 pixel to ensure full coverage of the image and spatial padding adapts to the convolutional layers input to preserve the spatial resolution. The pooling layers compress the features before extracting small details to understand the components of characters and words (Simonyan & Zisserman, 2014).

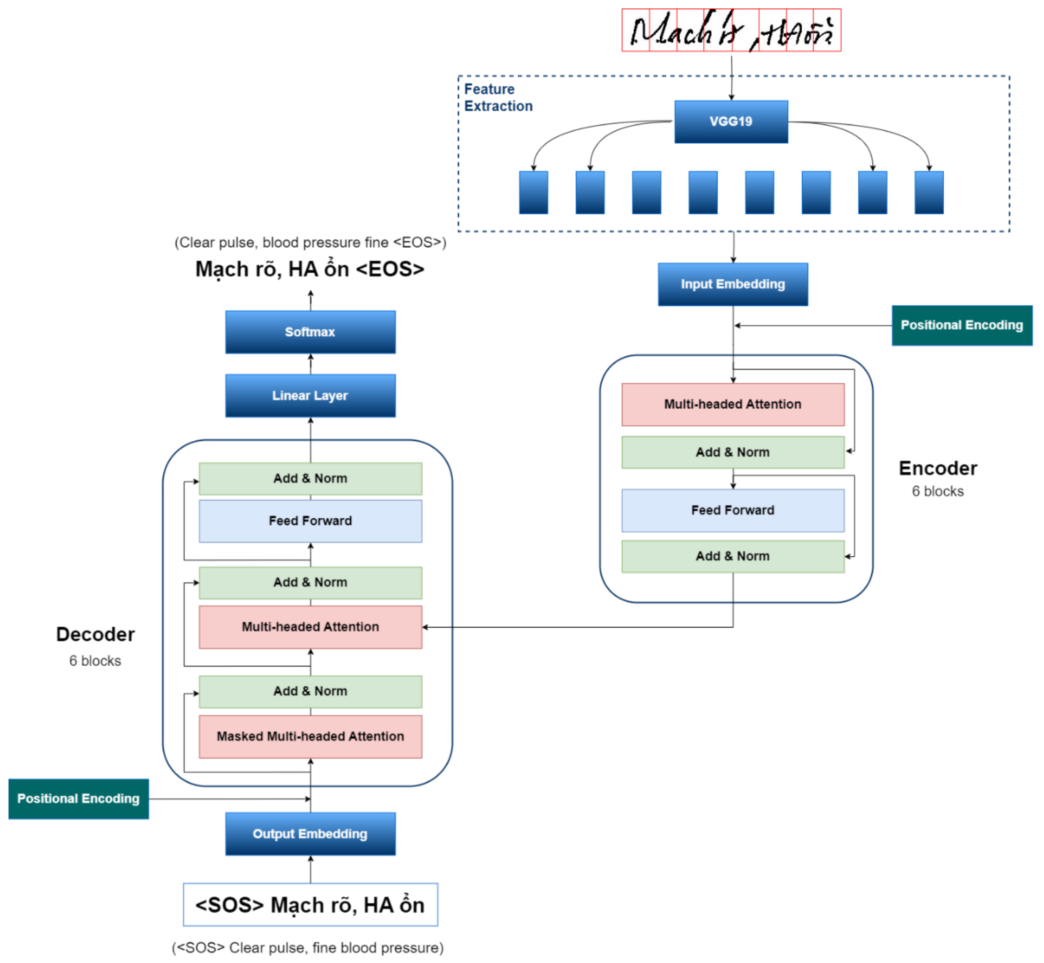


Figure 8. VGG19-Transformer architecture

Text recognition with Transformer

Objectives of a handwritten text recognition model include (1) capturing long-range dependencies, (2) learning the context behind the sequence, and (3) understanding the relationships between words and characters of a sentence. Objectives (2) and (3) can be achieved with the Bidirectional Long-Short Term Memory (Bi-LSTM) (A. Graves et al., 2009), which recurrently traverses a sentence in both directions and "remembers" information during the process. However, Bi-LSTM still suffers the vanishing gradient issue when dealing with very long-term dependencies. The transformer neural network can handle well long-range dependencies while performing sequence-to-sequence tasks (Vaswani et al., 2017).

Positional Encoding. The element positions determine the context of the sentence. Because Transformer takes a full sentence as input, each element needs to have a unique encoded value to portray its relative position. The formulas below produce two signals that when concatenated, form the Positional Encoding vectors. This mechanism captures the positional quality of elements, and it is placed at both embedding steps of the input (handwriting features) and the expected output (text features). The translation process continues with Encoder and Decoder phases.

$$PE[pos, 2i] = sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right) \qquad PE[pos, 2i+1] = cos(\frac{pos}{1000^{\frac{2i}{d_{model}}}})$$

Where pos is the current position, d_{model} is the model's fixed dimension and i is relative position in d_{model}.

Encoder Block. As shown in Figure 9, an encoder consists of a multi-head attention layer and a feed-forward layer, further connected by Residual connections, and followed by a normalization layer to reduce training time. The multi-head attention layer takes each word as input and generates three types of attention vectors: *Query (Q), Key (K),* and *Value (V)* by multiplying the packed embedded position with the weight matrices of the respective matrices (W_Q, W_K and W_V) (Alammar, 2018). We use a stack of 6 Encoder blocks, each with a dimension of 256 and consisting of a total of 8 heads for the Multi-head attention layer. We compute the weighted average of the resulting vectors as self-attention scores using the formula below, which contains the contextual relationship between the focused element and others of the same sequence. An example of self-attention scores for sequential elements ['Mạch', 'rõ', ',', 'HA', 'ổn'] is shown in Figure 10.

$$Z = softmax\left(\frac{QK^T}{\sqrt{Dimension\ of\ vector\ Q, K\ or\ V}}\right)V$$

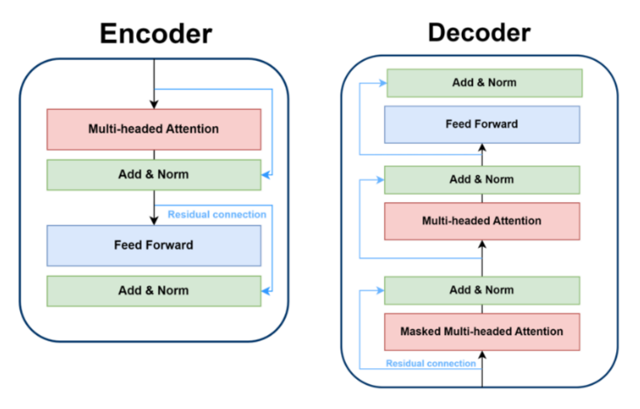


Figure 9. Encoder Block & Decoder Block

To direct more attention to other elements, multi-head attention prepares different Query/Key/Value weight matrices for each attention head. The system has 8 attention heads that directly influence the creation of 8 distinctive sets of attention scores. Ultimately, the overall attention benefits from receiving additional depths to its layers by learning from selective input elements that directly relate to a queried element or provide additional contextual information. Similar approaches show good performance (Bahdanau et al., 2014) in which for each generated unit, the model soft-search positions of the most relevant information in the original sentence to compose context vectors aiding the prediction. Then, the vectors are concatenated and passes through the feed-forward layer to be normalized before entering the next Encoder block or the first Decoder block (Ankit, 2022).

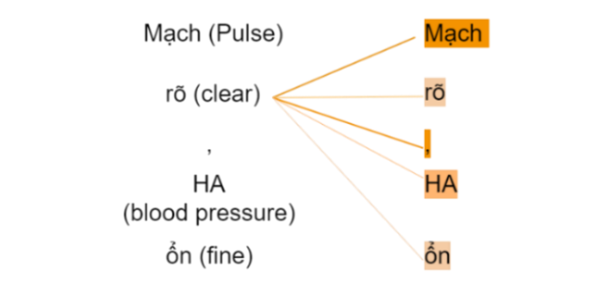


Figure 10. Attention Weight visualization

Decoder Block. Like the Encoder, there are 6 Decoder blocks in total, each consisting of a multi-headed attention layer with 8 attention heads receiving either embedding vectors of the expected output or the processed output from the previous Decoder block.

However, the main difference comes in two-fold:

1. Layer (1) of the Decoder – Masked multi-headed attention has a slightly different mechanism from Layer (1) of the Encoder. The learning process requires the next element in the sequence to be hidden (masked) so that the decoder attempts to predict it. This is necessary for the learning of the wording relation since exposure to a full sentence beforehand would be counterproductive (Ankit, 2022).
2. The extra multi-headed attention layer – Aside from receiving positional embedding machine text features at layer (1), the Decoders also utilized the Encoder's output to calculate mutual attention between optical handwriting features and predicted machine text features.

The direct connection between the Encoder and Decoder (Figure 8), where all sequence elements are transferred, addresses the long-range dependencies. Thus, the context is retained better, achieving both the second and third objectives. Outside of the Decoder Block, the Linear layer converts the predicted words to the encoded label of characters and symbols, follows by the Softmax layer to output the probability value.

Correction Algorithm

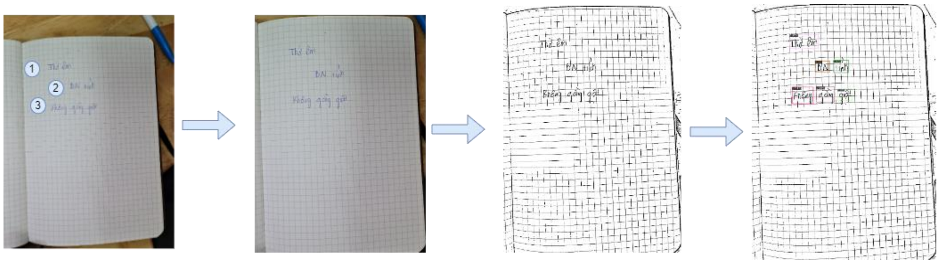
Even when the model achieves good prediction with typical Vietnamese words, we encounter cases where multiple words in conjunction with medical terms generate character errors. We introduce a correction mechanism that works similarly to a dictionary. First, we calculate of word occurrence in the dataset to capture the most occurred medical terms (for example: Midazolam, Arduan, Magnesulfate,…) along with pairs of words ("hết Magnesulfate" (out of Magnesulfate), "dao động" (vibrate)). For each word and those pairs in a prediction, the algorithm then calculates the Levenstein distance against terms and pairs prepared in the dictionary. The most likely choice with weighted edit distance under 5% - 15% (i.e., difference up to 2 to 3 characters based on the word length) of the dictionary term will be replaced. This addition helps reduces the Character Error Rate (CER) and Word Error Rate (WER) down to 2% and 12% respectively.

Evaluation Results

Image Adaptive Preprocessing

Apart from the typical image enhancement tasks, the image adaptive preprocessing module, which combines CLAHE, Sauvola, and OpenCV denoising technique, reduces the noises in the background and maintains the handwriting strokes, significantly. Initially, we considered horizontal/vertical lines of the medical form as noise, which needs to be removed before the text can be identified. However, the adaptive preprocessing, especially the adaptive Sauvola in combination with the OpenCV denoising function allows text features to be amplified while horizontal/vertical lines have minimal impact on recognition efficiency.

Figure 11 below illustrates the performance of the text detection step after the scanned document was preprocessed. The handwriting stroke still displays a good quality while most of the vertical and horizontal lines remain. A closer look shows that some parts of those lines were removed as noise. The output then comes through the text detection and results in the last image where all the texts are detected successfully.



Translation:

(1) Thở êm (breathe slowly)

(2) BN tỉnh (Patient awake)

(3) Không gồng giật (No jerking)

Figure 11. Preprocessing and Text Detection result on unfamiliar format

Text Recognition

We acquired the ground truth for 1800 images of handwriting lines taken from Tetanus patients' nursing diaries and checkup forms with 100% confidence in the annotation (verified by the hospital staff). Additionally, all images have been processed with all steps by the Preprocessing module. Another batch of 200 images was selected for validation. The evaluation experiment is carried out with 4 models below, which were trained on Tetanus handwriting scanned records. We capture the CER, WER, and Average Training Time as performance evaluation metrics for comparison.

- RestNet50 – Transformer
- VGG19 – Transformer
- Resnet50 – BiLSTM – CTC (Phung et al., 2020)
- A variant of VGG19–Transformer, pretrained using generic Vietnamese text images.

Architecture	CER	WER	Average Training Time (min)
Resnet50–Transformer	0.8152	0.9246	9.34
VGG19–Transformer	0.1864	0.4152	9.67
Resnet50–BiLSTM–CTC (Pretrained)	0.89	1.0	**20.21**
VGG19–Transformer (Pretrained)	**0.1339**	**0.2314**	10.5

Table 1. Performance of various Handwritten Text Recognition architectures

As shown in Table 1 above, we observe that a hybrid architecture VGG19–Transformer with no pretrained weights, produces fewer character and word errors in validation compared to a combined Resnet50–Transformer model. This result suggests that VGG19 makes a comparably better features extractor for the problem of recognizing doctor handwritten text. Furthermore, it seems that the combination of Resnet50–BiLSTM–CTC does not perform well in terms of training time because recurrent neural network architectures such as BiLSTM are usually more computationally expensive in training the prediction model.

The model that performs the best in terms of accuracy is VGG19 – Transformer with the pretrained weights loaded. It entails that weight training on a larger sample size and more diversity in terms of both contexts and text type (handwriting and machine) create a good foundation for the model to gain knowledge. In later stages of the training, the model adapts well and can infer new, unfamiliar medical lines with good accuracy. Figure 12 shows a noticeable trend in how the WER and CER steadily decrease as more data are passed into the training, indicating good adaptability to wider variants of handwriting styles, characters, and words.

Between training of size 1,000 and size 1,800, we decided to experiment further with the hyper-parameters such as batch size, iteration, and max learning rate for deeper understanding and to establish a clear set of parameters for future training. According to Table 2 and Table 3, we adjust our training process to use batch size 8 and a max learning rate of 0.00003 for later experiments. Furthermore, instead of retraining with different iteration settings, we decided to increase the iteration to 10,000 and observe that the model could still improve. The recommended parameters setting for the final training and future training are **{max_learning_rate: 0.00003, batch_size: 16, iterations: 10000}**. With the setting, we achieve 10% of CER and 21% of WER on the evaluation set at the end of this experiment. Upon further inspection, we detected character errors in medical term predictions and common phrases that could be avoided with a dictionary mechanism. This is the motivation for the Correction Algorithm module, as presented above.

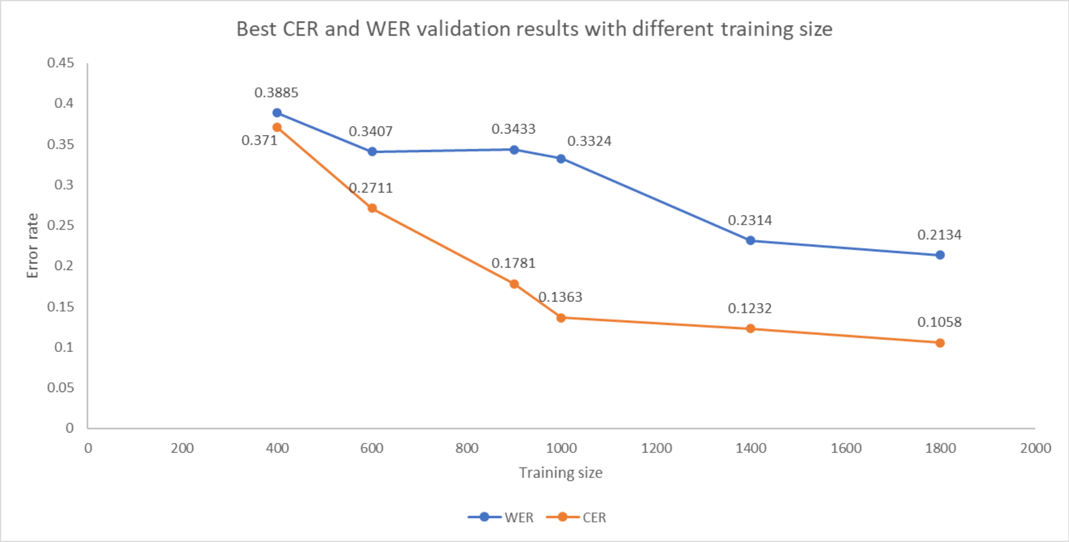


Figure 12. Accuracy improvement as the size of the training set increases

Batch size Experiment	CER_B4	WER_B4	CER_B8	WER_B8	CER_B16	WER_B16
Minimum Error Rate	0.1339	0.2763	0.1339	0.2763	**0.2308**	**0.1201**
Average Error Rate	0.1713	0.3110	0.1455	0.3040	**0.267278**	**0.12534**

Table 2. Experimenting with various Batch Size

Max Learning Rate (LR) Experiment	WER_ 0.003	CER_ 0.003	WER_ 0.0003	CER_ 0.0003	WER_ 0.00003	CER_ 0.00003
Minimum Error Rate	0.3114	0.1659	0.2432	0.1248	**0.1058**	**0.2134**
Average Error Rate	0.9764	0.9187	0.2840	0.166492	**0.1208**	**0.2465**

Table 3. Experimenting with various Learning Rate

Conclusion

Besides medical examination and treatment, a time-consuming job for doctors and nurses in healthcare facilities is to record medical notes and books. Importantly, medical staff in Vietnam must complete paper medical records and enter them into computers, which doubles the work. Furthermore, with doctors and nurses still having to hand-write the medical records, the data is scattered; thus, the statistics and synthesis of medical procedures take a lot of time. There have been great advances in tackling the handwritten text recognition (HTR) problem. However, beyond the challenges of Vietnamese handwritten text such as the presence of complex vocals and tonal symbols, physical medical notes and documents pose issues such as partial text coverage, blurry handwriting strokes, and/or unexpected vertical and horizontal lines.

In this work, we develop and evaluate a complete image processing pipeline to perform handwritten text recognition on scanned Vietnamese medical records for Tetanus cases. While the pipeline is specifically developed for recognizing Vietnamese handwriting, the study enables us to explore a wide range of image processing solutions, especially those applying deep learning techniques. From convolutional neural networks (CNN) to recurrent neural network (RNN) to the latest transformer architectures, deep learning shows a significant advancement in performing instance segmentation and features extraction, to text recognition. Nevertheless, we also acknowledge the effectiveness of traditional image processing algorithms, especially in enhancing the quality of images, in preparation for a more efficient training process. The proposed pipeline demonstrates promising results, especially in the context where works addressing the difficulties of digitizing Vietnamese handwritten medical records remain scarce. The success of this HTR technology is beneficial to other public hospitals in Vietnam because (1) they rely on handwriting documentation process, and (2) they use a standardized paper medical records throughout the hospital system. As a result, the solution proposed here can easily be adapted to other hospitals. Finally, as this work targets a specific medical context in a low-/middle income country like Vietnam, facilitates the digital transformation of medical centers and hospitals, and hence furthers the readiness for adopting modern EHR management systems, and creating conditions for the formation of a national health database.

We acknowledge the limitation of our current work and suggest the following future work. First, more annotated data will improve the accuracy of our deep learning models. However, labeled datasets are costly to acquire, especially in a medical setting. We aim to integrate Active Learning to minimize the number of required training labels while continuing to improve the model's performance. Especially, we plan to develop an application for online learning where high-confidence samples are automatically selected, iteratively assigned, and pseudo-labels are updated. Second, while the presented system performs well to digitize physical paper-based documents, the process to capture photos (or scans) of those physical documents could be time-consuming and non-scalable for larger healthcare facilities. We suggest extending the data processing pipeline to connect with peripheral devices such as scanners and mobile devices and supporting batch processing.

References

A. Graves, M. L., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A Novel Connectionist System for Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(5), 855-868. **https://doi.org/10.1109/TPAMI.2008.137**

Agarwal, A., Garg, R., & Chaudhury, S. (2013). Greedy Search for Active Learning of OCR. 12th International Conference on Document Analysis and Recognition, Washington DC.

Ahmad, F., & Farooqui, Z. (2021). Deep Learning Based Optical Character Recognition in Natural Images. International Research Journal of Modernization in Engineering Technology and Science, 3(8), 731-737.

Alammar, J. (2018). The Illustrated Transformer. Retrieved 10/06/2022 from **https://jalammar.github.io/illustrated-transformer/**

Ankit, U. (2022). Transformer Neural Networks: A Step-by-Step Breakdown. Retrieved 12 Sep from **https://builtin.com/artificial-intelligence/transformer-neural-network**

Cuk, S., Wimmer, H., & Powell, L. M. (2017). Investigating Problems Associated with Patient Care Reports and Transferring Data Between Ambulance and Hospitals from the Perspective of Emergency Medical Technicians.

Deng, D., Liu, H., Li, X., & Cai, D. (2018). PixelLink: Detecting Scene Text via Instance Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 32. **https://doi.org/10.1609/aaai.v32i1.12269**

Gong, -. X.-Y., Su, -. H., Xu, -. D., Zhang, -. Z.-T., Shen, -. F., & Yang, -. H.-B. (2018). - An Overview of Contour Detection Approaches. - International Journal of Automation and Computing%V - 15(- 6), - 656. **https://doi.org/- 10.1007/s11633-018-1117-z**

Google Inc. (2022). Cloud Vision API. Retrieved 06/12/2022 from https://cloud.google.com/vision

He, W., Zhang, X. Y., Yin, F., & Liu, C. L. (2017, 22-29 Oct. 2017). Deep Direct Regression for Multi-oriented Scene Text Detection. 2017 IEEE International Conference on Computer Vision (ICCV),

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of IEEE, 86(11), 2278-2324.

Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., & Wei, F. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. ArXiv. **https://doi.org/10.48550/arxiv.2109.10282**

Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. (2016). TextBoxes: A Fast Text Detector with a Single Deep Neural Network. Proceedings of the AAAI Conference on Artificial Intelligence, 31. **https://doi.org/10.1609/aaai.v31i1.11196**

Liao, M., Zhu, Z., Shi, B., Xia, G. s., & Bai, X. (2018, 18-23 June 2018). Rotation-Sensitive Regression for Oriented Scene Text Detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016, 2016//). SSD: Single Shot MultiBox Detector. Computer Vision – ECCV 2016, Cham.

Liu, Y., & Jin, L. (2017). Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298965**

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing, 28(5), 823-870. **https://doi.org/10.1080/01431160600746456**

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing, 22(10), 761-767. **https://doi.org/https://doi.org/10.1016/j.imavis.2004.02.006**

Microsoft Inc. (2022). Computer Vision. Retrieved 06/12/2022 from **https://azure.microsoft.com/en-us/products/cognitive-services/computer-vision/**

Nguyen, H. T., Nguyen, C. T., Bao, P. T., & Nakagawa, M. (2018). A database of unconstrained Vietnamese online handwriting and recognition experiments by recurrent neural networks. Pattern Recognition, 78, 291-306. **https://doi.org/10.1016/j.patcog.2018.01.013**

Phung, T. M., Dinh, M. N., Dang, D. P. T., Van, H. M. T., & C. Louise Thwaites. (2020). A Machine Learning-based Approach to Vietnamese Handwritten Medical Record Recognition. Australasian Conference on Information Systems, Wellington, New Zealand.

Pu, T. (2021). Application of active learning algorithm in handwriting recognition numbers. Journal of Physics: Conference Series, 1861(1). **https://doi.org/10.1088/1742-6596/1861/1/012060**

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis &amp; Machine Intelligence, 39(06), 1137-1149. **https://doi.org/10.1109/tpami.2016.2577031**

Samuel, P. (2021). Vietnam's Digital Transformation Plan Through 2025. Retrieved 20/10/2022 from https://www.vietnam-briefing.com/news/vietnams-digital-transformation-plan-through-2025.html/

Sauvola, J., & Pietikäinen, M. (2000). Adaptive document image binarization. Pattern Recognition, 33(2), 225-236. **https://doi.org/10.1016/S0031-3203(99)00055-2**

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Tabassum, S., Abedin, N., Rahman, M. M., Rahman, M. M., Ahmed, M. T., Islam, R., & Ahmed, A. (2022). An online cursive handwritten medical words recognition system for busy doctors in developing countries for ensuring efficient healthcare service delivery. Scientific Reports, 12(1), 3601. **https://doi.org/10.1038/s41598-022-07571-z**

Thu Vien Phap Luat. (2018). THÔNG TƯ QUY ĐỊNH HỒ SƠ BỆNH ÁN ĐIỆN TỬ. Retrieved 21/011/2022 from **https://thuvienphapluat.vn/van-ban/Cong-nghe-thong-tin/Thong-tu-46-2018-TT-BYT-su-dung-va-quan-ly-ho-so-benh-an-dien-tu-391438.aspx**

Tian, Z., Huang, W., Tong, H., He, P., & Qiao, Y. (2016). Detecting Text in Natural Image with Connectionist Text Proposal Network (Vol. 9912). **https://doi.org/10.1007/978-3-319-46484-8_4**

Toet, A., & Wu, T. (2014). Efficient contrast enhancement through log-power histogram modification. Journal of Electronic Imaging, 23(6), 063017-063017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wu, H., Zhou, R. G., & Li, Y. (2021). A Neural Network Model for Text Detection in Chinese Drug Package Insert. IEEE Access, 9, 39781-39791. **https://doi.org/10.1109/ACCESS.2021.3064564**

Xue, W., Li, Q., & Xue, Q. (2020). Text Detection and Recognition for Images of Medical Laboratory Reports With a Deep Learning Approach. IEEE Access, 8, 407-416. **https://doi.org/10.1109/ACCESS.2019.2961964**

Yao, C., Bai, X., Sang, N., Zhou, X., Zhou, S., & Cao, Z. (2016). Scene Text Detection via Holistic, Multi-Channel Prediction.