OCR Application for Cancer Care

Settha Tangkawanit, Jiraporn Pooksook, Jirarat Ieamsaard, and Panupong Sornkhom*
Department of Electrical and Computer Engineering, Faculty of Engineering, Naresuan University, Thailand
* E-mail: panupongs@nu.ac.th

Abstract— **The number of cancer patients in Thailand is increasing every year. During the cancer treatment process, patients might need to go to the hospital often, causing time consumption and the risk of getting germs. Instead of going to the hospital for blood tests, patients could have a blood test at any available blood test lab near their accommodation and send it to the nurse or medical staff. This paper presents an OCR application to collect blood test results from patients where the medical staff could use this information for cancer care plans. From experimental results, the proposed application provided 81.5% accuracy.**

I. Introduction

The number of healthcare workers involved in treating cancer patients in Thailand may still have a small ratio compared to the number of cancer patients. Thailand's population of 68 million is served by 927 government hospitals and 363 private hospitals [1]. However, access to medical care in rural areas still lags far behind that in the cities. There are 108 hospitals specifically for cancer patients nationwide, and they are not located in rural areas, causing patients to travel to the provinces for treatment. For cancer patients to receive treatment, they must have blood test results from their local lab/clinic/hospital before traveling to the cancer hospital in the cities so that medical staff can analyze the body's readiness and whether the patients' conditions can be treated according to the cancer treatment program. If the patients' criteria do not pass the requirements, they cannot get the treatment, and they must make an appointment to perform a blood test again after a specific time. The patients will have to come to the hospital several times. This can make traveling inconvenient for cancer patients who live in rural areas.

Therefore, the researchers developed a blood test reading system by applying OCR. To help patients submit blood test results from local hospitals near the residence to be submitted to doctors or caregiver nurses to check and make assessments for making an appointment when the blood test results meet the criteria, thus preventing patients from having to travel to the hospitals treated for cancer patients. The researchers studied OCR and image preprocessing to prepare the image of the blood test results with good quality for extracting blood test result data. Overview of the application, the patient/user agrees to the consent form, selects the blood test results image, and then uploads images to the storage. The image taken must not blur and have a good lighting condition. The image will not be uploaded if it does not meet the quality requirements of the application, and the application will alert and ask the patient/user to choose a new image to upload. The researchers study OCR tools to choose the suitable OCR to extract blood test results information. The OCR extracts text from the image, which is information about the blood test result, including the name of the test, result value, unit of result value, and reference range of the result value. Then analyze whether the test result is normal or relative to the reference criteria, and the severity of the symptoms was assessed according to the criteria of CTCAE Version 5.0 (Common Terminology Criteria for Adverse Events Version 5.0) [2]. This paper focuses on selecting an OCR algorithm for the cancer care mobile application.

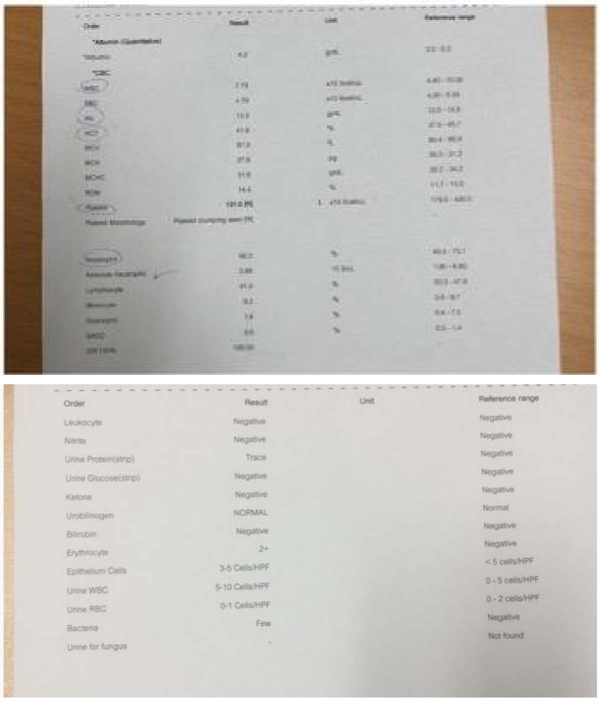
II. Related works

We designed a mobile application for reviewing patients' blood test results using OCR (Optical Character Recognition) to reduce the cause of traveling and time consumption. The application extracted information from the image of blood test results. OCR technology is used to convert virtually images containing text information into computer digitally encoded text. During the Covid-19 pandemic, OCR was applied to generate a medical report from handwriting documents [3]. OCR has played a role in reducing medical waste, which is the medicine that has been thrown away because it does not contain readable information on the medicine packages [4]. Moreover, to help patients understand their prescribed medicines before purchasing them, OCR was applied to read the doctor's handwriting in the medical prescription, then display the result, which is the name of the medicine, to the patients [5]. In [6], OCR is used to extract information from hard copy test reports and then predict the name of the diseases using with Bag of Words (BoW) model as feature selection algorithm, Naïve Bayes as classification algorithm, and AdaBoost technique to enhance the performance of the Naïve Bayes classifier [6].

III. Blood test results

Patients can take blood tests in a doctor's office, clinic, lab, or hospital located near their accommodations. The blood test result shows the value, unit, and reference range as shown in Fig. 1. For cancer care and treatment plan, medical staff, usually focus on 14 values of blood test results, including white blood cell, red blood cell, hemoglobin, platelet count, lymphocyte count, neutrophil count, creatinine, aspartate aminotransferase, alanine aminotransferase, alkaline phosphatase, blood bilirubin, hematocrit, absolute neutrophil, and serum creatinine.

978-616-590-477-3 ©2022 APSIPA



Order	Result	Unit	Reference range
*Albumin (Quantitative)			
*Albumin	4.2	g/dL	3.5 - 5.2
*CBC			
WBC	7.13	x10 3cells/uL	4.40 - 10.30
RBC	4.78	x10 6cells/uL	[illegible]
Hb	13.3	g/dL	12.0 - 14.9
HCT	41.8	%	37.0 - 45.7
MCV	87.3	fL	80.4 - 95.9
MCH	27.8	pg	25.0 - 31.2
MCHC	31.8	g/dL	30.2 - 34.2
RDW	14.4	%	11.7 - 15.0
Platelet	131.0 [R]	L x10 3cells/uL	170.0 - 435.0
Platelet Morphology	Platelet clumping seen [R]		-
Neutrophil	50.3	%	40.0 - 73.1
Absolute Neutrophil	3.89	10 3/uL	1.80 - 6.80
Lymphocyte	41.4	%	20.5 - 47.9
Monocyte	6.2	%	3.4 - 9.7
Eosinophil	1.6	%	0.4 - 7.5
BASO	0.5	%	0.2 - 1.4
Diff 100%	100.00		-

Order	Result	Unit	Reference range
Leukocyte	Negative		Negative
Nitrite	Negative		Negative
Urine Protein(strip)	Trace		Negative
Urine Glucose(strip)	Negative		Negative
Ketone	Negative		Negative
Urobilinogen	NORMAL		Normal
Bilirubin	Negative		Negative
Erythrocyte	2+		Negative
Epithelium Cells	3-5 Cells/HPF		< 5 cells/HPF
Urine WBC	5-10 Cells/HPF		0 - 5 cells/HPF
Urine RBC	0-1 Cells/HPF		0 - 2 cells/HPF
Bacteria	Few		Negative
Urine for fungus	-		Not found

Fig. 1 Example of blood test results.

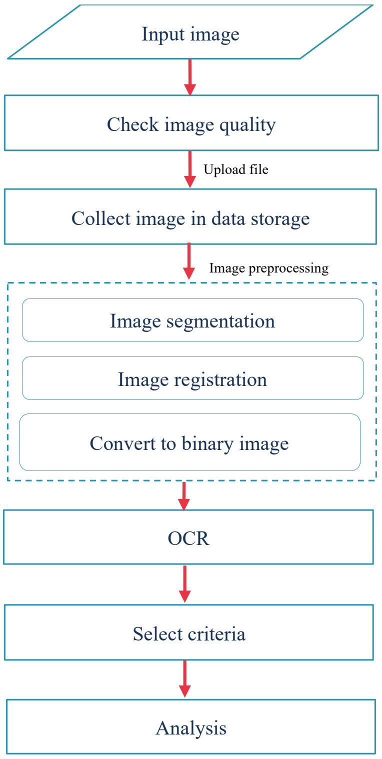


Fig. 2 System overview

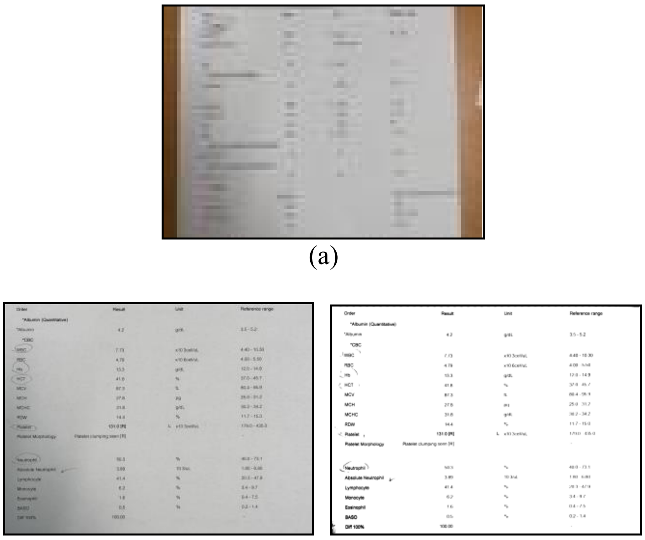
IV. Methodology

A. System Overview

Before using the application, the patients/user must agree on the consent form to allow the application to collect the blood test results images captured from their smartphone. After the user selects the input image, the application starts to process by checking image quality, sending image data to the data storage, and performing image preprocessing. Then the OCR tool performs to extract information from the image and uses that information to analyze the patient's condition according to CTCAE Version 5 [2]. The overall system is shown in Fig2.

While capturing images using a smartphone, some factors might cause the image blur, such as camera movement, subject movement, and missed focus. Blur images might reduce the performance of the OCR algorithm in the extraction of blood test results. To check the quality of the image, we calculate the variance of Laplacian of all pixels of the image and determine whether the test image passes the image quality test by using the blurriness score [7]; if the blurriness scores are less than 50, that image is not passed the checking process.

The image that passes the quality check will be sent to the data storage to prepares image preprocessing. The image preprocessing step prepared the image for the OCR algorithm. In the preprocessing, image segmentation is used to select the area that contains text information about test results using blob detection, then aligns the image and to transform it into a binary image. An example of the output of image preprocessing is shown in Fig3.



(b) (c)

Fig. 3 Output of the image preprocessing; a) original image, b) result of image segmentation and registration, c) binary image

Next, the OCR algorithm will perform to extract the test results for each line and collect data according to the 14 criteria along with the value, unit, and the reference range. The final step is to analyze whether the information is in the correct form and unit.

B. OCR Tools

To extract the blood test result from the captured image, we studied three open-source OCR tools; Keras-OCR, EasyOCR, and Tesseract-OCR.

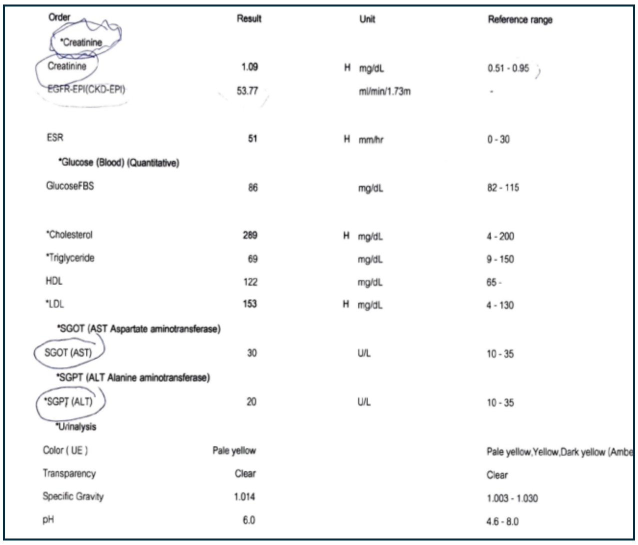
Keras-OCR has been applied to detect medicine information from blister packages to help pharmacies reduce the error that leads to patients receiving the wrong medicine [8]. In this work, we use the pre-trained mode, which is a CRNN model [9].

EasyOCR is an open-source, Ready-to-use OCR, one of the most widely used OCRs. The pre-trained recognition model is CRNN [10]. EasyOCR support more than 80 languages, including Thai.

Tesseract-OCR is one of the most famous open-source OCRs. Tesseract4 has implemented a Long Short Term Memory (LSTM) based recognition engine, which is RNN[11-12]. Tesseract-OCR has been studied to develop a system detecting invoice data [13].

V. Experimental Results

In this work, we studied three pre-trained models, Keras-OCR, EasyOCR, and Tesseract-OCR, to detect data from the image of blood test results. The blood test results image is segmented, aligned, and transformed into a binary image. The OCR tools detect characters using the binary image as an input image, and the output is the detected characters separated by a line. Example of blood test results in the binary image shown in Fig 4.



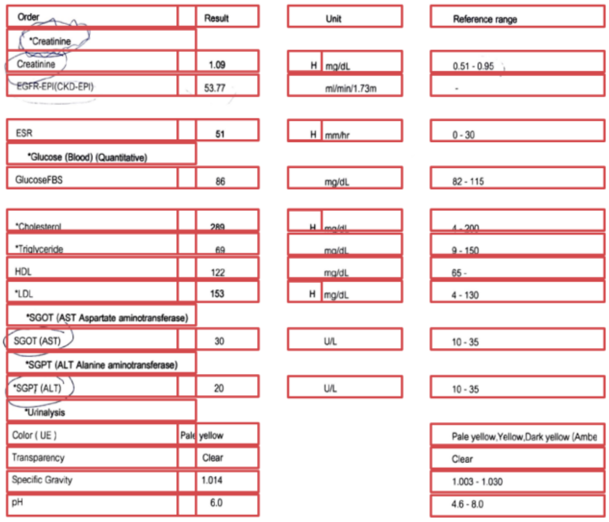
Order	Result		Unit	Reference range
*Creatinine				
Creatinine	1.09	H	mg/dL	0.51 - 0.95
EGFR-EPI(CKD-EPI)	53.77		ml/min/1.73m	-
ESR	51	H	mm/hr	0 - 30
*Glucose (Blood) (Quantitative)				
GlucoseFBS	86		mg/dL	82 - 115
*Cholesterol	289	H	mg/dL	4 - 200
*Triglyceride	69		mg/dL	9 - 150
HDL	122		mg/dL	65 -
*LDL	153	H	mg/dL	4 - 130
*SGOT (AST Aspartate aminotransferase)				
SGOT (AST)	30		U/L	10 - 35
*SGPT (ALT Alanine aminotransferase)				
*SGPT (ALT)	20		U/L	10 - 35
*Urinalysis				
Color (UE)	Pale yellow			Pale yellow,Yellow,Dark yellow (Ambe
Transparency	Clear			Clear
Specific Gravity	1.014			1.003 - 1.030
pH	6.0			4.6 - 8.0

Fig. 4 Binary image of blood test results

A. Blood test results detection

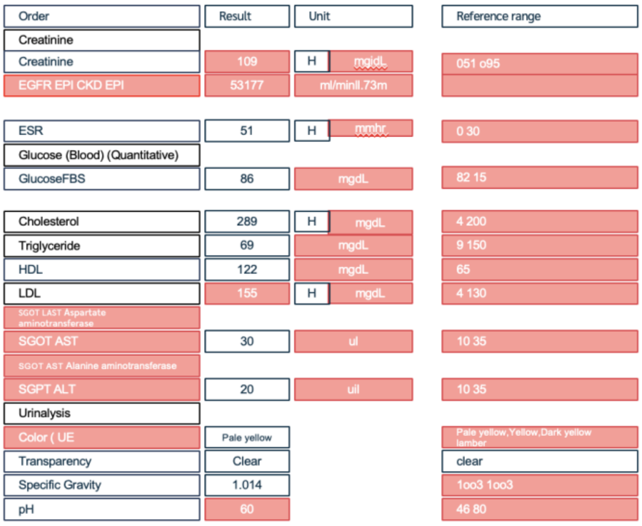
To measure OCR tools' performance, we determine the region of each word/value/range as a box, as shown in Fig 5. The blood test result image contains information about the test in four columns, including column Order, column Results, column Unit, and column Reference range. There are 65 boxes in the binary image, 20 boxes in the first column, and 15 in each remaining column.

Column Order is the test's name, containing upper case and lower case letters and special letters. Column Result shows the result value as a number, only the result of Color (UE) is shown as letters. Column Unit shows the unit of the result. In the last column, the Reference range shows the reference value of the result, and only the Color (UE) shown as a list of possible colors. If the OCR can correctly detect the characters in the box, that box will be a correct detection. Note that the special character "*" in this binary image is very small. If the OCR does not detect this special character at the box located in the first column and the remaining characters are correctly detected, that box will be a correct detection. The results of all three OCR tools, Keras-OCR, EasyOCR, and Tesseract-OCR, are shown in Fig 6 a), b), and c), respectively.



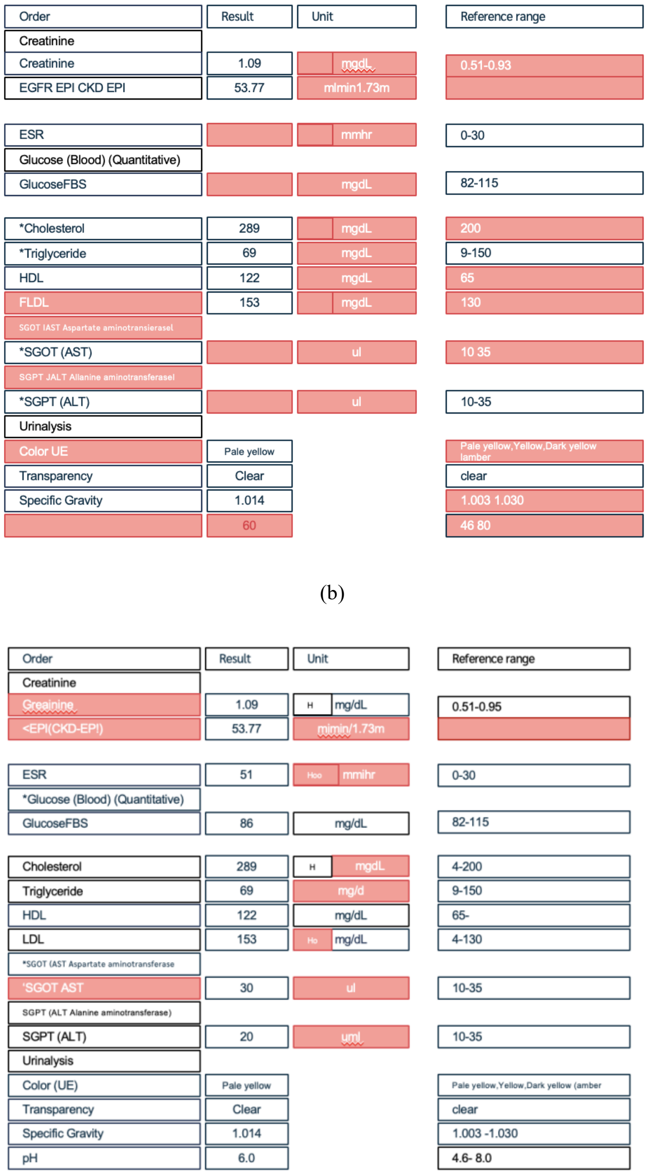
Order	Result		Unit	Reference range
*Creatinine				
Creatinine	1.09	H	mg/dL	0.51 - 0.95
EGFR-EPI(CKD-EPI)	53.77		ml/min/1.73m	-
ESR	51	H	mm/hr	0 - 30
*Glucose (Blood) (Quantitative)				
GlucoseFBS	86		mg/dL	82 - 115
*Cholesterol	289	H	mg/dL	4 - 200
*Triglyceride	69		mg/dL	9 - 150
HDL	122		mg/dL	65 -
*LDL	153	H	mg/dL	4 - 130
*SGOT (AST Aspartate aminotransferase)				
SGOT (AST)	30		U/L	10 - 35
*SGPT (ALT Alanine aminotransferase)				
*SGPT (ALT)	20		U/L	10 - 35
*Urinalysis				
Color (UE)	Pale yellow			Pale yellow,Yellow,Dark yellow (Ambe
Transparency	Clear			Clear
Specific Gravity	1.014			1.003 - 1.030
pH	6.0			4.6 - 8.0

Fig. 5 Box of word/value/range total 65 boxes.



Order	Result	Unit		Reference range
Creatinine				
Creatinine	109	H	mgidL	051 o95
EGFR EPI CKD EPI	53177	ml/minll.73m		
ESR	51	H	mmhr	0 30
Glucose (Blood) (Quantitative)				
GlucoseFBS	86	mgdL		82 15
Cholesterol	289	H	mgdL	4 200
Triglyceride	69	mgdL		9 150
HDL	122	mgdL		65
LDL	155	H	mgdL	4 130
SGOT LAST Aspartate aminotransferase				
SGOT AST	30	ul		10 35
SGOT AST Alanine aminotransferase				
SGPT ALT	20	uil		10 35
Urinalysis				
Color (UE	Pale yellow			Pale yellow,Yellow,Dark yellow lamber
Transparency	Clear			clear
Specific Gravity	1.014			1oo3 1oo3
pH	60			46 80

(a)



Order	Result	Unit	Reference range
Creatinine			
Creatinine	1.09	mgdL	0.51-0.93
EGFR EPI CKD EPI	53.77	mlmin1.73m	
ESR		mmhr	0-30
Glucose (Blood) (Quantitative)			
GlucoseFBS		mgdL	82-115
*Cholesterol	289	mgdL	200
*Triglyceride	69	mgdL	9-150
HDL	122	mgdL	65
FLDL	153	mgdL	130
SGOT IAST Aspartate aminotransierasel			
*SGOT (AST)		ul	10 35
SGPT JALT Allanine aminotransferasel			
*SGPT (ALT)		ul	10-35
Urinalysis			
Color UE	Pale yellow		Pale yellow,Yellow,Dark yellow lamber
Transparency	Clear		clear
Specific Gravity	1.014		1.003 1.030
	60		46 80

(b)

Order	Result	Unit	Reference range
Creatinine			
Greainine	1.09	H mg/dL	0.51-0.95
<EPI(CKD-EP!)	53.77	mimin/1.73m	
ESR	51	Hoo mmihr	0-30
*Glucose (Blood) (Quantitative)			
GlucoseFBS	86	mg/dL	82-115
Cholesterol	289	H mgdL	4-200
Triglyceride	69	mg/d	9-150
HDL	122	mg/dL	65-
LDL	153	Ho mg/dL	4-130
*SGOT (AST Aspartate aminotransferase			
'SGOT AST	30	ul	10-35
SGPT (ALT Alanine aminotransferase)			
SGPT (ALT)	20	uml	10-35
Urinalysis			
Color (UE)	Pale yellow		Pale yellow,Yellow,Dark yellow (amber
Transparency	Clear		clear
Specific Gravity	1.014		1.003 -1.030
pH	6.0		4.6- 8.0

(c)

Fig. 6 Result of character detection in each box; a) Keras-OCR, b) EasyOCR, and c) Tesseract-OCR

In Fig 6, a), b), and c), the detected characters are shown in each boxes. The box filled with red is a false detection box, meaning that at least one character in the boxes is incorrectly detected. The total number of correct detection box are shown in Table 1.

From Table1, Tesseract-OCR achieved the highest accuracy in all columns and provided 81.5% accuracy. Keras-OCR and EasyOCR both provided 49.2% of accuracy.

Table 1 Experimental results

Test Column	Keras OCR		EasyOCR		Tesseract OCR	
	correct	%acc	correct	%acc	correct	%acc
Column order 20 boxes	14	70%	15	75%	17	85%
Column result 15 boxes	11	73%	10	66%	15	100%
Column unit 15 boxes	5	33%	1	6%	7	43%
Column reference range 15 boxes	2	13%	6	40%	14	93%
Total 65 boxes	32	49.2%	32	49.2%	53	81.5%

B. Result discussion

Fig 6 shows that all three OCR tools can perform well in detection letters, but Keras-OCR and EasyOCR detect special characters incorrectly in many boxes, which are important to extract information from the blood test results such as "/", "." and "-". This error might cause by the size of the character.

VI. Conclusions

This paper studies the performance of three pre-trained models of OCR tools, Keras-OCR, EasyOCR, and Tesseract-OCR, to extract information from the image of blood test results to be the part of cancer care applications. The experimental result shows that the Tesseract-OCR provided the highest accuracy, with 81.5% accuracy.

For the future research, we desire to design the post-processing to predict the word in each text box. This improvement might increase the accuracy of text reading. Furthermore, we will use the extracted text to grade the condition of the patients according to the CTCAE Version 5.0.

Acknowledgment

This paper is supported by Faculty of Engineering, Naresuan University [grant number R2564E013].

References

[1] R. Thanyanan, S. Oranratnachai, P. Puataweepong, V. Tangsujaritvijit,and P. Cherntanomwong, "Lung cancer in Thailand." in *Journal of Thoracic Oncology* 15.11 (2020): 1714-1721.

[2] National Institutes of Health. "Common Terminology Criteria for Adverse Events (CTCAE), version 5.0 (2017)." (2021).

[3] S. Karthikeyan, A. G. S. de Herrera, F. Doctor, and A. Mirza, "An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2574-2581, May 2022, doi: 10.1109/TCSVT.2021.3087641.

[4] I S. Godbole, D. Joijode, K. Kadam and S. Karoshi, “Detection of Medicine Information with Optical Character Recognition using Android,” 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), 2020, pp. 1-6, doi: 10.1109/B-HTC50970.2020.9298016.

[5] E. Hassan, H. Tarek, M. Hazem, S. Bahnacy, L. Shaheen and W. H. Elashmwai, “Medical Prescription Recognition using Machine Learning,” 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 0973-0979, doi: 10.1109/CCWC51732.2021.9376141.

[6] W. A. Qader and M. M. Ameen, “Diagnosis of Diseases from Medical Check-up Test Reports Using OCR Technology with BoW and AdaBoost algorithms,” 2019 International Engineering Conference (IEC), 2019, pp. 205-210, doi: 10.1109/IEC47844.2019.8950605.

[7] L. Mary Francis, and N. Sreenath, “Pre-processing Techniques for Detection of Blurred Images”, Proceedings of International Conference on Computational Intelligence and Data Engineering, 2019, pp 59–66

[8] N. Maitrichit and N. Hnoohom, “Intelligent Medicine Identification System Using a Combination of Image Recognition and Optical Character Recognition,” 2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2020, pp. 1-5, doi: 10.1109/iSAI-NLP51646.2020.9376816.

[9] Keras-OCR, “GitHub”, [Online]. Available: https://github.com/faustomorales/keras-ocr. [Access 31 July 2022]

[10] EasyOCR. “GitHub”, [Online]. Available: https://github.com/JaidedAI/EasyOCR.git. [Access 31 July 2022].

[11] Tesseract. “GitHub”, [Online]. Available: https://github.com/tesseract-ocr/tesseract [Access 31 July 2022].

[12] R. Smith, “An overview of the tesseract ocr engine,” in Ninth international conference on document analysis and recognition (ICDAR 2007), vol. 2, pp. 629–633, IEEE, 2007.

[13] V. N. Sai Rakesh Kamisetty, B. Sohan Chidvilas, S. Revathy, P. Jeyanthi, V. M. Anu and L. Mary Gladence, “Digitization of Data from Invoice using OCR,” 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022, pp. 1-10, doi: 10.1109/ICCMC53470.2022.9754117.