BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Acute Lymphoblastic Leukemia blood cell classification using federated learning

– ITSG report –

Muntean Andrei, Data Science, andrei.muntean@stud.ubbcluj.ro
Chiorean Alexandra, Data Science, maria.alexandra.chiorean@stud.ubbcluj.ro

2023-2024

Abstract

Text of abstract. Short info about:

- project relevance/importance,
- inteligent methods used for solving,
- data involved in the numerical experiments;
- conclude by the the results obtained.
- Please add a graphical abstract of your work.

Remind that a good report should:

- be fun to read with many figures and visualizations;
- be easy to follow even for AI/ML novice;
- clearly convey the potential of AI/ML to the application domain;
- around 10 minutes to read (although this is not a hard constraint).

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 What? Why? How?

In US, approximately at each 3 minutes, a person is diagnosed with blood cancer. This disease is caused by the dysfunction of the bone marrow and can affect different blood types, resulting in different types of cancer like: Leukemia, lymphoma, myeloma, myelodysplastic syndromes (MDS), and myeloproliferative neoplasms (MPNs). In our project, we aim to reduce the diagnosis time by providing a computer based analysis of a microscopic image using an AI model. Moreover, to enable the model accuracy, we implement a federated learning approach, such that data does not leave the medical units.

- What is the (scientific) problem? Whenever it's the case for a cancer diagnosis, multiple doctors are called to give their professional opinion. The blood cell cancer is even more complex, most of the time needing different tests, which are also influenced by the stage of the cancer.

- Why is it important? Using peripheral blood smear (PBS) images, earlier detection of the ALL can be done, therefore more patients could be saved. Besides that, having a model that can predict the stage of the ALL, could help the doctors with a pragmatic opinion, obtaining a more precise and correct diagnosis for patients.

The examination of these PBS images by laboratory users is riddled with problems such as diagnostic error because the non-specific nature of ALL signs and symptoms often leads to misdiagnosis.

- What is your basic approach? The solution we proposed is based on a Convolutional Neural Network, which is known to obtain very accurate predictions on image tasks, having different architectures for classification tasks. In the medical field, patients confidence is really important, therefore datasets generation can be a very slow and bureaucratic. To avoid this impediment, we'll implement a federated learning approach that would speedup the entire process.

A short discussion of how it fits into related work in the area is also desirable. Summarize the basic results and conclusions that you will present.

1.2 Paper structure and original contribution(s)

The research presented in this paper advances the theory, design, and implementation of several particular models.

The main contribution of this report is to present an intelligent algorithm for solving the problem of blood cell cancer, concretely Acute Lymphoblastic Leukemia, which is very important since more and more patients are diagnosed with it. The algorithm is able to classify the blood samples in 4 classes: Benign, Pre-B, Pro-B, early Pre-B, all three being malign.

The second contribution of this report consists of building an intuitive, easy-to-use and user friendly software application. Our aim is to build a product that would enable doctors or patients to upload an image with the PBS on our website, where the model would predict the class of the blood sample. This application can be seen as a tool for doctors when they need a second, objective opinion, but also for the patients who could find a possible interpretation of their medical test before going to a specialist.

The third contribution in this report is related to the learning procedure. Since most of the medical data is very sensitive and confidential, we aim to train our model in a

Federated Learning approach, making the development and accuracy of the algorithm increase sooner. The data that the medical units is already having we'll be fed to the model, encoded, such that the data cannot be retrieved from the model. In this way, we take advantage of the quality and quantity of data, that each unit has, but it reaches to a greater purpose where patients from different location can be helped, on click away.

The present work contains xyz bibliographical references and is structured in five chapters as follows.

The first chapter is a short introduction in the blood cell cancer classification and its purpose in the current medical context.

The second chapter describes the problem in more details, having also more information about the design decision taken for implementing the software application.

The chapter 3 details briefly presents the current state of the art algorithms in both medical image classification and federated learning approaches.

The chapter 4 details the solution we are going to implement, starting from a smaller model, on which more complex layers we'll be added later. Moreover, this section will detail the necessary changes and adaptations done in order to synchronise model's learning results.

In the chapter 5 an analytical analysis will be drawn. Result obtain during the training process and a comparison with SOTA, and also describing the methodology of obtaining them should be documented.

Chapter 6 will describe the strengths, the opportunities, threads and weaknesses of our algorithm, as well as of our software application.

Finally, a conclusion chapter will be summarizing our results, future plans, why and how our approach is helping the society around us.

Chapter 2

Scientific Problem

2.1 Problem definition

Precisely define the problem you are addressing (i.e. formally specify the inputs and outputs). Elaborate on why this is an interesting and important problem.

Blood cell cancers, also known as hematologic cancers primarily affect the blood, bone marrow, lymph nodes, and spleen. These cancers are classified based on the specific type of blood cell affected and whether they are acute (develop rapidly) or chronic (develop gradually). Our main focus is to identify if a patient suffers from a benign or malign Acute Lymphoblastic Leukemia (ALL) which affects immature lymphocytes. To do this we are using a CNN which can analyze vast amounts of medical images, to detect subtle patterns indicative of early-stage ALL. Early diagnosis often leads to more effective treatments and improved outcomes.

CNNs excel at recognizing intricate patterns within images, even subtle ones that might be challenging for the human eye to discern. In PBS images, CNNs can identify specific cell types, anomalies, and morphological features. They can efficiently process large datasets, including thousands of PBS images, in a relatively short amount of time. This ability is essential for quick and accurate diagnosis, especially in busy clinical settings. CNNs can undergo continuous learning with new data, allowing them to adapt and improve their performance over time. As more annotated PBS images become available,

CNNs can refine their algorithms, leading to enhanced accuracy in diagnosis.

However, while CNNs can outperform doctors in specific tasks related to PBS image analysis, it's important to note that these technologies are most effective when used in conjunction with medical professionals. Doctors can provide critical context, interpret complex clinical situations, and make informed decisions based on the combined analysis of imaging data, patient history, and other relevant information. The synergy between intelligent algorithms and medical expertise can lead to the most accurate and comprehensive diagnoses.

Training such a model can become a very challenging task, due to the data sensitivity. Even though hospitals are possessing it, the collecting process can be very complicate. A solution for this problem is the federated learning. This method addresses privacy concerns, allowing institutions to collaborate without sharing sensitive patient data. Local devices train models on their data, sending only model updates to a central server, reducing communication overhead and ensuring data security.

As mention before, our CNN model has an PBS image as input and it returns 4 probabilities, each for the classes available: benign, Pre-B, early Pre-B and Pro-B. The class with the heights probability, will be the result for the processed sample. The image is uploaded in our web-base application and the result is shown there, after the inference on the model is done.

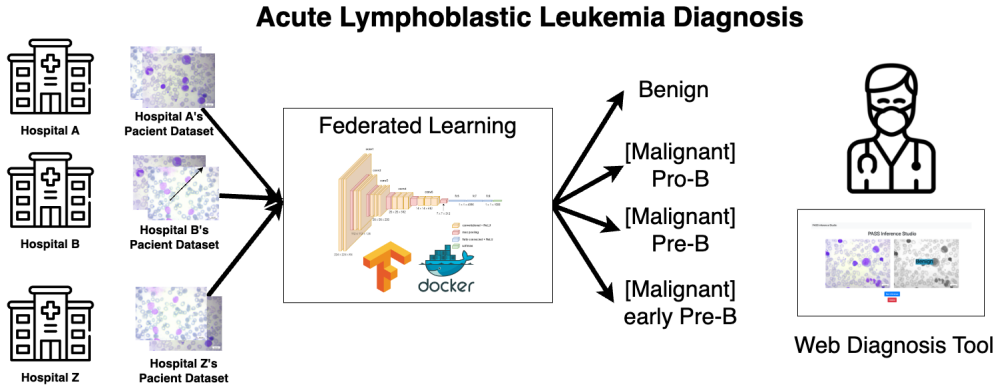
Chapter 3

State of the art/Related work

The application development is enabled by TensorFlow, a machine learning framework which provides the backbone for building intricate neural networks and implementing the CNN model, framework that is build upon Python, a versatile and widely-used programming language.

On the application side we use Django, a high-level Python web framework, which excels in the creation of robust and scalable web applications, providing a structured, efficient approach to backend development.

Last but not least, we use Docker, for the federated learning process, which ensures consistency and portability, encapsulating the entire application and its dependencies within containers.



Chapter 4

Investigated approach

Describe your approach!

Describe in reasonable detail the algorithm you are using to address this problem. A psuedocode description of the algorithm you are using is frequently useful. Trace through a concrete example, showing how your algorithm processes this example. The example should be complex enough to illustrate all of the important aspects of the problem but simple enough to be easily understood. If possible, an intuitively meaningful example is better than one with meaningless symbols.

Chapter 5

Application (numerical validation)

Explain the experimental methodology and the numerical results obtained with your approach and the state of art approache(s).

Try to perform a comparison of several approaches.

Statistical validation of the results.

5.1 Methodology

- What are criteria you are using to evaluate your method?
- What specific hypotheses does your experiment test? Describe the experimental methodology that you used.
- What are the dependent and independent variables?
- What is the training/test data that was used, and why is it realistic or interesting? Exactly what performance data did you collect and how are you presenting and analyzing it? Comparisons to competing methods that address the same problem are particularly useful.

5.2 Data

Describe the used data.

5.3 Results

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data. Are they statistically significant?

5.4 Discussion

- Is your hypothesis supported?
- What conclusions do the results support about the strengths and weaknesses of your method compared to other methods?
- How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

Chapter 6

SWOT Analysis

Chapter 7

Conclusion and future work

Try to emphasise the strengths and the weaknesses of your approach. What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.

Briefly summarize the important results and conclusions presented in the paper.

- What are the most important points illustrated by your work?
- How will your results improve future research and applications in the area?

Chapter 8

Latex examples

Item example:

- content of item1
- content of item2
- content of item3

Figure example

... (see Figure 8.1)

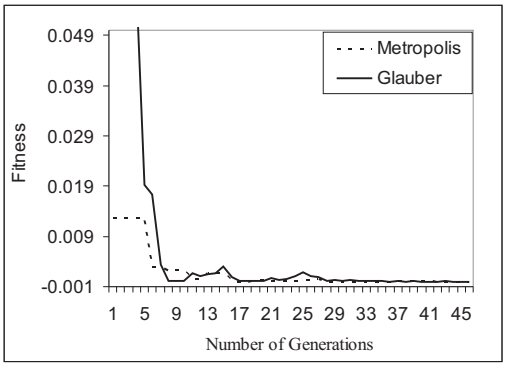


Figure 8.1: The evolution of the swarm size during the GA generations. This results were obtained for the f_2 test function with 5 dimensions.

Table example: (see Table 8.1)

Table 8.1: The parameters of the PSO algorithm (the micro level algorithm) used to compute the fitness of a GA chromosome.

Parameter	**Value**
Number of generations	50
Number of function evaluations/generation	10
Number of dimensions of the function to be optimized	5
Learning factor c_1	2
Learning factor c_2	1.8
Inertia weight	$0.5 + \frac{rand()}{2}$

Algorithm example

... (see Algorithm 1).

Algorithm 1 SGA - Spin based Genetic AQlgorithm

```
BEGIN
@ Randomly create the initial GA population.
@ Compute the fitness of each individual.
for i=1 TO NoOfGenerations do
  for j=1 TO PopulationSize do
    p ← RandomlySelectParticleFromGrid();
    n ← RandomlySelectParticleFromNeighbors(p);
    @ Crossover(p, n, off);
    @ Compute energy ΔH
    if ΔH satisfy the Ising condition then
      @ Replace(p,off);
    end if
  end for
end for
END
```

Bibliography