Extraction and Preprocessing of PPG Data from the MIMIC III Database

Viktor Meglenovski, Stojan Delenikov,
Bojan Dukovski and Hristina Miteva
Faculty of Computer Science and Engineeting
Skopje, North Macedonia
viktor.meglenovski@students.finki.ukim.mk
stojan.delenikov@students.finki.ukim.mk
bojan.dukovski@students.finki.ukim.mk
hristina.miteva@students.finki.ukim.mk

Bojana Koteska, Magdalena Kostoska and
Ana Madevska Bogdanova
Faculty of Computer Science and Engineeting
Skopje, North Macedonia
bojana.koteska@finki.ukim.mk
magdalena.kostoska@finki.ukim.mk
ana.madevska.bogdanova@finki.ukim.mk

***Abstract*—In this study, data was extracted from the Physionet MIMIC-III clinical database, which contained diverse medical records of patients who were admitted to the critical care units of Beth Israel Deaconess Medical Center between 2001 and 2012. Our research focused on PPG signals and SpO2 values, which were subjected to preprocessing and filtering in Python. The processed PPG data, together with the corresponding SpO2 values, were categorized based on the interval of SpO2 signal measurement, i.e., either one second or one minute. Subsequently, the filtered data was stored on a private ownCloud server, where it will be employed to enhance the database and facilitate the development of deep learning models for SpO2 prediction from one-channel PPG signals. At present, 340 GB of filtered data has been stored, which corresponds to approximately 2100 patients.**

***Index Terms*—photoplethysmography, oxygen saturation, database preprocessing, MIMIC III, Physionet**

I. Introduction

A photoplethysmogram (PPG) signal is a waveform generated through the use of a simple technique that employs infrared light to detect changes in blood volume in the microvascular bed of tissue. Stated differently, PPG signals are capable of sensing the rate of blood flow resulting from the heart's pumping action [1]. When assessing the graph representation of the PPG signal, each peak corresponds to a heartbeat; hence, if the graph displays 60 peaks, the heart rate is deemed to be 60 beats per minute. This technique provides crucial information pertaining to the cardiovascular system, facilitating the diagnosis, monitoring, and screening of a wide range of ailments, including heart attack, stroke, and heart failure [2].

The measurement of oxygen saturation (SpO2) involves the use of a pulse oximeter, which provides an indication of the percentage of oxygen present in the blood. Oxygen saturation represents the fraction of oxygen-saturated hemoglobin in relation to the total hemoglobin present in an individual's blood. The normal range of SpO2 levels in humans is between 97 and 100 percent. Should the SpO2 measurement fall below 95 percent, immediate medical attention is recommended. The SpO2 signals can be utilized to identify various lung diseases by measuring the amount of oxygen present in the blood. Moreover, during anesthesia and surgery, SpO2 monitoring is crucial in ensuring that the patient is receiving an adequate supply of oxygen [3], [4].

The calculation of SpO2 involves determining the ratio of the AC to DC components of the measured PPG signal. These components provide information on the heart rate during systole and diastole, as well as the respiratory rate during a specified time period, typically 60 seconds. By establishing this correlation over a specific duration, the SpO2 value can be estimated. This technique is widely used in the medical field for monitoring patients' oxygen levels, especially in critical care settings. To calibrate the measured photoplethysmographic signals for each type of commercial pulse-oximeter sensor, an empirical approach is employed, which involves in vitro measurement of SpO2 in extracted arterial blood through co-oximetry [5]. By utilizing Artificial Neural Networks (ANN) or Machine Learning models, it is possible to rapidly and accurately predict SpO2 from a single-channel PPG signal, thus overcoming the limitations imposed by the traditional R-value based calibration method utilized in signal processing methods [6].

Several research papers conduct preprocessing on the PPG signals obtained from the MIMIC-III Waveform Database for different purposes. As an instance, the authors in [7] retrieve blood pressure values from the PPG signals, which they preprocessed and filtered using Matlab. In [8], the authors used scalograms generated out of transmissive PPG signals collected from MIMIC-III database to diagnose diabetes. Lombardi at al. presented strategy for database preparation for training a sepsis detection system based on the utilization of only plethysmographic data from the MIMIC-III database [9].

This research paper aims to enhance the existing database presented in [10], which is used for building deep learning models to determine SpO2 values from one-channel PPG signals. To achieve this, we preprocess the PPG signals from the MIMIC-III Waveform Database in Python and create a filtered PPG signal database along with corresponding SpO2 values. The resultant database will serve as a valuable resource for researchers seeking to improve the accuracy of SpO2 prediction using one-channel PPG signals.

The subsequent sections of this paper are organized as follows: The Materials and Methods section presents a comprehensive description of the utilized database, elaborating on the complete process of data extraction, filtering, and storage on the server. In the Results section, various instances of preprocessed PPG data are demonstrated, along with an indication of the reasons for signal rejection such as flat lines, flat peaks, or NaN values. Lastly, the Conclusion section summarizes the paper's objectives and provides a brief overview of its contents.

II. Materials and Methods

A. MIMIC-III database

The MIMIC-III Waveform Database is an extensive centralized repository that contains crucial information regarding patients admitted to critical care units. The database comprises of 67,830 record sets pertaining to approximately 30,000 patients admitted to intensive care units. These record sets include physiologic waveforms and time series of vital signals, which were gathered from bedside patient monitors. The waveform signals consist of digitized signals, such as ECG, ABP, respiration, and PPG, while the numerics typically include vital signs such as heart and respiration rates, pulse, SpO2, systolic and diastolic blood pressure, etc. The database is the result of collaborative efforts by researchers at the Massachusetts Institute of Technology (MIT), Beth Israel Deaconess Medical Center (BIDMC), and Philips Healthcare, and was published on April 7, 2020 as version 1.0. The uncompressed size of this database is approximately 6.7 TB. The MIMIC-III database provides a valuable resource for researchers to develop algorithms and models for predicting outcomes, such as mortality, in critically ill patients. It has been widely used in the development of artificial intelligence models, as well as in clinical research studies [11]–[13].

B. Extraction of PPG signals and SpO2

Each database entry comprises a physiologic waveform and a corresponding numeric record, both of which are accompanied by header files that contain information about the measured signals. The physiologic waveform record captures up to eight signals that are simultaneously digitized at 125Hz. In contrast, the numeric record contains at least ten time-series of essential signals that are sampled either once per second or once per minute. It is essential to highlight that not all signals are monitored continuously throughout the entire duration of the record.

The signals that are of interest in this research are the PPG signals which if present can be found in the physiologic waveform records, as well as the SpO2 vital signals which if present can be found in the numeric records.

The length of the records is typically a few days in duration, but some records are shorter, and others which are several weeks long.

The extraction process begins by selecting only those records which contain both PPG and SpO2 signals. Subsequently, the PPG signal is divided into segments, with the number of segments equivalent to the length of the SpO2 time series in the corresponding record. This approach ensures that each PPG segment corresponds to one SpO2 value, with no skipped values. Examples of a raw PPG signal and a PPG segment with visible noise are presented in Fig. 1 and Fig. 2, respectively.

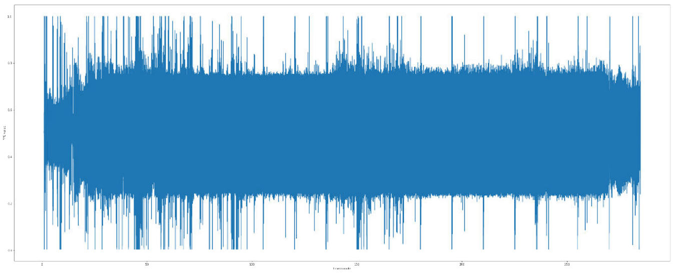

Fig. 1. Example of a raw PPG signal.

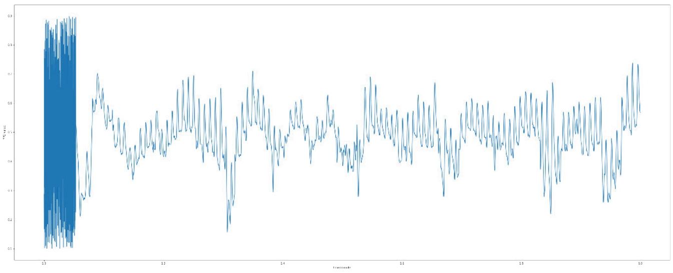

Fig. 2. Example of a raw PPG segment with noise.

Given that the PPG signal is sampled at a frequency of 125Hz, while the SpO2 is measured either once per second or once per minute, the length of each PPG segment should be either 125 or 7500 (125*60) samples. PPG segments that have lengths other than 125 or 7500 are excluded since they imply gaps in the monitoring of the signals that cannot be easily resolved.

On the other hand, if the length of the PPG segment is either 125 or 7500 samples, it indicates that there were no interruptions in the monitoring of the signals, and there is a corresponding SpO2 value available for every second or minute of PPG monitoring.

If the record passes this length check, its PPG signal is normalized. Each PPG segment, along with its corresponding SpO2 value, is passed through a series of filters that determine whether the data should be saved or discarded.

C. PPG signal filtering

The extracted PPG segments and corresponding SpO2 values are subjected to multiple filters to assess their quality. These filters are designed to determine if the data meets the required standards. The selection of the preprocessing steps and filters follows the methodology proposed in [7], with additional modifications to improve the results.

The initial filter evaluates whether the SpO2 value is NaN. In case of a NaN value, the corresponding PPG segment and SpO2 value are disregarded.

The subsequent filter applied is the ”flat lines” filter. It scans the PPG segment using a 15-window size and computes the proportion of flat lines present in the signal. For each PPG value, it determines whether the entire window of values, starting from that point, is either 0 or NaN. If so, it flags it as a flat line starting from that value. Finally, it calculates the percentage of PPG values that correspond to a flat line and discards the segment and corresponding SpO2 value if it exceeds 20%. This filter ensures that a maximum of 20% of the PPG segment comprises flat lines.

After passing the ”flat lines” filter, the PPG segment undergoes the ”flat peaks” filter, which follows a similar process. The filter analyzes the PPG segment by using a 5-window size and calculates the percentage of flat peaks in the signal. However, unlike the ”flat lines” filter, it identifies flat peaks when all PPG values in the window are the same but not equal to 0 or NaN. The same threshold of 20% is utilized for this filter.

Once the PPG segment has passed the preceding filters, it proceeds to the final filter that detects and removes noise at the start and end of the signal if present. Noise is characterized by an absolute difference of more than 10 units between consecutive PPG values.

D. Data storing format

The information is processed in batches of 25 records, and for each recored, two files are generated to store the data. One file holds the signals where SpO2 measurements are taken every second, while the other file corresponds to signals where SpO2 measurements are taken every minute. Additionally, a dictionary object consisting of the record ID, PPG segment represented as an array, and the corresponding SpO2 value is created. Based on whether the SpO2 was measured every second or every minute for that record, the dictionary object is placed into one of two resulting arrays.

Once all the data has been extracted and filtered, it is crucial to store it in a format that can be easily processed by a computer. In this study, Pickle format has been selected for its ability to serialize and deserialize almost any object using the Python programming language. It is particularly useful for storing and transferring large datasets in a compact, binary format, which satisfies this study’s requirements. Furthermore, compared to CSV files, Pickle format is considerably faster and uses compression techniques that can reduce file sizes by nearly half [14].

E. Data storage

The Python script responsible for processing and storing the data was executed on Google Colab, which had limited memory and storage resources. Consequently, some larger records were processed locally. The preprocessed data has been stored on an ownCloud server, and to accomplish this, the ownCloud Python library was imported. This library enabled the script to establish a connection with an ownCloud instance and upload the file contents to the server.

The ownCloud software is installed on virtual machine in our faculty data centre. We use the ownCloud Community Edition, as we are small to medium organization and our intended use is to run ownCloud with all basic functionalities on-premises by our self. The server is a virtualized resource with 16 cores, 16GB memory and has 1TB dedicated storage.

III. Results

So far, 8000 records have been preprocessed in 320 batches with 25 records each. Overall, this amounts to approximately 340 GB of accumulated data, which has been segregated into two folders. Each of these folders contains 320 files, corresponding to the 320 processed batches, and these files are created based on frequency of SpO2 readings. The organization of files in this structured manner facilitates efficient management of the data, making it more accessible for future use.

In the case when SpO2 is measured every second, the file sizes start from 9 KB to 3 GB, while in the case where SpO2 is measured every minute from 16.3 MB to 2 GB.

The figures below depict examples of accepted and rejected PPG segments and the results obtained from the filters applied. The plots were generated using the wfdb plot items function.

Figure 3 illustrates a PPG segment that has successfully passed through all filters, including the NaN filter, ”flat lines” filter, and ”flat peaks” filter.

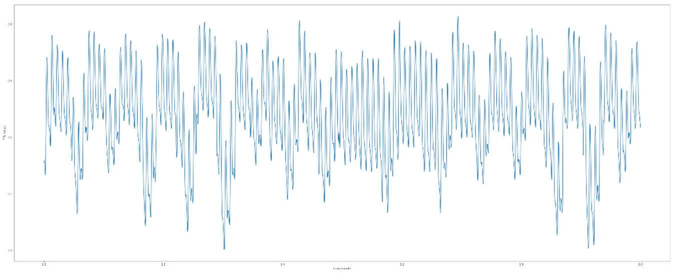

Fig. 3. Example of valid PPG segment that has passed all filters.

PPG segments that contain more than 20 percent NaN values or exhibit flat lines or flat peaks are deemed unsuitable and are consequently rejected. The figures below illustrate various causes of rejection of PPG segments. It is worth noting that these segments are not stored in the database to prevent any possible negative impact on the quality of data analysis.

Figure 4 displays an example of a rejected PPG segment due to the presence of NaN values.

Similarly, Fig. 5 illustrates a rejected PPG segment that contains flat lines.

Figure 6 depicts a rejected PPG segment that exhibits flat peaks.

Finally, Fig. 7 shows an example of a rejected PPG segment due to flat lines and flat peaks.

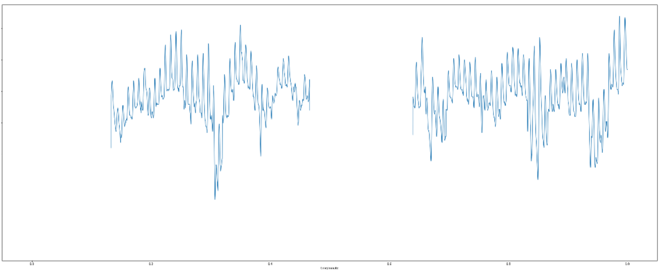

Fig. 4. Example of a rejected PPG segment due to NaN values.

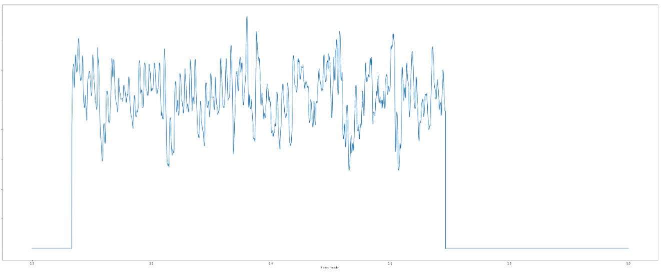

Fig. 5. Example of a rejected PPG segment due to flat lines.

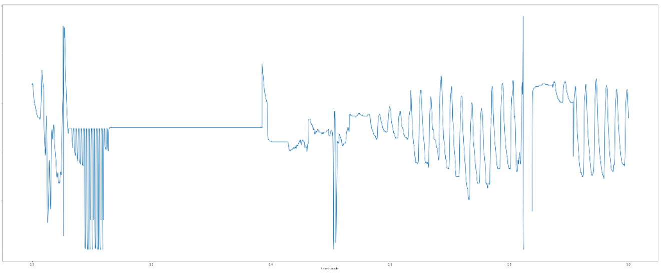

Fig. 6. Example of a rejected PPG segment due to flat peaks.

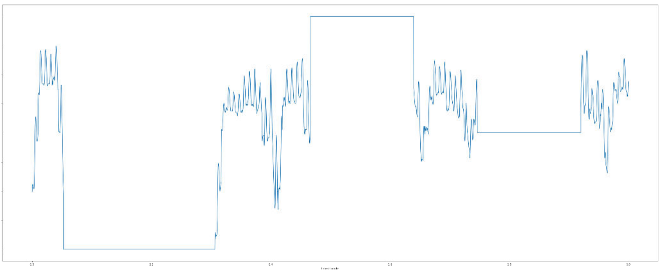

Fig. 7. Example of a rejected PPG segment due to flat lines and flat peaks.

Figure 8 and Fig. 9 demonstrate a PPG segment with noise before and after the noise-removing filter has been applied, respectively.

Currently, the database contains 320 files saved in separate directories for SpO2 measurements taken every second, and an additional 320 files for measurements taken every minute. Upon completing the preprocessing of the entire database, we anticipate generating 2400 more files for each SpO2 measurement, resulting in a total of 5400 files written to the server.

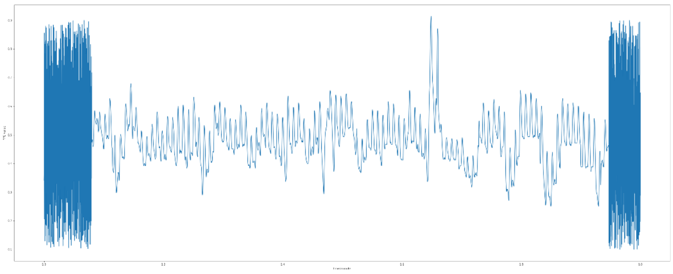

Fig. 8. Example of a PPG segment with noise before applying the noise removing filter.

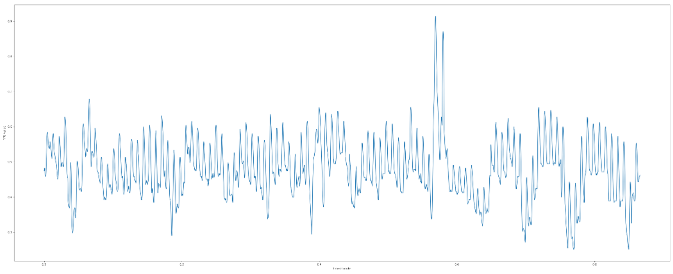

Fig. 9. The same PPG segment from Fig. 3. after applying the noise removing filter

IV. Conclusion

In this paper, we preprocess the PPG data from MIMIC-III Waveform Database, collected by bedside patient monitors in intensive care units, and create a database of PPG signals accompanied by the appropriate SpO2 numeric measurements.

For every record gathered in the database, after assuring that both SpO2 and PPG signals are present, PPG segments are divided into subsegments based on the SpO2 sampling frequency.

Firstly, the PPG signals are normalized and different filters are applied to identify flat lines and flat peaks, with a 20 percent threshold level. As a result of these steps, about 30 percent of the preprocessed records have been saved in the database. Currently, there is approximately 340 GB of saved data. This data is of big interest for building deep and machine learning models for prediction on the value of SpO2 from the PPG signals.

It is essential to note that the preprocessing of such medical noisy signals is a resource-intensive process that needs to be done carefully to ensure that the resulting data is of high quality and reliable. NaN values, which refer to missing data, can significantly affect the accuracy of the analysis. Similarly, flat lines and flat peaks can distort the shape of the waveform, leading to incorrect results and conclusions.

Acknowledgment

The work presented in this paper is partially supported by the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia and ”Smart Patch for Life Support Systems” - NATO project G5825 SP4LIFE.

References

[1] S. Cheriyedath, "Photoplethysmography (ppg)," *News-Medical. net*, 2019.

[2] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," *International journal of biosensors & bioelectronics*, vol. 4, no. 4, p. 195, 2018.

[3] D. Šoštarić, G. Mester, and S. Dorner, "Mobile ecg and spo2 chest pain subjective indicators of patient with gps location in smart cities," *Interdisciplinary Description of Complex Systems: INDECS*, vol. 17, no. 3-B, pp. 629–639, 2019.

[4] C. A. Haque, S. Hossain, T.-H. Kwon, and K.-D. Kim, "Comparison of different methods to estimate blood oxygen saturation using ppg," in *2021 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2021, pp. 792–794.

[5] M. Nitzan, A. Romem, and R. Koppel, "Pulse oximetry: fundamentals and technology update," *Medical Devices (Auckland, NZ)*, vol. 7, p. 231, 2014.

[6] S. Venkat, M. T. P. A. PS, A. Alex, S. Preejith, D. Christopher, J. Joseph, M. Sivaprakasam *et al.*, "Machine learning based SpO2 computation using reflectance pulse oximetry," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 482–485.

[7] G. Slapničar, N. Mlakar, and M. Luštrek, "Blood pressure estimation from photoplethysmogram using a spectro-temporal deep neural network," *Sensors*, vol. 19, no. 15, p. 3420, 2019.

[8] V. B. Srinivasan and F. Foroozan, "Deep learning based non-invasive diabetes predictor using photoplethysmography signals," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1256–1260.

[9] S. Lombardi, P. Partanen, and L. Bocchi, "Detecting sepsis from photoplethysmography: strategies for dataset preparation," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022, pp. 2286–2289.

[10] B. Koteska, A. M. Bodanova, H. Mitrova, M. Sidorenko, and F. Lehocki, "A deep learning approach to estimate spo2 from ppg signals," in *Proceedings of the 9th International Conference on Bioinformatics Research and Applications*, 2022, pp. 142–148.

[11] B. Moody, G. Moody, M. Villarroel, G. Clifford, and I. Silva III, "Mimic-iii waveform database (version 1.0)," *PhysioNet*, vol. 3, 2020.

[12] A. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database sci," *Data*, vol. 3, no. 160035, pp. 10–1038, 2016.

[13] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[14] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.