applied sciences

MDPI

Article

CIMI: Classify and Itemize Medical Image System for PFT Big Data Based on Deep Learning

Tong Min Kim [1], Seo-Joon Lee [2], Hwa Young Lee [3], Dong-Jin Chang [4], Chang Ii Yoon [5], In-Young Choi [2,*] and Kun-Ho Yoon [6,*]

1 Department of Biomedicine & Health Sciences, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea; dianakim@catholic.ac.kr
2 Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea; 22001362@cmcnu.or.kr
3 Division of Pulmonology and Allergy, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea; lehwyo@catholic.ac.kr
4 Department of Ophthalmology and Visual Science, The Catholic University of Korea College of Medicine, Seoul 06591, Korea; hpalways@catholic.ac.kr
5 Graduate School of Medicine, The Catholic University of Korea, Seoul 06591, Korea; ckddlf2073@catholic.ac.kr
6 Division of Endocrinology and Metabolism, Department of Internal Medicine, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea
* Correspondence: iychoi@catholic.ac.kr (I.-Y.C.); yoonk@catholic.ac.kr (K.-H.Y.)

Received: 10 November 2020; Accepted: 27 November 2020; Published: 30 November 2020

Abstract: The value of pulmonary function test (PFT) data is increasing due to the advent of the Coronavirus Infectious Disease 19 (COVID-19) and increased respiratory disease. However, these PFT data cannot be directly used in clinical studies, because PFT results are stored in raw image files. In this study, the classification and itemization medical image (CIMI) system generates valuable data from raw PFT images by automatically classifying various PFT results, extracting texts, and storing them in the PFT database and Excel files. The deep-learning-based optical character recognition (OCR) technology was mainly used in CIMI to classify and itemize PFT images in St. Mary's Hospital. CIMI classified seven types and itemized 913,059 texts from 14,720 PFT image sheets, which cannot be done by humans. The number, type, and location of texts that can be extracted by PFT type are all different, but CIMI solves this issue by classifying the PFT image sheets by type, allowing researchers to analyze the data. To demonstrate the superiority of CIMI, the validation results of CIMI were compared to the results of the other four algorithms. A total of 70 randomly selected sheets (ten sheets from each type) and 33,550 texts were used for the validation. The accuracy of CIMI was 95%, which was the highest accuracy among the other four algorithms.

Keywords: artificial intelligence; deep learning; big data; medical image; image processing; optical character recognition

1. Introduction

Respiratory disease is one of the leading causes of death worldwide [1,2] long before the advent of Coronavirus Infectious Disease 19 (COVID-19) [3], the main symptom of which is lung failure. According to a report by the World Health Organization (WHO), the top five major causes of lung related severe illness are chronic obstructive pulmonary disease (COPD), asthma, acute lower respiratory tract infections, tuberculosis (TB), and lung cancer [4]. According to their statistics, each year, 4 million people die prematurely from chronic respiratory disease [5], with infants and young children particularly susceptible [6].

Recently, due to the COVID-19 epidemic, many studies have been conducted around the world to predict the occurrence of cardiopulmonary diseases [7]. Artificial intelligence (AI) [8] has been rapidly being applied to healthcare such as pharmaceuticals [9] and precision medicine [10]. Its application to medical imaging in the recent medical industry [11] is especially used in terms of lung disease analysis, because most of the medical data obtained from lung examinations include medical images.

Standardized big data related to lung examinations are required for such studies to be properly conducted. Many data related to pulmonary functions in the past were stored in hard copies. Even if the data were stored in electronic medical records (EMRs), they were stored either in a non-standardized format or in such a rudimentary form from which it is impossible to conduct proper statistical analysis. This makes it difficult for researchers to make proper use of lung-related big data for applied research.

Therefore, to solve this problem, this study proposed a deep-learning-based medical image AI processing system to classify and itemize medical image (CIMI). The proposed AI algorithm in CIMI not only classifies and itemizes medical images obtained from lung testing, but also de-identifies information for security purposes and reads medical images in a standardized medical data form for big data or AI researchers to use. The actual pulmonary function testing (PFT) data obtained from patients at Seoul St. Mary's Hospital were used for evaluation. CIMI is envisioned to be used as an impactful tool for future research on the big data pulmonary disease analysis and future solutions to predict respiratory diseases.

2. Related Work

CIMI is developed based on optical character recognition (OCR) and computer vision (CV) technologies. Related technologies regarding this research are specified in Section 2.1 "Related Technology". Studies similar to CIMI are covered in Section 2.2 "Related Research", describing what distinguishes CIMI from these studies.

2.1. Related Technology

Optical character recognition (OCR) technology enables images, PDF files, or other types of files (if they include texts) to be conveniently converted into machine-encoded format texts [12]. With OCR being applied in various fields such as education, medicine, law, etc. [13–16], this research is mainly based on Tesseract optical character recognition. Tesseract OCR is an open-source-based optical character recognition engine that was developed by Hewlett-Packard (HP) as a result of 10 years of development from 1984 to 1994. It was first started as a research project at Bristol HP Labs. Its prototype was sent to the University of Nevada, Las Vegas, in 1995 for yearly testing, where its performance results greatly proved its worth [17]. By the end of 2005, HP was able to officially launch an open-source-based Tesseract that is now available (web link: http://code.google.com/p/tesseract-ocr). Tesseract's function is mainly consisted of "Line and Word Finding" algorithm, "Word Recognition" algorithm, "Static Character Classifier", "Linguistic Analysis", and a more font-sensitive version of static character classifier "Adaptive Classifier".

Tesseract OCR technology is based on long short-term memory (LSTM) that is a recurrent neural network (RNN) deep learning algorithm. In RNN, hidden nodes are connected to each other through direction edges, forming a circulation architecture such as an artificial neural network [18]. This enables information to be contained, similar to the human brain. RNNs are especially being applied in fields such as voice recognition, language modeling, language translation, and image annotation creation [19]. However, the shortcomings of the RNN lie in its long-term dependencies [20]. Hochreiter and Schmidhuber proposed the idea of LSTM, which overcomes such shortcomings [21,22], and consequently, LSTM is being widely used in fields [23–25].

The Open Source Computer Vision Library (OpenCV) was first adopted by Intel in the early 2000s. As a result of many programmers contributing to enriching the library, it was years later that its true value shone. Its most updated version is "OpenCV 2" that was mainly modified in 2009. To date, over 2500 OCR optimized algorithms exist. It is also one of the most acknowledged libraries

worldwide [26], with more than 2.5 million downloads and over 40 thousand user experiences. In this research, Tesseract and OpenCV technologies were converged to maximize the accuracy of image processing. Our specifications are further explained in Section 3.

2.2. Related Research

Laique et al. [27] demonstrated a new hybrid approach using natural language processing of charts that has been elucidated with optical character recognition processing (OCR and Natural Language Processing, NPL hybrid) to obtain relevant clinical information from scanned colonoscopy and pathology reports. The limitation of this research was that their solution did not consider various types of PFT test sheets, and fundamentally, their solution does not extract text from image files.

Park et al. [28] proposed an automated method to construct a PFT database with various clinical information using optical character recognition and regular expression techniques. Pethidine-related patient case-control research data were used as samples. Their solution not only allowed anyone to easily construct a soft-copy database extracted from hard-copy test results, but also provided a de-identification technique to protect personal information.

Hinchcliff et al. [29] developed text data extraction for a prospective, research-focused data mart. They provided "Regextractor", an open source structured query language server integration services package that allows data extraction from PFT testing reports. The evaluation results showed that Regextractor successfully constructed a PFT data mart accurately extracted from the test charts that allowed clinical researchers and bioinformatics researchers to conduct the analysis.

However, although these recent studies may contribute to the future development of AI-based image processing solutions, no research has so far focused on itemizing past testing results. The proposed solution, CIMI, is not only compatible with future clinical test data, but also compatible with past accumulated data. This is one of the most unique implications of CIMI in the field of pulmonary clinical research.

Moreover, the shortcomings of the related research mentioned above were that none provided the graphic user interface (GUI) suitable for non-professionals, such as Excel-based database results. In contrast, our proposed research solution, CIMI, provides an application as simple and clear as possible for use by non-professionals. Clinicians do not need any engineering background to access the DB and can easily confirm the DB results in Excel format. In addition, the aforementioned studies simply used parsing techniques that only allow regular expression, whereas CIMI adopts a more dynamic technique based on coordinate measurement using mouse event handlers.

3. System Architecture

In this section, the overall system of CIMI will be first described. Then, details of the functional specifications of CIMI will be handled. Lastly, the proposed CIMI's database architecture will be explained.

3.1. Overall System

The flowchart of the overall system is shown in Figure 1. CIMI mainly consists of three major functions: the image classification function (blue squared), the data extraction function (red squared), and the database creation function (green squared). The first function that is the image classification function can be considered as the "pre-processing" process. Normally, the test results are divided into several different types, and the location of the data (text and image) differs according to these types. Rather than dumping the entire test sheet into the AI algorithm, CIMI first classifies test sheets into several types prior to algorithm processing to maximize efficiency. The data extraction function extracts data from the test result images [30]. The CIMI's algorithm is based on the Tesseract OCR (4.1.1, Google, Mountain View, CA, USA, 2019) library that adopts the AI LSTM method. The CIMI's main algorithm was also converged using OpenCV (4.1.1, Intel, 2200 Mission College Blvd, Santa Clara, CA, USA, 2019)-based techniques to maximize accuracy. The third major function finally creates a

database in the form of a highly compatible common-separated values (CSV) or Excel format. GUI was developed so that it was convenient for clinicians and researchers, and its application was developed based on PyQt5 (5.15.1, Riverbank Computing, Dorchester, UK, 2020).

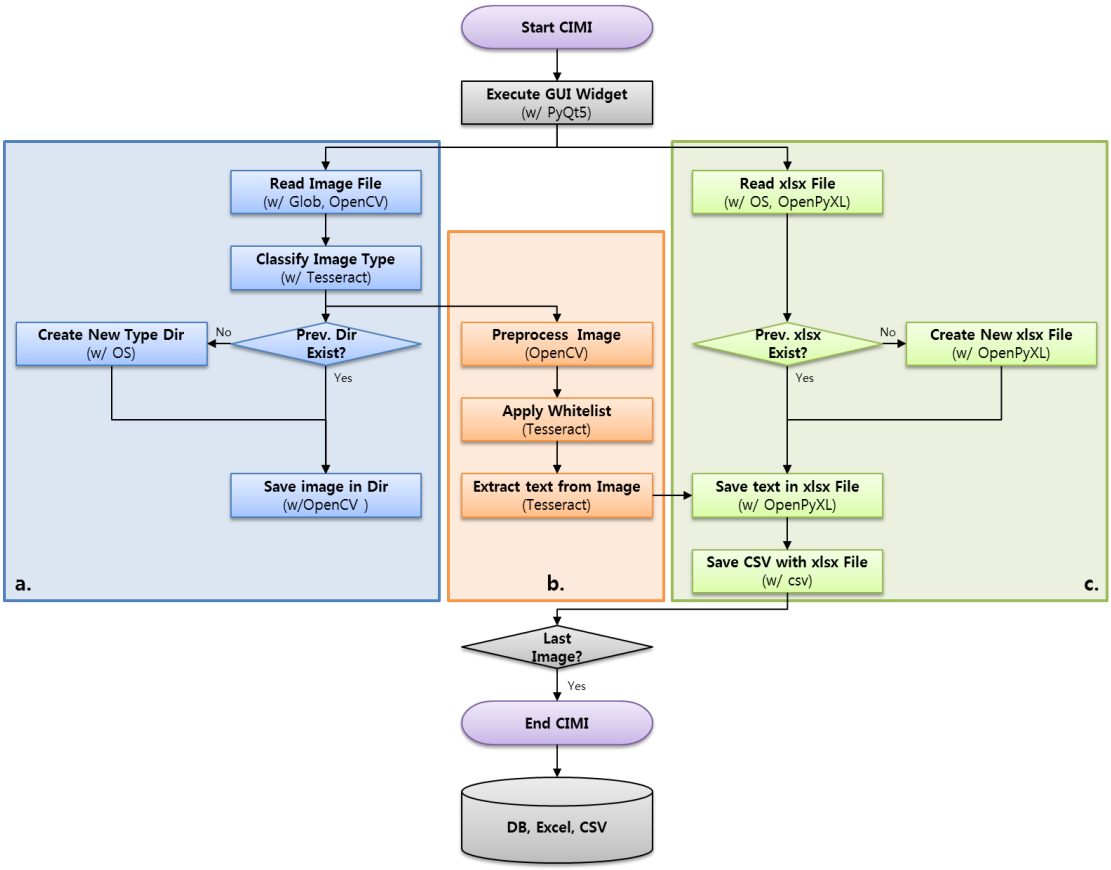

Figure 1. Flowchart of the overall classification and itemization medical image (CIMI) system; (**a**) PFT sheet classification systems; (**b**) PFT sheet preprocessing and PFT text extraction system; (**c**) PFT excel generation system.

When CIMI application starts, the original image-captured files of the medical test results are read via the OpenCV library (Figure 1a). Next, the CIMI's modified Tesseract algorithm classifies the results into several types (Figure 1b). Technically, classification here is defined as creating a new type of folder and moving the image result file from the original location to the destination. At the same time, CIMI sets a memory pointer to create a database in real time (OpenPyXL (web link: https://openpyxl.readthedocs.io/en/stable/) mainly used, Figure 1c). Now, the environmental settings are set. Next, the CIMI's OpenCV algorithm pre-processes the image results data to maximize reading accuracy. With the converged Tesseract algorithm, all characters are extracted and saved to the designated database created. This process is repeated until the CIMI AI reaches the last image file.

3.2. Major Function Specifications

In this section, the major functions of PFT image classification, data extraction, and database creation are explained.

The PFT image classification is shown in Figure 2. For the PFT image classification, the standards of the types were set by an actual professional clinician. Any type of test sheet that does not contain texts (for example, only graphs) were classified as "Unknown". The backbone of the CIMI classification algorithm was developed using the "keyword and coordinates" method according to the standard. The best representative keyword (Figure 2a) was set for each type and was unique to the classified

type, so that the word does not overlap with other types' keywords. After this keyword was set, the coordinates (Figure 2b,c) of this keyword were extracted by dragging the keyword's location with CIMI's mouse handler. Next, any random images that are input were classified according to the "keywords and coordinates". As long as this keyword's coordinates match, incoming image results were classified as the same type. The practical implementation results of this first function are shown in Section 4.

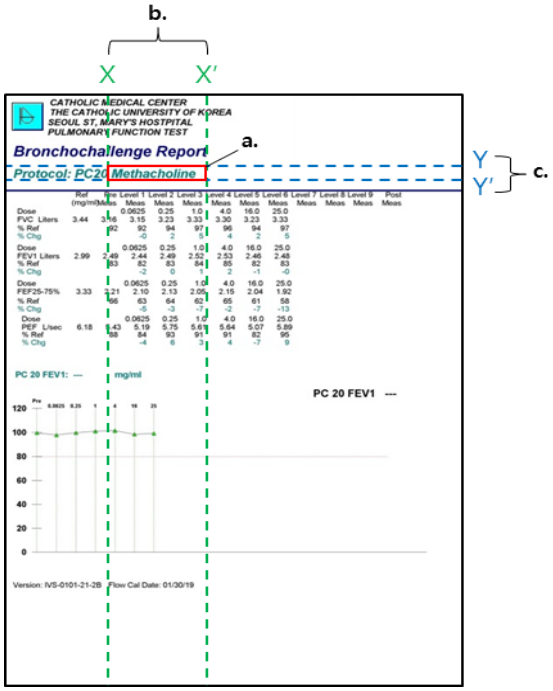

Figure 2. Keyword and coordinates; (**a**) the best representative keyword; (**b**) x coordinate values at the beginning and end of the keyword; (**c**) y coordinate values at the beginning and end of the keyword.

The data extractor functional process is shown in Figure 3. The first step of this functional algorithm is to extract images and coordinate values from the original PFT test sheets (Figure 3a). Next, the CIMI machine learning algorithm processes images into text (Figure 3b) and inputs the text into the newly created database according to column and row (Figure 3c,d, in this case as Excel). Finally, the steps are repeated from a to d until all data in the sheet are extracted.

Although Tesseract is a widely used solution with high accuracy, it was found that its accuracy further increases when pre-processed. There are several pre-processing methods such as de-noise, thresholding, dilate, erode, open, canny, and deskew. The CIMI AI's unique algorithm based on Tesseract technology was modified to maximize reading accuracy. Several methods were selectively converged in this proposed research to provide the best outcome optimized for PFT results sheets. Mainly, CIMI converged the gray scale method and rescale method. The proposed original pre-processing algorithm is shown in Figure 4. First, the original image was resized to twice the original size, and then, it was gray-scaled. Originally developed whitelists were applied per field, considering the characteristics of PFT that its test results normally do not contain special characters and mostly contain only alphabet characters. The related performance results will be further discussed in the results and discussion section.

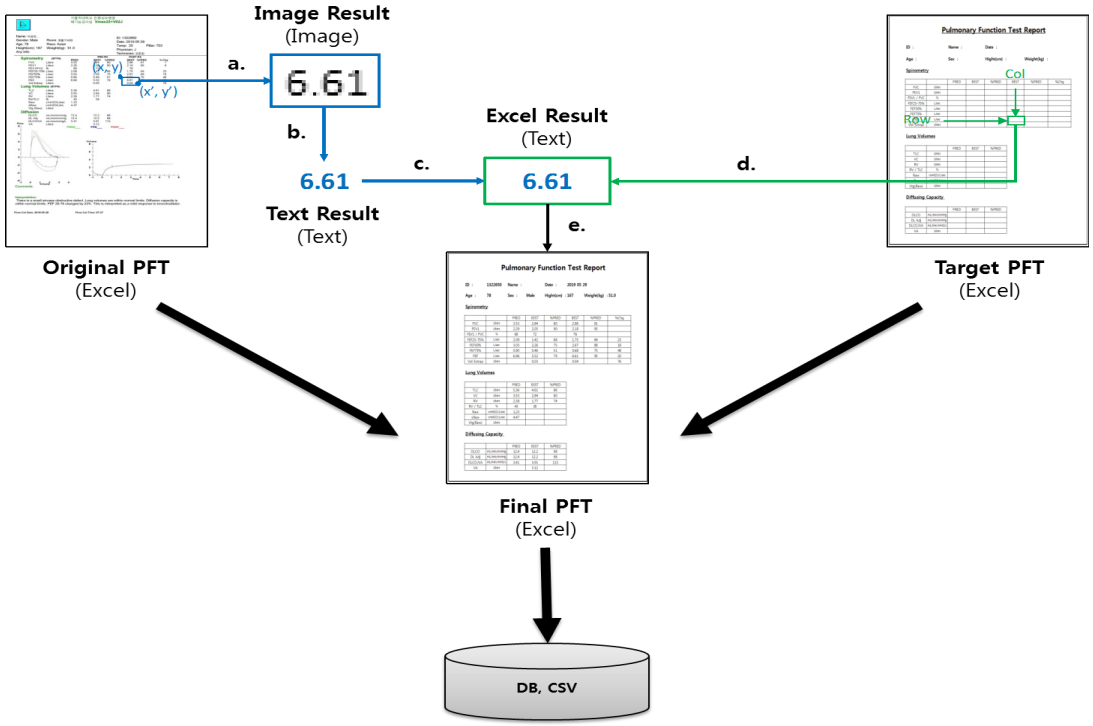

Figure 3. Data extractor algorithm functional process; (**a**) measure x and y coordinate values from the original PFT sheets; (**b**) extract text from the PFT sheet within corresponding coordinates; (**c**) insert the extracted text into the cell of corresponding excel; (**d**) get the row and column value from the target PFT excel file; (**e**) insert extracted text into the final PFT Excel file.

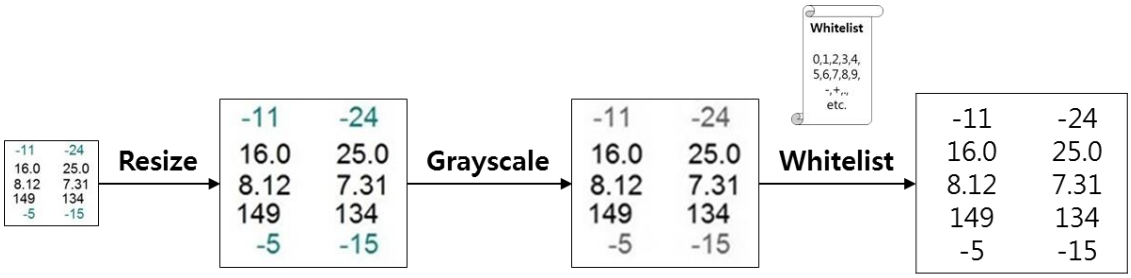

Figure 4. Optimized CIMI pre-processing algorithm for pulmonary function testing (PFT) test sheets.

The PFT image big data (sheets) are very dynamic in the sense that these data vary in types even if it is the same prescription. Additionally, PFT results contain numerous data (texts) per sheet. Therefore, an algorithm that automatically classifies and extracts big data texts within the big data sheets is essential. However, the speed of processing image data is much slower than the speed of processing text data. This issue is especially important in the case of processing PFT test results, because by nature, PFT test result sheets are numerous image files that take up a large memory space. Moreover, OCR itself takes a lot of central processing units (CPUs), memory, and disk space [31]; therefore, it is very important to use CPU parallel computing technology in order to maximize processing speed. The proposed CIMI solves this problem with its original simultaneous processing algorithm based on Python's "concurrent.futures" library that enables multi-core processing to maximize processing speed.

3.3. Database Architecture

The database creation function not only creates a text extracted version of the PFT testing result sheet, but also creates an overall database (with columns as variables and rows as case numbers) so that researchers may conduct statistical analysis. The database architecture of the proposed solution is

shown in entity relationship diagram (ERD) in Figure 5. The developed database was optimized for maximum efficiency when using the proposed solution. The database consists of six tables: a table that includes PFT test results (PFT_IMG_INFO), a table that includes patient information (PATIENT), and four other PFT test-related tables (SPIROMETRY, LUNG_VOLUMES, DIFFUSING_CAPACITY, and RESISTANCE).

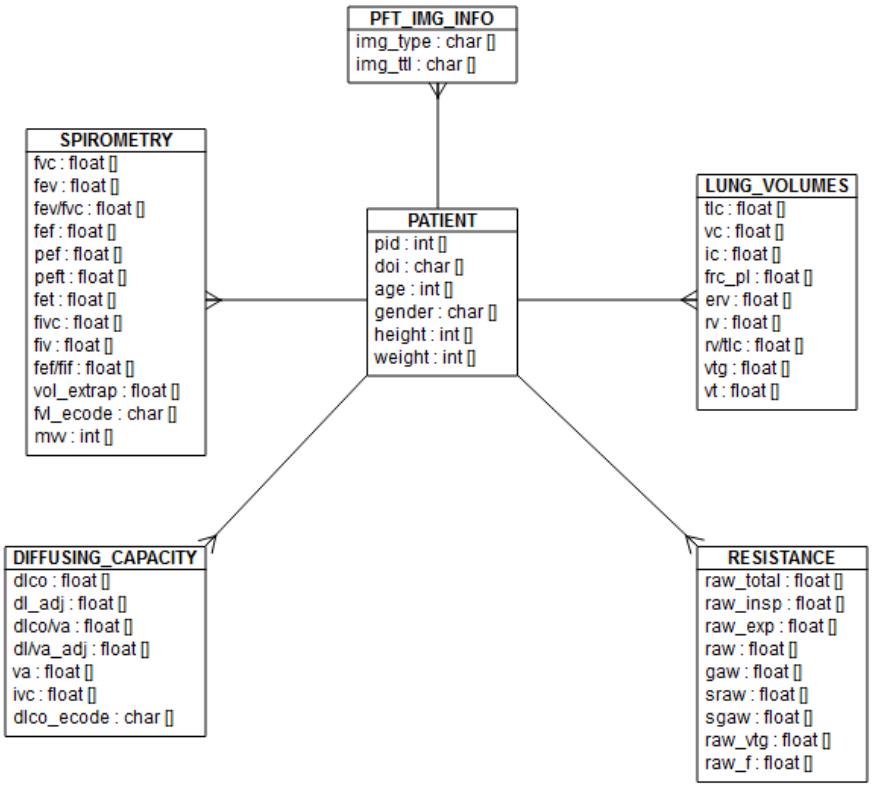

Figure 5. Entity relationship diagram (ERD) of the CIMI database architecture.

4. Evaluation

Actual PFT big data were used to evaluate the CIMI system for accuracy. Additionally, other algorithms were compared and analyzed to highlight the strength of CIMI. Details are described below.

4.1. Materials and Methods

A total of 14,720 sample sheets were used from St. Mary's Hospital, Seoul, South Korea. These samples consisted of PFT test results from actual patients that have been registered in the hospital's information system (including hard copies and soft copies), acquired from 1st January, 2019 to 31st December, 2019 (one year). All samples were input into the trained AI solution CIMI to show the implementation results to assess the performance of the major functions. All sampling and evaluation processes were approved by the Institutional Review Board (IRB) of St. Mary's Hospital, Seoul.

All 14,720 sheets consisted of eight types (including unknown type) that contained different numbers and sorts of texts. Note that unknown classified types were excluded from this research evaluation, because they only contained non-texts. On the other hand, sheets classified as type 1, 2, 3, 4, 5, 6, and 7 contained a sub-total of 116, 164, 21, 130, 37, 73, and 130 texts in each sheet, respectively. Related evaluation will be further discussed in Section 4.2.1.

In addition, the CIMI validation tool was originally developed for the scientific validation of accuracy performance (Figure 6). CIMI Validator allows efficient evaluation of text extraction accuracy. The user can load the extracted text data via the "Load Excel" button and set the path of the PFT sheet using the "Image Path" button (Figure 6a). "Load Image" button of Figure 6a randomly selects

a sheet, which is visually shown in Figure 6c. The texture of the sheet is then aligned in an editable format (Figure 6b). In every single text there is a "confirm" button, in which if the user simply clicks, it increases the "correct count" by one, and if the user modifies then clicks, it increases the "error count" by one (Figure 6d).

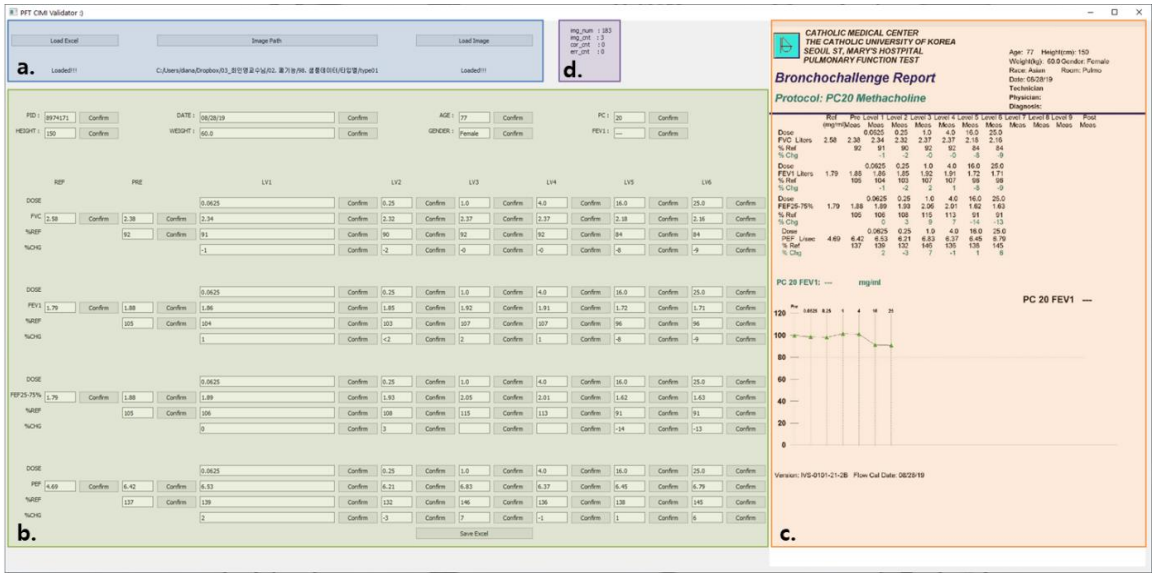

Figure 6. UX/UI screen of CIMI validation tool; (**a**) set the excel and the image file to validate; (**b**) the results from the excel; (**c**) the results of the image; (**d**) validation count information.

4.2. Performance Evaluation

First, CIMI's PFT type classification function results were handled. Then, CIMI's text extraction function accuracy was evaluated separately.

4.2.1. Type Classification Results

Specifically, there are several types of PFT test results that vary according to the test objectives. Therefore, type classification must be conducted before text extraction. The PFT-type classification results using CIMI are shown in Table 1. In summary, in the case of used samples, images were classified into seven types (excluding unknown type).

Table 1. Type classification results using CIMI.

Type	01	02	03	04	05	06	07	Unknown	Total
PFT Sheets	213	12	20	**5728**	1956	697	139	5955	14,720
Texts in Each Sheet	116	164	21	**130**	37	73	**130**	0	671
Extracted Texts	24,708	1968	420	**744,640**	72,372	50,881	18,070	0	913,059

The CIMI's classification results showed that type 4 had the highest number of sheets, with a total of 5728 sheets. The second most common type was type 5, with a total of 1956 sheets. The least number of sheets was found in type 2, with only 12. Type 2 contained the most texts in each sheet by 164, with type 4 and type 6 at second rank with 130 each. Results in Figure 7 also showed that some PFT types were similar to each other. The format of type 1 and type 2 were similar; type 3, type 4, and type 7 were similar; type 5 and type 6 were similar. These similarity classifications details are shown in Figure 7.

Figure 7. Samples of de-identified PFT sheets classified by CIMI system; (**a**) similar form of Type 1 and Type 2; (**b**) similar form of Type 3, Type 4 and Type 7; (**c**) similar form of Type 5 and Type 6.

4.2.2. Accuracy Performance Results

Each PFT type consisted of different number of data texts that were extracted: 116 data for type 1, 164 data for type 2, 21 data for type 3, 130 data for type 4, 37 data for type 5, 73 data for type 6, and 130 data for type 7. Ten sheets were randomly chosen by our developed validator for each of the seven types (n = 70) to validate the accuracy. In addition, four more widely used image processing algorithms were compared. Therefore, the researcher validated the total of 33,550 texts hands on. The vanilla algorithm refers to the pure Tesseract OCR. Resize and grayscale were pre-processed to vanilla for comparison. Finally, the whitelist-applied vanilla was used for comparison. The results are shown in Table 2.

CIMI showed the highest accuracy performance in type 5, with an average accuracy of 99.7%. It performed poorly in type 1 and type 2, in which the accuracy performance was 86 and 87%, respectively. The accuracy results of the vanilla, resize pre-processed, grayscale pre-processed, and whitelist-applied algorithm showed accuracies of 73, 89, 73, and 89%, respectively. Resizing the pre-processed algorithm and the whitelist-applied algorithm's accuracy were the highest contenders of CIMI, whereas Tesseract only and grayscale pre-processed showed the poorest performance. The only notable subtotal performance that was better than CIMI was for type 2 in the whitelist-applied algorithm, outperforming CIMI by 89 to 0.87%.

Table 2. Accuracy performance of CIMI and its comparison with other algorithms.

		Type01 (texts = 116)	Type02 (texts = 164)	Type03 (texts = 21)	Type04 (texts = 130)	Type05 (texts = 37)	Type06 (texts = 73)	Type07 (texts = 130)	Average Accuracy
CIMI	Correct (Average)	100.3	142.9	20.7	127.4	36.9	72.2	128.1	
	Omitted (Average)	9.7	19.0	0.1	1.2	0.0	0.0	0.1	
	Error (Average)	6.0	2.3	0.2	1.4	0.1	0.8	1.7	
	Accuracy (%)	86.5	87.1	98.6	98.0	99.7	98.9	98.5	95.3
Vanilla	Correct (Average)	83.3	122.5	10.8	106.4	34.0	59.8	76.3	
	Omitted (Average)	14.6	25.8	9.5	13.5	0.0	8.9	38.4	
	Error (Average)	18.7	15.3	0.7	10.1	3.0	4.3	15.2	
	Accuracy (%)	71.8	74.7	51.4	81.8	91.9	81.9	58.7	73.2
Resize Pre-processed	Correct (Average)	98.4	142.5	19.7	126.2	36.2	64.0	101.1	
	Omitted (Average)	7.8	18.4	1.0	1.8	0.5	9.0	16.8	
	Error (Average)	10.4	2.7	0.3	2.0	0.3	0.0	12.0	
	Accuracy (%)	84.8	86.9	93.8	97.1	97.8	87.7	77.8	89.4
Grayscale Pre-Processed	Correct (Average)	82.8	122.2	10.8	106.4	34.1	59.9	77.7	
	Omitted (Average)	14.0	25.2	9.5	13.5	0.0	8.9	37.1	
	Error (Average)	19.2	16.2	0.7	10.1	2.9	4.2	15.1	
	Accuracy (%)	71.4	74.5	51.4	81.8	92.2	82.1	59.8	73.3
Whitelist Applied	Correct (Average)	93.1	145	19.8	117.9	34.8	65.5	111.0	
	Omitted (Average)	4.0	0.9	0.1	3.4	0.0	0.8	1.9	
	Error (Average)	19.5	17.7	1.1	8.7	2.2	6.7	17.0	
	Accuracy (%)	80.3	88.4	94.3	90.7	94.1	89.7	85.4	89.0

5. Discussion and Conclusions

This study proposed CIMI, a deep-learning-based medical image AI processing system that classifies and extracts text from medical images. CIMI not only classifies PFT results, but also extracts medical images to text in a standardized medical data form for big data or AI researchers. We also provided a CIMI validation tool for related researchers to assess the accuracy of their algorithm.

The accuracy of Tesseract OCR, which is mainly used in CIMI, is not always constant and is greatly affected by the application of various pre-processes and whitelists [32]. The PFT used in our study was particularly affected by the number of characters, size, thickness, color, and margins between the letters. The higher the number of characters, the larger the size, the thicker the characters, and the greater the color difference between the characters and the background, the higher the accuracy. The margins between the letters showed greatest impact in accuracy when characters and margins were best distinguished. In other words, if the margins between characters were extremely high or small, the boundaries between characters became blurred, which resulted in lowered accuracy. Because of these characteristics, the same Tesseract OCR base algorithm showed different accuracy according to PFT type or algorithm. (Table 2)

Type 1 and type 2 format was found to be similar (Figure 7a). This resulted in little difference between the CIMI algorithm (86%) and vanilla algorithm (72%) accuracy. This was because the letter of these PFTs are thicker, larger in size, and have proper margins between the letters than other PFTs. These types generally showed lower accuracy than other types; for example, CIMI with a standard of 80%. This was because within these PFTs, there were more "%Chg" items that consist mostly of one word compared to other PFTs.

For types 3, 4, and 7, where format is similar to each other (Figure 7b), the accuracy difference between CIMI and vanilla, grayscale algorithms is large, because the letters in these PFTs are thin and small. Difference is most severe in type 3, moderately severe in type 7, less in type 4, because type 3 has the narrowest margins between letters, whereas type 7 has more "%Chg" items than other PFTs. Type 4 had a similar font size and thickness compared to type 3 and type 7, but the margins between the characters were wide, and there was no "%Chg" items, therefore showing a moderately severe difference.

For type 5 and type 6, which were similar (Figure 7c), despite the smallest letter size compared to other PFTs, the accuracy between CIMI and other algorithms were relatively small, because the letter

thickness was thicker than other PFTs, and the margins between letters is reasonable. The reason type 6 had lower accuracy than type 5 was because type 6 had more "%Chg" items than type 5.

Evaluation based on PFT data obtained from patients in St. Mary's Hospital, Seoul, showed successful classification of PFT sheets. The CIMI accuracy comparison results were superior to those of other widely used original algorithms, with an accuracy of 95.3%. This was the result of numerous systematic comparison and combination test attempts to create the CIMI's algorithm accuracy to be optimized in the field of PFT.

The limitation of this study was that CIMI was only applied to one medical institution for only one year (2019). Since medical data are prone to discrepancies among facilities, more data from other institutions and past/future years should be used and validate for future research. Moreover, the accuracy of CIMI was low for certain data types (especially in cases of type 1 and type 2); therefore, future investigation is needed to further enhance accuracy. CIMI is envisioned to be used as an impactful tool for future research on big data pulmonary disease analysis and future solutions to predict respiratory diseases.

Author Contributions: Conceptualization, T.M.K. and S.-J.L.; Data curation, T.M.K.; Formal analysis, T.M.K. and S.-J.L.; Funding acquisition, I.-Y.C.; Investigation, T.M.K.; Methodology, T.M.K., D.-J.C.; Project administration, I.-Y.C.; Resources H.Y.L.; Software, T.M.K.; Supervision, S.-J.L., I.-Y.C. and K.-H.Y.; Validation C.I.Y.; Visualization, T.M.K.; Writing—original draft, T.M.K.; Writing—review & editing, T.M.K. and S.-J.L. All authors have read and agree to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A5A2027588).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gibson, G.J.; Loddenkemper, R.; Lundbäck, B.; Sibille, Y. Respiratory health and disease in Europe: The new European Lung White Book. *Eur. Respir. Soc.* **2013**, *42*, 559–563. [CrossRef] [PubMed]
2. Bowen, T.S.; Aakerøy, L.; Eisenkolb, S.; Kunth, P.; Bakkerud, F.; Wohlwend, M.; Ormbostad, A.M.; Fischer, T.; Wisloff, U.; Schuler, G.; et al. Exercise Training Reverses Extrapulmonary Impairments in Smoke-exposed Mice. *Med. Sci. Sports Exerc.* **2017**, *49*, 879–887. [CrossRef] [PubMed]
3. Hadi, A.G.; Kadhom, M.; Hairunisa, N.; Yousif, E.; Mohammed, S.A. A Review on COVID-19: Origin, Spread, Symptoms, Treatment, and Prevention. *Biointerface Res. Appl. Chem.* **2020**, *10*, 7234–7242.
4. World Health Organization. *The Global Impact of Respiratory Disease*, 2nd ed.; World Health Organization: Geneva, Switzerland, 2017.
5. World Health Organization. *Global Status Report on Noncommunicable Diseases 2014*; no. WHO/NMH/NVI/15.1; World Health Organization: Geneva, Switzerland, 2014.
6. Galode, F.; Dournes, G.; Chateil, J.-F.; Fayon, M.; Collet, C.; Bui, S. Impact at school age of early chronic methicillin-sensitiveStaphylococcus aureusinfection in children with cystic fibrosis. *Pediatr. Pulmonol.* **2020**, *55*, 2641–2645. [CrossRef]
7. Wang, C.; Qi, Y.; Zhu, G. Deep learning for predicting the occurrence of cardiopulmonary diseases in Nanjing, China. *Chemosphere* **2020**, *257*, 127176. [CrossRef]
8. Vergeer, M. Artificial Intelligence in the Dutch Press: An Analysis of Topics and Trends. *Commun. Stud.* **2020**, *71*, 1–20. [CrossRef]
9. Zhavoronkov, A.; Vanhaelen, Q.; Oprea, T.I. Will Artificial Intelligence for Drug Discovery Impact Clinical Pharmacology? *Clin. Pharmacol. Ther.* **2020**, *107*, 780–785. [CrossRef]
10. Ahmed, Z.; Mohamed, K.; Zeeshan, S.; Dong, X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database J. Biol. Databases Curation* **2020**, *2020*. [CrossRef]
11. Willemink, M.J.; Koszek, W.A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L.R.; Summers, R.M.; Rubin, D.L.; Lungren, M.P. Preparing Medical Imaging Data for Machine Learning. *Radiology* **2020**, *295*, 4–15. [CrossRef]

12. Tappert, C.C.; Suen, C.Y.; Wakahara, T. The state of the art in online handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 787–808. [CrossRef]
13. Zanibbi, R.; Blostein, D. Recognition and retrieval of mathematical expressions. *Int. J. Doc. Anal. Recognit.* **2012**, *15*, 331–357. [CrossRef]
14. Thompson, P.; Batista-Navarro, R.T.; Kontonatsios, G.; Carter, J.; Toon, E.; McNaught, J.; Timmermann, C.; Worboys, M.; Ananiadou, S. Text Mining the History of Medicine. *PLoS ONE* **2016**, *11*, e0144717. [CrossRef] [PubMed]
15. Ashley, K.D.; Bridewell, W. Emerging AI & Law approaches to automating analysis and retrieval of electronically stored information in discovery proceedings. *Artif. Intell. Law* **2010**, *18*, 311–320.
16. Memon, J.; Sami, M.; Khan, R.A.; Uddin, M. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access* **2020**, *8*, 142642–142668. [CrossRef]
17. Smith, R. An overview of the Tesseract OCR engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Parana, 23–26 September 2007; Volume 2, pp. 629–633.
18. Socher, R.; Lin, C.C.-Y.; Ng, A.Y.; Manning, C.D. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the ICML—28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011.
19. Fabbri, M.; Moro, G. Dow Jones Trading with Deep Learning: The Unreasonable Effectiveness of Recurrent Neural Networks. In Proceedings of the 7th International Conference on Data Science, Technologies and Applications (DATA 2018), Porto, Portugal, 26–28 July 2018; pp. 142–153.
20. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef]
21. Olah, C. Understanding lstm Networks. 2015. Available online: http//colah.github.io/posts/2015-08-Understanding-LSTMs (accessed on 30 November 2020).
22. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
23. Sundermeyer, M.; Schlüter, R.; Ney, H. LSTM neural networks for language modeling. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
24. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. In Proceedings of the 9th International Conference on Artificial Neural Networks—ICANN '991999, Edinburgh, UK, 7–10 September 1999.
25. Gers, F.A.; Schraudolph, N.N.; Schmidhuber, J. Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **2002**, *3*, 115–143.
26. Culjak, I.; Abram, D.; Pribanic, T.; Dzapo, H.; Cifrek, M. A brief introduction to OpenCV. In Proceedings of the 35th International Convention MIPRO, Opatija, Croatia, 21–25 May 2012; pp. 1725–1730.
27. Laique, S.N.; Hayat, U.; Sarvepalli, S.; Vaughn, B.; Ibrahim, M.; McMichael, J.; Qaiser, K.N.; Burke, C.; Bhatt, A.; Rhodes, C.; et al. Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports. *Gastrointest. Endosc.* **2020**. [CrossRef]
28. Park, M.Y.; Park, R.W. Construction of an PFT database with various clinical information using optical character recognition and regular expression technique. *J. Internet Comput. Serv.* **2017**, *18*, 55–60.
29. Hinchcliff, M.; Just, E.; Podlusky, S.; Varga, J.; Chang, R.W.; Kibbe, W.A. Text data extraction for a prospective, research-focused data mart: Implementation and validation. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, 1–7. [CrossRef]
30. González, D.R.; Carpenter, T.; van Hemert, J.I.; Wardlaw, J. An open source toolkit for medical imaging de-identification. *Eur. Radiol.* **2010**, *20*, 1896–1904. [CrossRef] [PubMed]
31. Rybalkin, V.; Wehn, N.; Yousefi, M.R.; Stricker, D. Hardware architecture of bidirectional long short-term memory neural network for optical character recognition. In Proceedings of the Design, Automation & Test in Europe Conference & Exhibition, Lausanne, Switzerland, 27–31 March 2017; pp. 1390–1395.

32. Koistinen, M.; Kettunen, K.; Kervinen, J. How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine. In Proceedings of the 8th Language and Technology Conference LTC, Poznan, Poland, 17–19 November 2017; pp. 279–283.