XXXII INTERNATIONAL SCIENTIFIC STUDENT CONFERENCE
 SAKHAROV READINGS

**WEB APPLICATION FOR LEUKEMIA TYPE CLASSIFICATION USING CNN MACHINE LEARNING MODEL**

Pornsiwakul P., Juirnarongrit N., Kaiyawong K.

*Mahidol Wittayanusorn School, Nakhon Pathom, Thailand*

pasin.por\_g30@mwit.ac.th

Nowadays, the number of leukemia patients in Thailand is increasing rapidly. According to the Age-standardized Rate (ASR) of Thailand, the ratio of male is 3.9 : 100,000 and female is 3.2 : 100,000. In addition, the leukemia data of Thailand from WHO Cancer Country Profile 2020 is in the 9th place, compared with other cancers.

Leukemia is commonly classified into 4 types: Chronic lymphocytic leukemia (CLL), Chronic myeloid leukemia (CML), Acute lymphocytic leukemia (ALL), and Acute myeloid leukemia (AML). CLL affects lymphoid cells and usually grows slowly. It accounts for more than 15,000 new cases of leukemia each year. Most people who are diagnosed with the disease are over the age of 55. It almost never impacts children. CML impacts myeloid cells and typically grows slowly at first. It accounts for nearly 5,000 new cases of leukemia each year. ALL affects lymphoid cells and grows quickly. It accounts for more than 5,000 new cases of leukemia each year. ALL is the most common type of leukemia in young children. It also affects adults. AML affects myeloid cells and grows quickly. It accounts for more than 13,000 new cases of leukemia each year. It occurs in both adults and children. There are many ways to diagnose leukemia, but the most useful way is a complete blood count.

Complete blood count (CBC) is a blood test used to evaluate the overall health and detect a wide range of disorders, including anemia, infection, and leukemia. CBC can be done by taking a blood sample and counting its components which are red blood cells, white blood cells, hemoglobin, and platelets. Normal people have around 3,300 - 8,700 white blood cells per cubic millimeter. Neutrophil, Lymphocyte, Monocyte, and Eosinophil are the most common white blood cells that can be found in the blood sample respectively.

At present, both artificial intelligence (AI) and machine learning are being widely used in the medical field due to their simplicity and efficiency. Common applications include: patient diagnosis, end-to-end drug discovery and development, improvement on physician-patient communication, medical documents transcription, and remote patient treatment [1]. We are interested in patient diagnosis because the diagnosis processes for some diseases are time-consuming and exhausting for doctors; thus, AI and machine learning can help automate diagnosis and save time for doctors.

Our project, Web Application for Leukemia Type Classification Using CNN Machine Learning Model will be divided into three main parts. The data management, the machine learning, and the making of the web application.

The dataset we received is from Siriraj Piyamaharajkarun Hospital, including 3,392 images of blood smear (.png format), and 3,392 labels of the white blood cells inside each image (.txt format). Each class is one of twelve general types of leukemia cells: Atypical Lymphocyte, Band Neutrophil, Basophil, Blast, Eosinophil, Lymphocyte, Metamyelocyte, Monocyte, Myelocyte, NRC (Neo red cell), Promyelocyte, and Segmented Neutrophil. The class representation is quite imbalanced with a range of 524 as shown in Figure 1 (a). For this reason, we use a technique called oversampling using SMOTE (Synthetic Minority Oversampling Technique). The resulted dataset had a range of 66 as shown in Figure 1 (b). After the oversampling, we split the data into training data and validation data. The training data include 70% of the entire dataset and the rest will be for validation.

For machine learning, we used YOLOv5 to train our dataset. YOLOv5 is an object detection algorithm that divides images into a grid system. Each cell in the grid is responsible for detecting objects within itself. [2]. The finished model will return to us a weight.pt file (pytorch extension) which we can use with the detect.py file from the YOLOv5 to detect the object. To visualize the data of the runs, we used the Weights and Biases web application to help with this part. The performance measures used to determine the model efficiency include precision, recall, and F1 score. For our model, the best precision is 80.4%; the best recall is 79.2%; and the best F1 score is 0.797, as shown in Figure 2.

For the website, we created a form in HTML for receiving the images from the users. The images will be sent to the model by the Django library in Python. After the detection is done, the web application will redirect the user to the result page, which has the user-uploaded image with bounding boxes for each leukemia cell, and a detailed description of each cell, including the confidence level.

In conclusion, the web application we created allows the user to upload a single blood smear image into the database. The returned image is labeled with the bounding box of the detected white blood cell(s) and the classification types, which are the 12 general types of leukemia cells: Atypical Lymphocyte, Band Neutrophil, Basophil, Blast, Eosinophil, Lymphocyte, Metamyelocyte, Monocyte, Myelocyte, NRC (Neo red cell), Promyelocyte, and Segmented Neutrophil. The model used in the website in training using YOLOv5s model trained and validated with our dataset of 3,392 images retrieved from Siriraj Piyamaharajkarun Hospital. The final model has precision of 80.4%, recall of 79.2%, mean average precision over 0.5 threshold of 83.2% and F1 score of 0.797.



**Figure 1 (a) Class representation before oversampling**



**(b) Class representation after oversampling**



**Figure 2 Result of our trained model**

**References**

[1] K. Basu, R. Sinha, A. Ong, T. Basu. *Indian Journal of Dermatology*, 2020, **65**, 365.

[2] X. Zhu, S. Lyu, X. Wang, Q. Zhao. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2778-2788.

**ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ КЛАССИФИКАЦИИ ТИПОВ ЛЕЙКЕМИИ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ**

**НА ОСНОВЕ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ**

**Порнсивакул П., Жуирнаронгрит Н., Кайявонг К.**

**под руководством Саенга Т., Иттичаивонг П.**

**Школа Махидол Виттаянусорн**

На сегодняшний день количество больных лейкемией в Тайланде стремительно увеличивается. Согласно тайландским показателям с разбивкой по возрастам (ASR), болеют 3.9 из 100,000 мужчин и 3.2 из 100,000 женщин. Кроме того, заболевание лейкемией находится на 9-м месте в Cancer Country Profile ВОЗ для Таиланда по сравнению с другими раковыми заболеваниями.

Лейкемию обычно делят на 4 типа: хронический лимфолейкоз (CLL), хронический миелолейкоз (CML), острый лимфолейкоз (ALL) и острый миелолейкоз (AML). CLL поражает лимфоидные клетки и обычно распространяется медленно. Ежегодно на его долю приходится более 15,000 новых случаев лейкемии. В основном заболевание диагностируют у людей старше 55 лет, оно почти не поражает детей. CML, поражающее миелоидные клетки, начинает расти медленно. В год обнаруживается около 5,000 новых случаев CML. ALL, который поражает лимфоидные клетки и распространяется быстро, обнаруживают примерно у 5,000 пациентов в год. Это самый распространённый тип лейкемии у детей раннего возраста, но он поражает и взрослых. AML также распространяется быстро, но воздействует на миелоидные клетки. В год выявляется около 13,000 новых случаев. AML встречается как у взрослых, так и у детей. Существует много способов выявления лейкемии, но самый полезный — общий анализ крови.

Общий клинический анализ крови (ОАК) — это анализ крови, предназначенный для оценки общего состояния здоровья и выявления широкого спектра заболеваний, включая анемию, инфекции и лейкемию. Для ОАК необходимо взять образец крови и подсчитать количество эритроцитов, лейкоцитов и тромбоцитов. В норме у человека около 3,300 - 8,700 лейкоцитов на кубический миллиметр. Нейтрофилы, лимфоциты, моноциты и эозинофилы — виды лейкоцитов, которые встречаются в образцах крови наиболее часто.

На настоящий момент и искусственный интеллект (ИИ), и машинное обучение, благодаря своей простоте и эффективности, широко используются в медицине. Их основные сферы применения включают в себя диагностику заболеваний пациентов, сквозное (end-to-end) изобретение и разработку лекарств, улучшение коммуникации между врачом и пациентом, копирование медицинских документов и удалённое лечение пациентов [1]. Наша цель - помочь в постановке диагнозов, потому что этот процесс в ряде случае отнимает много времени и усилий врачей. Таким образом, ИИ и машинное обучение могут помочь автоматизировать этот процесс и высвободить время.

Наш проект “**Веб приложение для классификации типов лейкемии, использующее модель машинного обучения на основе сверточных сетей (CNN)”** разделён на три основные части: управление данными, машинное обучение, создание веб-приложения.

Полученный нами dataset из базы данных больницы Siriraj Piyamaharajkarun включает в себя 3,392 изображения мазка крови (в формате .png) и 3,392 маркировочные подписи лейкоцитов на каждом изображении (в формате .txt). Каждый класс относится к одному из 12 основных классов клеток лейкемии: атипичные лимфоциты, палочкоядерные гранулоциты, базофилы, бласты, эозинофилы, лимфоциты, метамиелоциты, моноциты, миелоциты, NRC (Neo red cell), промиелоциты и сегментоядерные гранулоциты. Классы представлены довольно неравномерно (см. диаграмму 1(a)), размах (range) составляет 524. Для устранения этой проблемы мы воспользовались передискретизацией SMOTE (метод передискретизации синтетического меньшинства). Полученный dataset более равномерен, размах уменьшился до 66 (см. диаграмму 1 (b)). После передискретизации мы разбили данные на тестовую (проверочную) и тренировочную части. Тренировочная часть включает 70% всей выборки, а оставшаяся будет использована для оценки качества модели.

В разделе машинного обучения мы использовали YOLOv5 для тренировки нашей модели. YOLOv5 — алгоритм распознавания объектов, использующий разделение изображения на ячейки прямоугольной сетки. Каждая ячейка ответственна за определение объектов, заключённых внутри неё [2]. В результате исполнения итоговая модель вернёт файл weight.pt (расширение pytorch), используя который программа detect.py, полученная с помощью YOLOv5, распознает объект. Для визуализации данных запусков мы использовали веб приложение Weights and Biases. Измерения производительности, используемые для определения эффективности модели, включали в себя точность (precision), полноту (recall) и F1-метрику. Лучшими результатами для нашей модели оказались точность - 80,4%; полнота - 79,2%; F1 - 0,797 (см. диаграмму 2).

Для сайта мы создали HTML-форму, куда пользователи могут загружать изображения, которые затем будут переданы модели с помощью библиотеки Django в Python. После обработки данных веб-приложение перенаправит пользователя на страницу результатов, где на загруженном пользователем изображении клетки лейкемии будут выделены и дополнены детальным описанием, включающим уровень достоверности.

Таким образом, созданное нами приложение позволяет пользователю загрузить в базу данных изображение мазка крови. На выходе пользователь получит изображение, где ограничивающим прямоугольником будут выделены лейкоцит(ы), а также подписаны результаты классификации, являющиеся одним из основных 12 типов клеток лейкемии: атипичные лимфоциты, палочкоядерные гранулоциты, базофилы, бласты, эозинофилы, лимфоциты, метамиелоциты, моноциты, миелоциты, NRC (Neo red cell), промиелоциты и сегментоядерные гранулоциты. Тренировка модели на сайте прошла с помощью алгоритма YOLOv5s, а качество нашей модели подтверждено с помощью dataset, который получен из больницы Siriraj Piyamaharajkarun. Окончательная модель имеет точность 80,4%, полноту 79,2%, среднюю точность выше 0,5 порога 83,2%, F1-метрику 0,797.



**Рис 1 (a) Классовые представления до передискретизации** 

**(b) Классовые представления после передискретизации**



**Рис 2 Результат обученной модели**

**Литература:**

[1] K. Basu, R. Sinha, A. Ong, T. Basu. *Indian Journal of Dermatology*, 2020, **65**, 365.

[2] X. Zhu, S. Lyu, X. Wang, Q. Zhao. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2778-2788.