The 10th IEEE International Conference on E-Health and Bioengineering - EHB 2022
Grigore T. Popa University of Medicine and Pharmacy, Iasi, Romania, November 17-18, 2022

Smart Scan of Blood Test Documents to be Integrated in a mHealth Application

Pedro Lobo[1], João Vilaça[2], Helena Torres[3], Bruno Oliveira[4] and Alberto Simões[5]
2AI – Applied Artificial Intelligence, IPCA – Polytechnic Institute of Cávado and Ave, Barcelos, Portugal,
pjlobo@ipca.pt[1], jvilaca@ipca.pt[2], htorres@ipca.pt[3], boliveira@ipca.pt[4], asimoes@ipca.pt[5]

***Abstract*—Controlling the physiological evolution of people plays an important role in the prevention and detection of health problems. Often, the medical analyzes performed by people are stored in documents that in the long term tend to be lost, suffer wear and tear and most importantly, do not facilitate the comparison of data to assess the physiological evolution. This work proposes an intelligent scanning system for blood test documents. This type of component can be found in digital (Digitally created PDFs, scanned PDFs and images) or physical format. It is intended to extract from this type of documents only the essential information, consisting of the blood components and their concentration. Optical Character Recognition (OCR) Machine Learning (ML) Kit was used to extract relevant information from documents in the case of scanned PDF's (SPDFs), images, and physical documents, and Apache's PdfBox for digitally created PDF's (DCPDFs). To filter the data and associate the blood constituents detected with the respective concentrations, a condition tree was developed. In the end, the methods used were able to detect an average of 95.38% of the blood compounds present in the different document formats. On average, 87.63% of the concentrations were correctly associated with the detected compounds.**

Keywords—mHealth, OCR, Data Collection, Blood Tests .

I. INTRODUCTION

In recent times, we have witnessed the massive adhesion by institutions and companies to the phenomenon of the digitalization era. However, despite the numerous advantages of digitally storing documents, the problem of consumed memory is increasingly taking place. The solution is not to stop digitizing information, but a more selective digitization.

Several works have been developed to explore medical information digitization tools to integrate mobile health (m-health) applications. Erin Sarzynski et al. developed a patient-centered medication management application, which uses an OCR tool that auto-populates drug names and dosing instructions directly from patients' medication labels [1]. During the work, numerous tests were carried out, including the adherence of the volunteers to the application for 6 months. After the first month, the adherence was 93% and at the end of the experience it was 78%.

The need to digitize medical documents selectively by detecting the main information in the documents is transversal at the individual and institutional level, despite having different motivations. Mehul Gupta and Kabir Soeny developed a work in which they focus on the digitization of printed and handwritten medical prescriptions of medicines [2]. The objective was to identify the prescribed drugs. Drug identification can later be used to provide detailed information on usage and Side Effects as presented in the work by Saumitra Godbole et al. [3]. Dinuka Kulathunga et al. also explored, in addition to medical prescriptions, blood analysis documents and extraction of the main information [4].

Still in the automatic reading of medical prescriptions, another interesting work arises motivated by medication consumption problems by elderly people. Roisul Islam Rumi et al. developed a work in which we proposed an intelligent Medicine Dispenser capable of reading medical prescriptions [5]. From the prescriptions read, it would detect the medicines listed therein and would give the user the respective ones.

In a different reality, but with similar purposes, the work of Sungrim Moon et al. explores the fact that some medical institutions share medical reports by fax among themselves for analysis purposes [6]. These documents are generally long and do not have searchable text. To reduce time spent and analysis errors by health professionals, Sungrim Moon et al developed a tool responsible for extracting text from them and identifying the most relevant information for analysis purposes.

Many times, medical analysis in documents may not have the necessary impact to clarify, alert or reassure people of their real state of health. Or on the other hand, there may be so many analysis documents that make the process of analyzing the health state difficult. Nan Liu el al. proposed a solution that consisted of digitalizing medical information from numerous sources (images and documents) so that they could be presented in didactic graphic environments that represent the physiological evolution of people [7].

The work developed in this paper will focus on a type of document that has been little explored, blood tests. Many of the works presented above show similarities in terms of objectives, and approach methods that can be very useful for my work. However, none meets the needs of this work.

This selective scanning process, presented in this paper, will also bring many advantages in monitoring physiological evolution for exemple in form of graphs over time as is already normal in mobile health (mhealth) applications.

2022 E-Health and Bioengineering Conference (EHB) | 978-1-6654-8557-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/EHB55594.2022.9991279

978-1-6654-8557-9/22/$31.00 ©2022 IEEE

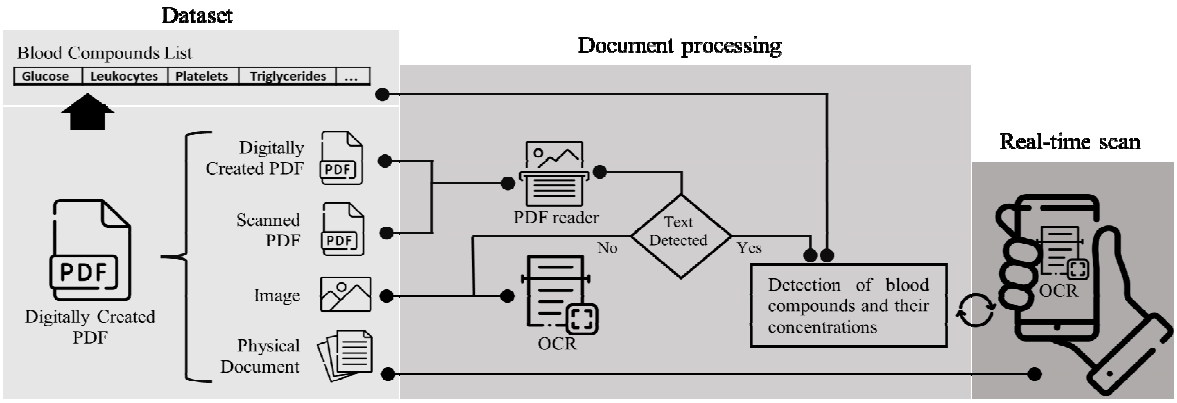

Figure 1 - Overview of the Smart Scan of Blood Test Documents.

However, the information contained in blood tests is not normally monitored in this way.

Currently, blood test results documents are provided via PDF and/or physical document. So, in the second part of this project, it is intended to develop a selective scanning system, capable of processing blood test documents in different formats so that later the information is aggregated on the same platform to streamline the process of monitoring the evolution of data. For this, an image processing method was developed to detect, in images of blood analysis results, the main information, namely the blood constituents and respective concentration. Thus, it is possible to create a database of people's medical information and later present these same aggregated data, for example in a graph to give a better idea of the physiological evolution.

The system was developed for mobile devices and its elaboration process was divided into three stages that are directly linked to the source of the document being processed. The first stage consisted of extracting text from DCPDFs. This type of document has searchable text and images, so it would only be necessary to use a PDF reader. The second stage would be dedicated to an image and SPDF documents. For these types, it was necessary to use an Optical Character Recognition (OCR) method to recognize the text contained in them. Finally, the last step was the real-time detection of analyzes contained in physical documents. The detection would be done through the camera of the mobile device.

The main contributions of this work are:

- Extract blood compounds and their concentration from DCPDFs, SPDFs and images;
- Vision system to extract blood compounds and their concentration in real time from physical documents;
- Validation of systems developed in mobile devices.

Henceforth, this paper has been organized as follows: In Section II, the framework is presented. A description of the procedures necessary for the implementation of the methods is made in Section III. In Section IV, the real-time scanning system was tested. The evaluation of all methods was explored in Section V. After all the experiences, Section VI performs an analysis of the results obtained. Finally, we conclude with some final remarks in Section VII.

II. Methods

A. General Overview

The development process of this work is based on three conceptual blocks, as described in Figure 1.

The first is the development of a dataset of DCPDFs of blood tests to be later used to obtain documents in different formats. Through these documents, a list of the main blood compounds presented will be created. Then documents in different formats will be forwarded to a text extraction block. A specific text extraction method is applied for each document format, namely: 1) DCPDFs will be processed by a PDF reader, 2) SPDFs and images will be processed by an OCR, and 3) physical documents will also be processed by an OCR in real-time through the camera of a mobile device. Third and last, the detected text is analyzed to verify if it contains any of the elements of the blood compounds dictionary. If this occurs, the element concentration value is also extracted.

B. Dataset Preparation

All documents acquired consisted of analyzes from the Unilabs company's medical laboratory. Although Unilabs has its own template for blood analysis documents, the layout of this type of documents is transversal to the various blood analysis clinics.

Since the acquired documents were DCPDFs, it was possible to generate new documents, through these, in the formats of images and physical documents. Later, through the physical documents, SPDFs were also generated. In the end, each document has four versions: DCPDFs, SPDFs, image, and physical document formats.

From the documents acquired, a list of all the blood compounds that were found was built. In total the list had 68 different elements and all in the Portuguese language (Figure 2).

ERITROCITOS, HEMOGLOBINA, HEMATOCRITO, VOL.GLOBULAR, HB.GLOBULAR, CONC.HB.G.MEDIA, RDW, LEUCOCITOS, NEUTROFILOS, EOSINOFILOS, BASOFILOS, LINFOCITOS, VOL.PLAQUETARIO, MONOCITOS, PLAQUETAS, VOL.PLAQUETARIO, CREATININA, TRIGLICERIDEOS, SODIO, POTASSIO, CLORETOS, GLICOSE EM JEJUM, COLESTEROL VLDL, COLESTEROL LDL, COLESTEROL TOTAL, COLESTEROL HDL, ACIDO URICO, TRANS.GLUT.PIRUVICA (TGP), GAMAGLUTAMILTRANSFERASE (GGT), SODIO, POTASSIO, CLORETOS, UREIA, T.TROMBOPLASTINA, TEMPO DE QUICK, TAXA DE PROTROMBINA, I.N.R, GLICOSE, GLICEMIA, TRIGLICERIDOS, MICROALBUMINURIA, CREATININEMIA, ALANINA, PSA TOTAL, PROTEINA C, CALCIO IONIZADO, VOLUME PLAQUETARIO, GLICADA (A1C), GLICADA IFCC, GME - ADA, OXALACETICA, PIRUVICA, TRANSFERASE, FOSFATASE ALCALINA, VITAMINA D, T4, T3, TSH, BILIRRUBINA TOTAL, BILIRRUBINA CONJUGADA, BILIRRUBINA NAO CONJUGADA, MAGNESIO, FOSFORO, ACIDO FOLICO, VITAMINA B12, CK TOTAL, FOSFATASE ALCALINA and DESIDROGENASE LACTICA.

Figure 2 - List of Blood Compounds

C. Text Extraction

As mentioned earlier, two methods of extracting text would be used depending on the format of the document in question. The documents would be divided into two groups, one would be DCPDFs and the other would be composed by images from SPDFs or real-time smart scan. The other would be DCPDFs.

1) DCPDFs

Regarding this format, they have searchable text and images, so it would only be necessary to use a PDF reader. The choice of the reading tool does not require comparing the efficiency of the different solutions on the market because they all meet the purpose. The choice fell on the Apache's PdfBox library[1] as it can be integrated into a mobile application (version 2.0.17.0).

2) SPDF, Images and Real Time

In the second group, extracting the text would not be as linear a process as in the case of DCPDFs. In this case, it is necessary to detect the text contained in the documents using an OCR method and it is more susceptible to extracting wrong text or even not detecting it. The state of the art resulted in the use of the OCR ML Kit[2] method, which is aimed at mobile devices.

3) Condition Tree

After extracting the text contained in the documents, it is necessary to perform filtering to obtain only the blood compounds and their respective concentrations. Based on the list of blood compounds mentioned above, was made the identification of blood compounds contained in the text extracted from the documents.

Subsequently, to identify the concentration of blood compounds detected, a method based on the layout of blood analysis documents was applied. After some study, it was verified that in several analysis laboratories, the disposition of the data is the same. Blood compounds are always accompanied by their concentration on the right and aligned horizontally. They may contain several values aligned horizontally like reference values, but the concentration is always the leftmost value and closest to the identification of the blood compound. The OCR ML Kit provides not only the detected text but also the coordinates where it is located. In this way, it was possible to identify the concentrations of each blood compound.

[1] https://github.com/TomRoush/PdfBox-Android (2022)

[2] https://developers.google.com/ml-kit/vision/text-recognition (2021)

III. Implementation Details

Each one of the 35 blood analysis documents initially acquired would have 4 versions of themselves that only differentiated the format. In these 35 documents, were recorded 260 appearances of blood compounds. To determine the detection accuracy of each one of the methods, they were tested in the 35 documents and verified how many compounds were detected out of the 260.

The OCR ML kit together with the developed condition tree would be applied to SPDFs, images, and real-time detection. The PSPDF kit with the condition tree would be applied to the DCPDFs. In the end, the blood compounds and respective concentrations detected would be verified, as well as the veracity of the values read.

These two tools, OCR ML Kit and PSPDF, have their own implementation guides provided by the developers.

IV. Experiments

All methods of extracting information from blood analysis documents were developed directly in a mobile application system. In order to validate them, a demo that allowed loading PDF documents, images or starting the camera to acquire images in real-time was developed.

In the case of PDF and image documents, the processing is straightforward. After being loaded, the desired information is extracted and shown to the user. On the other hand, in the case of physical documents, this process is done through camera live. The image processing is done in real-time as well as the sampling of the extracted data. As soon as a blood compound is detected, is enabled a button to save the data by the user. If it stops detecting and/or does not detect any blood compounds, the save button is disabled.

The demo application developed, as well as the scanning process can be seen in Figure 3.

V. RESULTS

Table 1 shows the comparison of the performance of the strategies proposed to deal with the extraction of blood compounds and their respective concentrations in the different document formats. In the average of all the methods, 95,38% of the blood compounds were detected as well as 87,63% of the respective concentrations in the 4 document formats.

Table 1 - Detected Compounds and Hit Rate of its Concentrations.

	DCPDFs	SPDFs	Images	Physical
Blood Compounds Detected	260/260 (100%)	232/260 (89.23%)	243/260 (93.46%)	257/260 (98.85%)
Correct Concentrations	249/260 (95.76%)	195/232 (84.05%)	204/243 (83.95%)	223/257 (86.77%)
False Detections	5	4	5	5

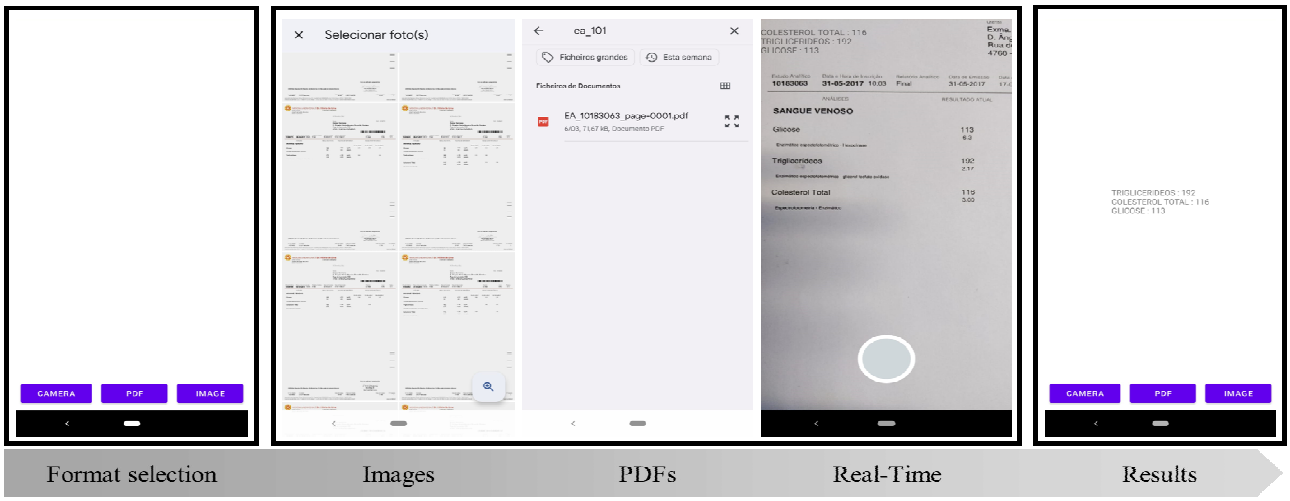

Figure 3 - Data Extraction Application.

VI. Discussion

It is possible to verify that the detection efficiency of blood compounds and their concentration is high in DCPDFs and physical documents. This is essentially due to two factors. In the case of text from DCPDFs, they do not contain errors because they were read and not detected by an OCR method. On the other hand, in the case of physical documents, it is justified by the fact that the digitization process is in real-time and presents the detected data. This allows users to verify the extracted data and only save the data when it is accurate.

In documents that use OCR (SPDFs, images, and physical documents) it was found that the extraction of the text by the ML kit has some flaws and consequently causes the association of compounds with wrong concentration values. A common example is when a "0" (zero) is read as "O". In this way, it is not possible to correctly associate a numerical value (concentration) with the respective blood compound. In the remaining group of documents (DCPDFs), the wrong concentrations that were obtained occurred in some cases where the letter size of the compounds was not the same as the respective concentration.

VII. Conclusions

In this work, a system for extracting the main information present in blood analysis documents was developed. The accuracy of the strategies used to extract the blood compounds and their respective concentrations was validated in a dataset of documents from a clinical analysis clinic. This work resulted in a public android library published in maven central repository. Indications for integration into future android projects can be found in the author's github repository[3].

In general, the strategies proved to be capable and with the potential to be applied in documents from other clinics and in other languages.

[3] https://github.com/pedrolobo98/SmH-BloodAnalysis-Library (2022)

Acknowledgment

This work was funded by the projects "NORTE-01-0145-FEDER-000045", supported by Northern Portugal Regional Operational Programme (Norte2020), under the Portugal 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). It was also funded by national funds, through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES in the scope of the project UIDB/05549/2020, UIDP/05549/2020 and LASI-LA/P/0104/2020.

References

[1] E. Sarzynski *et al.*, "Beta Testing a Novel Smartphone Application to Improve Medication Adherence," *https://home.liebertpub.com/tmj*, vol. 23, no. 4, pp. 339–348, Apr. 2017, doi: 10.1089/TMJ.2016.0100.

[2] M. Gupta and K. Soeny, "Algorithms for rapid digitalization of prescriptions," *Vis. Informatics*, vol. 5, no. 3, pp. 54–69, Sep. 2021, doi: 10.1016/J.VISINF.2021.07.002.

[3] S. Godbole, D. Joijode, K. Kadam, and S. Karoshi, "Detection of Medicine Information with Optical Character Recognition using Android," *Proc. B-HTC 2020 - 1st IEEE Bangalore Humanit. Technol. Conf.*, Oct. 2020, doi: 10.1109/B-HTC50970.2020.9298016.

[4] D. Kulathunga, C. Muthukumarana, U. Pasan, C. Hemachandra, M. Tissera, and H. De Silva, "PatientCare: Patient assistive tool with automatic hand-written prescription reader," *ICAC 2020 - 2nd Int. Conf. Adv. Comput. Proc.*, pp. 275–280, Dec. 2020, doi: 10.1109/ICAC51239.2020.9357136.

[5] R. I. Rumi, M. I. Pavel, E. Islam, M. B. Shakir, and M. A. Hossain, "IoT Enabled Prescription Reading Smart Medicine Dispenser Implementing Maximally Stable Extremal Regions and OCR," *Proc. 3rd Int. Conf. I-SMAC IoT Soc. Mobile, Anal. Cloud, I-SMAC 2019*, pp. 134–138, Dec. 2019, doi: 10.1109/I-SMAC47947.2019.9032709.

[6] S. Moon *et al.*, "Salience of Medical Concepts of Inside Clinical Texts and Outside Medical Records for Referred Cardiovascular Patients," *J. Healthc. Informatics Res.*, vol. 3, no. 2, pp. 200–219, Jun. 2019, doi: 10.1007/S41666-019-00044-5.

[7] N. Liu *et al.*, "A New Data Visualization and Digitization Method for Building Electronic Health Record," *Proc. - 2020 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2020*, pp. 2980–2982, Dec. 2020, doi: 10.1109/BIBM49941.2020.9313116.