DATA SAMPLING METHODS ON IMBALANCED DATASETS FOR PNEUMONIA DETECTION IN COVID-19 PATIENTS

SYASYA FARINA BINTI DZULKEFLI

UNIVERSITI TEKNOLOGI MALAYSIA

DATA SAMPLING METHODS ON IMBALANCED DATASETS FOR PNEUMONIA DETECTION IN COVID-19 PATIENTS

SYASYA FARINA BINTI DZULKEFLI

A project report submitted in partial fulfilment of
the requirements for the award of the degree of
Master of Engineering (Computer and Microelectronic Systems)

School of Electrical Engineering
Faculty of Engineering
Universiti Teknologi Malaysia

FEBRUARY 2022

DEDICATION

This thesis is dedicated to my husband and my daughter who have been my pillar of strength every time I am thinking of giving up. Also, to my family, friends and those who have been contributed directly or indirectly during the phase of completing this project.

ACKNOWLEDGEMENT

In the name of Allah S.W.T, the Most Beneficent and Merciful. All praise is due to Allah, Lord of the Worlds who grants me the strength and determination to complete this research and final year project thesis despite all the difficulties and challenges I have faced. First and foremost, I would like to raise my deepest gratitude to my research's supervisor, Dr Ab Al-Hadi Ab Rahman for his great knowledge, his continuous support for my final year project 1 & 2 and for his patience in guiding me throughout the journey of completing this research and writing the thesis.

My earnest appreciations to my friends especially Asyikin, Amira and Adib and the rest of my Intel colleagues for the strong support, advice, and assistance through various stages along the master journey.

Lastly, my sincere appreciation and thank you goes to my husband, daughter, my parents, and my family who always become my pillar of strength, give endless encouragement and support to complete this final year project.

ABSTRACT

Data classification is one of the important aspects in the real-world decision-making support function which can be severely affected by an imbalance class distribution in the training data especially in the medical field. In medical datasets, the data are mainly had imbalanced datasets problem which composed of minority of normal samples and majority of abnormal samples. As for today's example, the outbreak of novel coronavirus disease or also called as COVID-19 in late 2019 is still on-going which we can see new variants have been discovered from time to time and this can lead to increasing of number of cases around the world. The medical staffs can detect the patients by checking on the symptoms but one of the common COVID-19 symptoms that will be investigating in this research is pneumonia. It is important to detect the pneumonia faster at early stage to avoid it become more severe. Thus, Chest Xray scan images can be considered as one of the confirmatory approaches as they are fast to obtain and easily accessible. Diagnosing diseases in general is a considerable application of data analysis for medical science. In this research, data sampling methods will be explored and implemented for pneumonia detection for imbalanced datasets. The imbalanced datasets of pneumonia X-Ray images from Kaggle dataset will be obtained and different existing data sampling methods also new proposed methods that are achieved by combining or modifying exiting methods will be implemented to balance the images between majority and minority classes of the datasets. After achieved a balanced dataset, CNN model will be implemented to set benchmark of detection accuracy in terms of confusion matrix, precision, accuracy, F1-score and recall for each method and those results will be compared to choose which method will give the highest accuracy in detecting pneumonia. The best undersampling method is near miss with 85.47% accuracy, the best oversampling method is data augmentation with 88.78% accuracy and the best combination method is SMOTE + Tomek with 83.20% accuracy compared to 79% of accuracy when there is no method being implemented on the imbalanced dataset. Implementing data sampling methods will boost the performance of data classification in all applications especially in detecting pneumonia in COVID-19 patients.

ABSTRAK

Klasifikasi data merupakan salah satu aspek yang sangat penting dalam membuat keputusan sebenar yang boleh terjejas teruk berpunca daripada ketidakseimbangan data terutamanya dalam bidang perubatan. Dalam set data perubatan, kebiasaannya data perubatan yang sedia ada mempunyai masalah set data tidak seimbang yang terdiri daripada minoriti sampel yang normal dan majoriti sampel yang tidak normal. Contoh hari ini yang dapat dilihat, wabak penyakit novel coronavirus atau juga dipanggil sebagai COVID-19 yang telah dikesan pada penghujung tahun 2019 dan kini masih merebak dengan pantas dimana varian baru telah ditemui dari semasa ke semasa dan ini boleh menyebabkan peningkatan jumlah kes di sekeliling dunia. Kakitangan perubatan boleh mengesan pesakit dengan memeriksa simptom yang dihadapi tetapi salah satu simptom COVID-19 yang biasa menjangkiti pesakit-pesakit COVID-19 dan yang akan disiasat dalam penyelidikan ini adalah radang paru-paru. Adalah penting untuk mengesan radang paru-paru lebih cepat pada peringkat awal untuk mengelakkan ia menjadi lebih teruk. Oleh itu, imej imbasan X-Ray dada boleh dianggap sebagai salah satu pendekatan pengesahan kerana ia cepat diperoleh dan mudah diakses. Mendiagnosis penyakit secara umum adalah aplikasi analisis data yang besar untuk sains perubatan. Dalam penyelidikan ini, kaedah pensampelan data akan diterokai dan dilaksanakan untuk mengesan radang paru-paru bagi set data yang tidak seimbang. Set data tidak seimbang yang merangkumi imej X-Ray pneumonia daripada laman web Kaggle akan diperoleh dan pelbagai kaedah pensampelan data sedia ada dan juga kaedah baru iaitu gabungan antara kaedah-kaedah sedia ada akan dilaksanakan untuk mengimbangi imej antara kelas majoriti dan minoriti. Selepas mencapai set data yang seimbang, model CNN akan dilaksanakan untuk menetapkan penanda aras ketepatan pengesanan dari segi matriks kekeliruan, ketepatan, skor F1 bagi setiap kaedah dan keputusan tersebut akan dibandingkan untuk memilih kaedah yang akan memberikan ketepatan tertinggi. dalam mengesan radang paru-paru. Kaedah pensampelan terkurang terbaik ialah nyaris hampir dengan ketepatan 85.47%, kaedah pensampelan berlebihan terbaik ialah penambahan data dengan ketepatan 88.78% dan kaedah gabungan terbaik ialah SMOTE + Tomek dengan ketepatan 83.20% berbanding ketepatan 79% apabila tiada kaedah dilaksanakan pada set data yang tidak seimbang. Melaksanakan kaedah pensampelan data akan meningkatkan prestasi klasifikasi data dalam semua aplikasi terutamanya dalam mengesan radang paru-paru dalam pesakit COVID-19.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	-	Covolutional Neural Network
CNN	-	Condensed Nearest Neighbor
SMOTE	-	Synthetic Minority Oversampling Technique
ADASYN	-	Adaptive Synthetic Sampling Method
OSS	-	One Sided Solution
NCR	-	Neighbor Cleaning Rule
VGG16	-	Visual Geometry Group 16
COVID-19	-	Coronavirus Disease 2019
GAN	-	Generative Adversarial Networks
ILSVR	-	ImageNet Large Scale Visual Recognition
RGB	-	Red Green Blue

LIST OF APPENDICES

CHAPTER 1

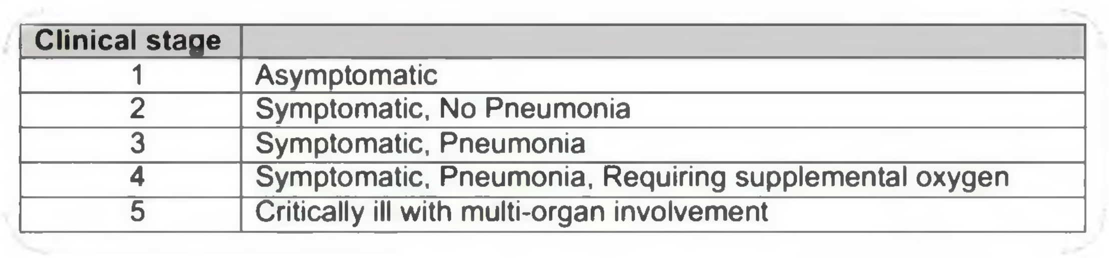
INTRODUCTION

1.1 Research Background

1.1.1 COVID-19

COVID-19 or also known as Coronavirus disease 2019 has been firstly discovered in late 2019 during the respiratory illness outbreak in Wuhan City, one of the cities in Hubei Province, China.[1] This coronavirus was initially reported to the World Health Organization (WHO) on December 31, 2019. However, one month later, WHO has declared it as an outbreak of global health emergency. [2,3] This COVID-19 virus is very contagious and has quickly spread all over the world.

COVID-19 most common symptoms are such respiratory symptoms which can much feel like common flu, cold and pneumonia. COVID-19 also may attack more than the lungs and respiratory system. Figure 1.1 below shows the 5 categories of COVID-19 patients with the symptoms that can be detected in the patient's body.



Clinical stage	
1	Asymptomatic
2	Symptomatic, No Pneumonia
3	Symptomatic, Pneumonia
4	Symptomatic, Pneumonia, Requiring supplemental oxygen
5	Critically ill with multi-organ involvement

Figure 1.1: Categories of COVID-19 patients with symptoms

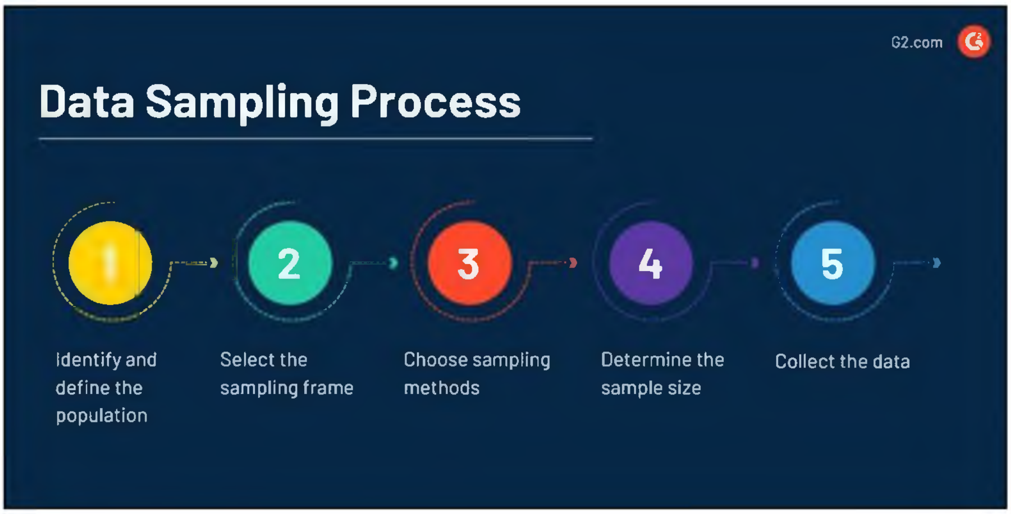
Viruses can be constantly changed through mutations. When a virus has one or more mutations, it is called a variant of the original virus. The Centres for Diseases Control and Prevention currently has identified two new variants which have caused chaos and rapidly increasing in daily cases all round the world which are Delta (B.1.617.2) variant and Omicron (B.1.1.529) variant. Even with the new variants have been discovered, the most common symptom that can be detected is respiratory illness or pneumonia. With the rapid increasing in daily COVID-19 cases, it has become chaos especially in healthcare state as there will be tonnes of X-Ray images of pneumonia or non-pneumonia cases that need to be analysed by the specialists and this issue can cause imbalanced of the data.

1.1.2 Data Sampling in Machine Learning

Generally, skewed data distributions or imbalanced datasets is the major challenge in medical applications especially in image processing. In any medical applications of data classification, the specialist often facing the imbalanced number of data samples where at least one of the classes consist of very minimal amount of data. At the same time, this problem also represents the same problem facing in the machine learning algorithms. An imbalanced data distribution can cause bias in healthcare data analysis which than lead to a false result. Thus, it is very crucial to consider the representation of both majority and minority classes of data and make sure the data is well balanced before proceeding with the data analysis.

Therefore, to overcome the problem of skewed data distributions or imbalanced datasets, any data sampling method can be implemented. Data sampling is a collection of different kind of statistic methods or techniques which analyze the patterns and trends in a dataset and transform it into a balanced set. Most machine learning algorithms are designed to operate on a balanced of data distribution. Therefore, if that is not the case, the algorithm will ignore the differences betweenthe imbalanced classes just to achieve a good performance. This is not a good approach as it will lead to poor predictive performance of data analysis.

The overall process of a data sampling is a statistic analysis method which helps to draw a conclusion of a sample's population. The first step of data sampling is to define and identify the population of samples or also referred to data collection. This step can be done by carrying out various observations, surveys, interviews, or questionnaires. It is also very important to set a parameter as a benchmark. Next, select the sampling frame of the data which is the list of items forming a population where the sample is taken. An example of sampling frame is the X-Ray images containing pneumonia cases of COVID-19 patients. Then, choose a sampling method need to be applied on the sample's population where this method will be broken down into two categories which are profitability and non-profitability sampling. The next step is to determine the sample size that need to be analyzed. In data sampling, the size of a sample is the same number of samples that need to be measured for an observation to be made. Finally, it is the time to collect data from the sample after all the process. Based on the data presents, it will then either make a conclusion, actionable plan, or decision. Figure 1.2 shows the overall data sampling process.



G2.com

Data Sampling Process

1 Identify and define the population

2 Select the sampling frame

3 Choose sampling methods

4 Determine the sample size

5 Collect the data

Figure 1.2: Data Sampling Process

1.1.2.1 Data Sampling Method

There are various sampling methods can be chosen within the two main categories: probability and non-probability sampling.

1.1.2.1.2 Probability Sampling

Every aspect of the sample population will have an equal chance of being selected as one of the data to be analysed and studied. This category is typically providing the best chance of creating a representative of sample as good as possible. The methods lying under this category are:

i. Simple random sampling – Each sample is chosen by random or by chance and each sample has an equal chance of being selected. This method is the most straight forward to perform probability sampling

ii. Systematic sampling – The first individual of sample is selected randomly while other sample will be selected using a fixed sampling interval. This method is more complicated than simple random sampling, but it is easy to execute and understand as executer has full control of the process which there will be low risk factor to data contamination.

iii. Stratified sampling – A method where samples are divided into several small subgroups called strata based on their common factor and similarities, Then, the samples will be randomly collected and selected from each of the subgroup. This method provides greater precision than other methods as it can be chosen to test a smaller sample instead of the whole population of samples.

iv. Cluster sampling – This method divides the whole population into small sections or cluster based on a defining factor where the clusters then will be randomly selected to be put in the sample and analysed. This method allows a larger sample of data to be studied and it is also quick and a less expensive method.

v. Multistage sampling – A complicated form of cluster sampling. It works by dividing a large population of samples into many small clusters and it will be broken down further based on secondary factor before being sampled and analysed. The staging of this method continues as those small clusters will be continued to be identified and analysed.

1.1.2.1.2 Non-Probability Sampling

For this type of sampling method, every aspect of the sample population will not have an equal chance of being selected as one of the data to be analysed and studied or in other words it won't rely on randomization. This category is relying on the ability of the specialists to choose the elements for a sample. This could result in having a very general conclusion. The methods lying under this category are:

i. Convenience sampling – This method sometimes is being called as availability or an accidental sampling as the data is collected from an easily available or accessible group. The data is selected based on the availability to be part of the sample. It is prone to be a bias method because the sample may not be representing the actual specific characteristics needed but it is known to be easy to carry out at a low cost in a timely manner.

ii. Quota sampling – By applying this method, the data is chosen based on predetermined standards. The specialists handle this method must ensure an equal representation within the sample from all subgroups of the dataset. This method is an easy method once characters are determined, and it is also a cost-effective method.

iii. Judgment sampling – A method also known as selective sampling where it is based on the assessment of experts when choosing the data to be included in the sample. This method takes lesser time compared to other methods.

iv. Snowball sampling – This method also called as a referral or chain referral sampling where it is being used where a population is unknown or very rare. A small group of data is being selected based on a specific criterion. It is called snowball because the data is chosen specifically to be in the sample the same increases in size like a snowball.

1.1.2.2 Advantages and Disadvantages of Data Sampling

There are reasons why data sampling is very popular as there are many advantages of this method. The first advantage is low cost of sampling as the data of a sample is collected from a small proportion of entire population. Data sampling method also consume lesser time compared to census or survey technique. Lastly, as a sample is drawn, it permits high degree of accuracy as due to a small proportion of population and more careful execution of data analysis can be done. Thus, the results of the data sampling turn out to be sufficiently accurate.

However, there are also some disadvantages and challenges of the data sampling method that might come up during the sampling process. False employment of data sampling method might lead to a chance of bias in sampling results which then can lead to erroneous conclusions. Besides that, a representative of a sample is difficult to be selected if the phenomena under the study are of a very complex nature. Lastly, to apply a sampling method requires an adequate specific knowledge in sampling techniques. This is because sampling is involving a statistic analysis and probable error calculation. Lack of knowledge in sampling might commit to serious mistakes and lead to misleading results. Thus, it will be going to take both time and effort no matter which method is being chosen.

1.2 Problem Statements

Transmission of COVID-19 virus can occur through direct, indirect, or even close contact with the infected people through any infected secretions like saliva and respiratory droplets which are expelled when people are coughing, sneezing, singing, or talking. The virus will latch its spiky surface protein onto the receptors of healthy cells especially those cells in the lungs. Therefore, there is high possibility that the lungs are the most affected organ when being infected by COVID-19. This can lead to respiratory illness or pneumonia. One of the methods to detect pneumonia patients is through analyzing of X-ray images. However, rapid increasing in daily COVID-19 cases all over the world has made the healthcare service become chaos as it will also increase the number of samples for X-ray images for COVID-19 patients and thus results in imbalanced of data where there will be a huge gap between pneumonia and non-pneumonia cases. At the same time, it will lead to delay of data analysis and treatment as there are shorts of medical staff to handle the huge amount of X-Ray images to be analyzed. This situation also will give hard time to the healthcare in developing countries where some of the countries are lacking specialists as only the specialists can analyze the X-Ray images.

Other than that, mostly hospitals or clinics in rural area or small town don't have a computer-aided diagnosis system to diagnose and analyze X-Ray images. As X-Ray images have some limitations, such as the images might have lower resolutions which can be difficult to be interpreted and diagnosed by rough eyes. Thus, the specialists might tend to do some mistakes during the analysis and lead to incorrect results. Lastly, as different people will show different stages of pneumonia, the appearance of pneumonia on the images might be uncertain which can lead to subjective decisions. Thus, all these problems could be solved by developing a computer-aided system which can automatically detect the pneumonia in COVID-19 patients with the help of data sampling methods.

1.3 Research Objectives

The objectives of the research are:

i. To improve the performance of learning models and give accurate results.

ii. To compare which existing methods can give high accuracy in detecting pneumonia.

iii. To explore new method to detect pneumonia by combining different data sampling methods to give better performance.

1.4 Research Scope

The scopes of this research are:

- The algorithms used are undersampling algorithm and oversampling algorithm.
- The variation of algorithms is undersampling algorithm + oversampling algorithm.
- The input data of this research is a sample of pneumonia and non-pneumonia X-ray images from Kaggle datasets.
- The performance indicators of this research are confusion matrix, accuracy, recall, precision, f1-score.
- The codes are written in Python language and ran with Kaggle Notebook and with Intel® Core™ i7-8665U CPU @1.90GHz 8GB RAM

1.5 Thesis Organization

This whole thesis is divided into 5 main chapters as below:

- Chapter 1 provides an introduction of the research including the COVID-19 virus, data sampling methods in machine learning with the examples. The problems, objectives and scopes of the research are also being highlighted in this chapter.

- Chapter 2 elaborates in detail on literature review of the research that includes the implementation of data sampling methods in previous works.

- Chapter 3 describes the data sampling algorithms, input dataset, and design methodology implemented in this research.

- Chapter 4 explains the results obtained and discussion for each data sampling method in aspect of performance indicators.

- Chapter 5 concludes overall findings of the research with the research objective listed and recommendations for future work.

REFERENCES

[1] CDC. 2019 Novel Coronavirus, Wuhan, China. CDC. Available at https://www.cdc.gov/coronavirus/2019-ncov/about/index.html. January 26, 2020.

[2] Gallegos A. WHO Declares Public Health Emergency for Novel Coronavirus. Medscape Medical News. Available at https://www.medscape.com/viewarticle/924596. January 30, 2020.

[3] Ramzy A, McNeil DG. W.H.O. Declares Global Emergency as Wuhan Coronavirus Spreads. The New York Times. Available at https://nyti.ms/2RER70M. January 30, 2020.

[4] Devi, D., Biswas, S. K., & Purkayastha, B. (2020). A review on solution to class imbalance problem: Undersampling approaches. *2020 InternationalConference on Computational Performance Evaluation (ComPE)*.

[5] Krishnan, U. and Sangar, P., 2021. A Rebalancing Framework for Classification of Imbalanced Medical Appointment No-show Data. Journal of Data and Information Science, 6(1), pp.178-192.

[6] Luján-García, J., Yáñez-Márquez, C., Villuendas-Rey, Y. and Camacho-Nieto, O., 2020. A Transfer Learning Method for Pneumonia Classification and Visualization. Applied Sciences, 10(8), p.2908.

[7] Vuttipittayamongkol, Pattaramon & Elyan, Eyad. (2020). Overlap-Based Undersampling Method for Classification of Imbalanced Medical Datasets. 10.1007/978-3-030-49186-4_30.

[8] Kaur, Prabhjot & Gosain, Anjana. (2018). Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise. 10.1007/978-981-10-6602-3_3.

[9] Masud, Mehedi & Bairagi, Anupam & Nahid, Abdullah & Sikder, Niloy & Rubaiee, Saeed & Ahmed, Anas. (2021). A Pneumonia Diagnosis Scheme Based on Hybrid Features Extracted from Chest Radiographs Using an Ensemble Learning Algorithm. Journal of Healthcare Engineering. 2021. 1-11. 10.1155/2021/8862089.

[10] Yang, Zaifeng & Hou, Yubo & Chen, Zhenghua & Zhang, Le & Chen, Jie. (2021). A Multi-Stage Progressive Learning Strategy for Covid-19 Diagnosis Using Chest Computed Tomography with Imbalanced Data. 8578-8582. 10.1109/ICASSP39728.2021.9414745.

[11] Fajardo, Val & Findlay, David & Jaiswal, Charu & Yin, Xinshang & Houmanfar, Roshanak & Xie, Honglei & Liang, Jiaxi & She, Xichen & Emerson, D.B.. (2020). On oversampling imbalanced data with deep conditional generative models. Expert Systems with Applications. 169. 114463. 10.1016/j.eswa.2020.114463.

[12] García-Ordás, María & Benítez-Andrades, José & García, Isaías & Benavides, Carmen & Alaiz Moreton, Hector. (2020). Detecting Respiratory Pathologies Using Convolutional Neural Networks and Variational Autoencoders for Unbalancing Data. Sensors. 20. 10.3390/s20041214.

[13] Nishio, Mizuho & Noguchi, Shunjiro & Matsuo, Hidetoshi & Murakami, Takamichi. (2020). Automatic classification between COVID-19 pneumonia, non-COVID-19 pneumonia, and the healthy on chest X-ray image: combination of data augmentation methods. Scientific Reports. 10. 17532. 10.1038/s41598-020-74539-2.

[14] Habib, Nahida & Hasan, Md & Reza, Md & Mohammad, · & Rahman, Mohammad Motiur. (2020). Ensemble of CheXNet and VGG-19 Feature Extractor with Random Forest Classifier for Pediatric Pneumonia Detection. SN Computer Science. 1. 359. 10.1007/s42979-020-00373-y.

[15] Darici, Muazzez & Dokur, Zumray & Ölmez, Tamer. (2020). Pneumonia Detection and Classification Using Deep Learning on Chest X-Ray Images.

International Journal of Intelligent Systems and Applications in Engineering. 8. 177. 10.18201/ijisae.2020466310.

[16] Qu, Wendi & Balki, Indranil & Méndez, Mauro & Valen, John & Levman, Jacob & Tyrrell, Pascal. (2020). Assessing and mitigating the effects of class imbalance in machine learning with application to X-ray imaging. International journal of computer assisted radiology and surgery. 15. 10.1007/s11548-020-02260-6.

[17] Win, Khin & Maneerat, Noppadol & Sreng, Syna & Hamamoto, Kazuhiko. (2021). Ensemble Deep Learning for the Detection of COVID-19 in Unbalanced Chest X-ray Dataset. Applied Sciences. 11. 10528. 10.3390/app112210528.

[18] Fujiwara, Koichi & Huang, Yukun & Hori, Kentaro & Nishioji, Kenichi & Kobayashi, Masao & Kamaguchi, Mai & Kano, Manabu. (2020). Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. Frontiers in Public Health. 8. 178. 10.3389/fpubh.2020.00178.

[19] Park, Seunghyun & Park, Hyunhee. (2021). Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic. Computing. 103. 1-24. 10.1007/s00607-020-00854-1.

[20] Koziarski, Michał. (2021). CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification. 1-8. 10.1109/IJCNN52387.2021.9533415.

[21] Junsomboon, Nutthaporn & Phienthrakul, Tanasanee. (2017). Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset. 243-247. 10.1145/3055635.3056643.

[22] Blagus, Rok & Lusa, Lara. (2015). Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics. 16. 10.1186/s12859-015-0784-9.